The error exponent with delay for lossless source coding

Cheng Chang and Anant Sahai
Wireless Foundations,
University of California at Berkeley
cchang@eecs.berkeley.edu, sahai@eecs.berkeley.edu

***Abstract*— In channel coding, reliable communication takes place at rates below capacity at the fundamental cost of end-to-end delay. Error exponents tell us how much faster convergence is when we settle for less rate. For lossless source coding, entropy takes the place of capacity and error exponents tell us how much faster convergence is when we use more rate. While in channel coding without feedback the block error exponent is a good proxy for studying the more fundamental tradeoff with fixed end-to-end delay, it is not so in source coding. Block-coding error exponents are quite conservative (despite being tight!) when it comes to the tradeoff with delay. Nonblock codes can achieve much better performance with fixed delay and we present both the fundamental bound and how to achieve it in a delay-universal manner. The proof gives substance to Shannon's cryptic statement about how the duality between source and channel coding is like the duality between the past and the future.**

I. Introduction and Problem Setup

Shannon closed his seminal paper on lossy source coding [1] with an intriguing comment:

> "[The duality between 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."

While there has been much work exploring the relationship between source and channel coding, no connection to the past and the future has emerged so far. In this paper, we demonstrate such a connection by looking at the fundamental error exponent with respect to fixed delay between the time a message is first made known to the encoder and its deadline at the decoder. For channel coding without feedback, the dominant error event is caused by the channel's "future" behavior in that errors can be forced by mild channel misbehavior during the period between the message arrival time and its deadline [2], [3]. This sort of mild misbehavior is exactly what is bounded by the traditional sphere-packing arguments for block coding. When feedback is allowed, the encoder can do flow-control and become robust to such mild misbehavior even when facing fixed-delay deadlines [4], [3]. The fundamental limits are instead given by the "focusing bound" [3].

This paper develops a lossless source-coding counterpart to the focusing bound. This reveals that the dominant source of errors is the "past" misbehavior of the source, over which the encoder has no control. Since the communication medium is a noiseless fixed-rate bit-pipe, the future is entirely under the control of the encoder. This bound is also asymptotically achievable using fixed-to-variable codes whose variable-rate nature is smoothed out by using a queue.

A. Review of block source coding results

The discrete memoryless source $\mathcal{S}$ generates iid random variables x_i from a finite alphabet $\mathcal{X}$ according to distribution p_{x}. Without loss of generality, assume $p_{\mathsf{x}}(x) > 0$, $\forall x \in \mathcal{X}$. A rate R block source coding system for n source symbols consists of a encoder-decoder pair $(\mathcal{E}_n, \mathcal{D}_n)$ where

$$\mathcal{E}_n : \mathcal{X}^n \longrightarrow \{0,1\}^{\lfloor nR \rfloor}, \qquad \mathcal{E}_n(x_1^n) = b_1^{\lfloor nR \rfloor}$$
$$\mathcal{D}_n : \{0,1\}^{\lfloor nR \rfloor} \longrightarrow \mathcal{X}^n, \qquad \mathcal{D}_n(b_1^{\lfloor nR \rfloor}) = \widehat{x}_1^n$$

The probability of block decoding error is $P(\mathsf{x}_1^n \neq \widehat{\mathsf{x}}_1^n) = P(\mathsf{x}_1^n \neq \mathcal{D}_n(\mathcal{E}_n(\mathsf{x}_1^n)))$.

Shannon proved that arbitrarily small error probabilities are achievable by letting n get big as long as the encoder rate is larger than the entropy of the source, $R > H(p_{\mathsf{x}})$. Furthermore, it turns out that the error probability goes to zero exponentially in n.

Lemma 1: (From [5]) For a discrete memoryless source $\mathsf{x} \sim p_{\mathsf{x}}$ and encoder rate $R < \log_2 |\mathcal{X}|$,

$\forall \epsilon > 0$, $\exists K(\epsilon) < \infty$, $\forall n \geq 0$, $\exists$ a block encoder-decoder pair $\mathcal{E}_n, \mathcal{D}_n$ satisfying

$$P(\mathsf{x}_1^n \neq \widehat{\mathsf{x}}_1^n) \leq K(\epsilon) 2^{-n(E_b(R)-\epsilon)} \tag{1}$$

This result is asymptotically tight, in the sense that for any sequence of encoder-decoder pairs $\mathcal{E}_n, \mathcal{D}_n$,

$$\limsup_{n\to\infty} -\frac{1}{n} \log_2 P(\mathsf{x}_1^n \neq \widehat{\mathsf{x}}_1^n) \leq E_b(R) \tag{2}$$

where $E_b(R)$ is defined as the block source coding reliability function with the form:

$$E_b(R) = \min_{q: H(q) \geq R} D(q \| p_{\mathsf{x}}) \tag{3}$$

Paralleling the definition of the Gallager function for channel coding [6], define

$$E_0(\rho) = (1+\rho) \log_2 [\sum_x p_{\mathsf{x}}(x)^{\frac{1}{1+\rho}}] \tag{4}$$

Then, as mentioned as an exercise in [5]:

$$\begin{aligned} E_b(R) &= \sup_{\rho \geq 0} \{\rho R - E_0(\rho)\} \qquad (5) \\ &= D(p_{\mathsf{x}_{\rho_R}} \| p_{\mathsf{x}}) \end{aligned}$$

where $p_{\mathsf{x}_{\rho_R}}$ is the tilted distribution of parameter ρ_R satisfying $H(p_{\mathsf{x}_{\rho_R}}) = R$. The tilted distribution of parameter $\rho \in (-1, \infty)$ of a distribution p_{x} is: $\forall x \in \mathcal{X}$

$$p_{\mathsf{x}_\rho}(x) = \frac{p_{\mathsf{x}}(x)^{\frac{1}{1+\rho}}}{\sum_s p_{\mathsf{x}}(s)^{\frac{1}{1+\rho}}} \tag{6}$$

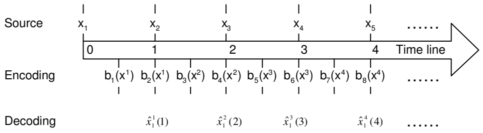

Fig. 1. Delay-universal sequential source coding at $R = 2$

$p_{x_0} = p_x$ and as shown in [7], $\frac{\partial H(p_{x_\rho})}{\partial \rho} \geq 0$ and is generally strictly positive unless p_x is uniform on $\mathcal{X}$. Thus the optimal ρ_R corresponding to R is unique unless p_x is uniform over $\mathcal{X}$ which is a uninteresting case.

B. Sequential Source Coding

Rather than being known in advance, the source symbols enter the encoder in a real-time fashion. We assume that the source $\mathcal{S}$ generates one source symbol x per second from a finite alphabet $\mathcal{X}$. The j'th source symbol x_j is not known at the encoder until time j. Rate R operation means that the encoder sends 1 binary bit to the decoder every $\frac{1}{R}$ seconds. For obvious reasons, we focus on cases with $H(p_x) < R < \log_2 |\mathcal{X}|$.

Definition 1: A sequential encoder-decoder pair $\mathcal{E}, \mathcal{D}$ are sequence of maps. $\{\mathcal{E}_j\}, j = 1, 2, ...$ and $\{\mathcal{D}_j\}, j = 1, 2, ...$ The outputs of $\mathcal{E}_j$ are the outputs of the encoder $\mathcal{E}$ from time $j-1$ to j.

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

The outputs of $\mathcal{D}_j$ are the decoding decisions of all the arrived source symbols by time j based on the received binary bits up to time j.

$$\mathcal{D}_j : \{0,1\}^{\lfloor jR \rfloor} \longrightarrow \mathcal{X}$$
$$\mathcal{D}_j(b_1^{\lfloor jR \rfloor}) = \widehat{x}_1^{j-d}$$

Where $\widehat{x}_1^{j-d}(j)$ is the estimation, at time j, of x_1^{j-d} and thus has end-to-end delay of d seconds. In a delay-universal scheme, the decoder emits revised estimates for all source-symbols so far. A rate $R = 2$ delay-universal sequential source coding system is illustrated in Figure 1.

For sequential source coding, it is important to study the symbol by symbol decoding error probability instead of the block coding error probability.

Definition 2: A family of rate R sequential source codes $\{(\mathcal{E}^d, \mathcal{D}^d)\}$ are said to achieve delay-reliability $E_s(R)$ if and only if: $\forall i$

$$\liminf_{d \to \infty} \frac{-1}{d} \log_2 P(x_i \neq \widehat{x}_i(i+d)) \geq E_s(R)$$

II. Upper bound on $E_s(R)$

To bound the best possible error exponent with fixed delay, we consider a genie-aided encoder/decoder pair and translate the block-coding bounds of [5] to the fixed delay context. The arguments are analogous to the "focusing bound" derivation in [3] for the case of channel coding with feedback.

Theorem 1: For fixed-rate encodings of discrete memoryless sources, it is not possible to achieve an error exponent with fixed-delay better than

$$E_s^*(R) = \inf_{\alpha > 1} \frac{1}{\alpha - 1} E_b(\alpha R) \tag{7}$$

Proof: For simplicity of exposition, we ignore integer effects arising from the finite nature of $d, R,$ etc. For every $\alpha > 1$ and delay d, consider a code running over its fixed-rate noiseless channel till time $\frac{\alpha d}{\alpha - 1}$. By this time, the decoder will have committed to estimates for the source symbols up to time $i = \frac{d}{\alpha - 1}$. The total number of bits used during this period is $\frac{\alpha d}{\alpha - 1} R$.

Now consider a genie that gives the encoder access to the first i source symbols at the beginning of time, rather than forcing the encoder to get the source symbols one at a time. Simultaneously, loosen the requirements on the decoder by only demanding correct estimates for the first i source symbols by the time $\frac{\alpha}{\alpha - 1} d$. 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 error probability of the new problem is clearly also a bound for the original problem. Furthermore, the new problem is just a fixed-length block-coding problem requiring the encoding of i source symbols into $\frac{\alpha}{\alpha - 1} dR$ bits. The rate per symbol is

$$\begin{aligned} (\frac{\alpha}{\alpha - 1} dR) \frac{1}{i} &= (\frac{\alpha}{\alpha - 1} dR) \frac{\alpha - 1}{d} \\ &= \alpha R \end{aligned}$$

Theorem 2.15 in [5] tells us that such a code has a probability of error that is at least exponential in $iE_b(\alpha R)$. Since $i = \frac{d}{\alpha - 1}$, this translates into an error exponent of at most $\frac{E_b(\alpha R)}{\alpha - 1}$ with parameter d.

Since this is true for all $\alpha > 1$, we have a bound on the reliability function $E_s(R)$ with fixed delay d:

$$E_s(R) \leq \inf_{\alpha > 1} \frac{1}{\alpha - 1} E_b(\alpha R)$$

The right hand side is defined to be $E_s^*(R)$. The minimizing α tells how much of the past ($\frac{d}{\alpha - 1}$) is involved in the dominant error event. □

This bound can be rewritten in terms of the ρ parameter to get a form paralleling the symmetric channel case from [3].

Corollary 1:

$$E_s(R) \leq E_0(\rho^*) \tag{8}$$

where ρ^* satisfies $R = \frac{E_0(\rho^*)}{\rho^*}$.

Proof: For convenience, define

$$G_R(\rho) = \rho R - E_0(\rho) \tag{9}$$

and notice that $G_R(\rho^*) = 0$. As shown in [7]:

$$\begin{aligned} &\frac{dG_R(\rho)}{d\rho} = R - H(p_{x_\rho}), \quad \frac{d^2 G_R(\rho)}{d\rho^2} \leq 0 \\ &G_R(0) = 0, \quad G_R(\infty) < 0 \end{aligned} \tag{10}$$

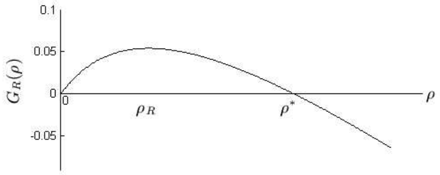

Fig. 2. $G_R(\rho)$

$G_R(\rho)$ is a concave ∩ function as illustrated in Figure 2. Consider $\rho_R \ge 0$ for which $R - H(p_{x_{\rho_R}}) = 0$. By concavity, we know that $\rho_R < \rho^*$. Write:

$$\begin{aligned} F_R(\alpha,\rho) &= \frac{1}{\alpha-1}(\rho\alpha R - E_0(\rho)) \\ &= \rho R + \frac{G_R(\rho)}{\alpha-1} \end{aligned}$$

Then for any $\alpha \in (1, \frac{\log_2|\mathcal{X}|}{R})$,

$$\begin{aligned} \tfrac{\partial F_R(\alpha,\rho)}{\partial\rho} = \tfrac{\alpha R - H(p_{x_\rho})}{\alpha-1}, &\qquad \frac{\partial^2 F_R(\alpha,\rho)}{\partial\rho^2} \le 0 \\ F_R(\alpha,0) = 0, &\qquad F_R(\alpha,\infty) < 0 \end{aligned} \qquad (11)$$

So for any $\alpha \in (1, \frac{\log_2|\mathcal{X}|}{R})$ let $\rho(\alpha)$ be so $F_R(\alpha,\rho(\alpha))$ is maximized. $\rho(\alpha)$ is thus the unique solution to:

$$\alpha R - H(p_{x_{\rho(\alpha)}}) = 0$$

Define $\alpha^* = \frac{H(p_{x_{\rho^*}})}{R}$ which satisfies

$$\begin{aligned} \alpha^* &= \frac{H(p_{x_{\rho^*}})}{R} \le \frac{H(p_{x_\infty})}{R} = \frac{\log_2|\mathcal{X}|}{R} \\ \alpha^* &= \frac{H(p_{x_{\rho^*}})}{R} \ge \frac{H(p_{x_{\rho_R}})}{R} = 1 \end{aligned}$$

Since $\alpha^* R - H(p_{x_{\rho^*}}) = 0$, ρ^* maximizes $F_R(\alpha^*,\rho)$ over all ρ.

Using (5) and the above analysis:

$$\begin{aligned} E_s(R) &\le \inf_{\alpha>1} \frac{1}{\alpha-1} E_b(\alpha R) \\ &= \inf_{\alpha>1} \sup_{\rho>0} \frac{1}{\alpha-1}(\rho\alpha R - E_0(\rho)) \\ &= \inf_{\alpha>1} \sup_{\rho>0} F_R(\alpha,\rho) \\ &\le \sup_{\rho>0} F_R(\alpha^*,\rho) \\ &= F_R(\alpha^*,\rho^*) \\ &= \rho^* R \end{aligned} \qquad (12)$$

This proves the desired upper bound. □

III. Achievable exponents

To lower-bound the error exponent with delay, we give an explicit fixed-rate coding scheme that is a minor variation of the scheme analyzed in [8]. In [3], an achievable scheme was given for channel coding with feedback when there was an additional low-rate noise-free link that could carry flow-control information. In source-coding, there is just a noise-free link and the flow control information is made implicit by using a prefix-free code. Just as in [3], the queuing behavior is what will determine the achieved delay error exponent. Rather than repeating the from first principles' derivation of [3], here we give an alternate derivation by applying Cramáer's theorem.

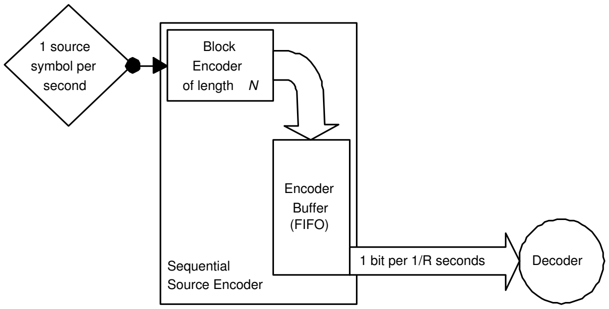

Fig. 3. Sequential source coding using a variable-length code

A. A sequential variable length source coding scheme

We are interested in the performance with asymptotically large delays d. A block-length N is chosen that is much smaller than the target end-to-end delays, while still being large enough. For a discrete memoryless source and large block-lengths N, the best possible variable-length code is given in [5] and consists of two stages: first describing the type of the block $\vec{x}$ using $O(|\mathcal{X}|\log N)$ bits and then describing which particular realization has occurred by using a variable $H(\vec{x})$ bits. The overhead $O(|\mathcal{X}|\log N)$ is asymptotically negligible and the code is also universal in nature.

While the above code will also work, for simplicity of analysis, we instead consider a Shannon-code built for a particular tilted distribution for X.

Definition 3: The instantaneous code $\mathcal{C}_{N,\lambda}$ for $\lambda > -1$ is a mapping from $\mathcal{X}^N$ to a variable number of binary bits.

$$\mathcal{C}_{N_\lambda}(x_1^N) = b_1^{l(x_1^N)}$$

where $l(x_1^N)$ is the codeword length for source sequence x_1^N. The first bit is always 1 and the rest of the codewords are the Shannon codewords based on the λ tilted distribution of p_x

$$\begin{aligned} l(x_1^N) &= 1 + \lceil -\log_2 \frac{p_x(x_1^N)^{\frac{1}{1+\lambda}}}{\sum_{s_1^N \in \mathcal{X}^N} p_x(s_1^N)^{\frac{1}{1+\lambda}}} \rceil \\ &\le 2 - \sum_{i=1}^N \log_2 \frac{p_x(x_i)^{\frac{1}{1+\lambda}}}{\sum_{s\in\mathcal{X}} p_x(s)^{\frac{1}{1+\lambda}}} \end{aligned} \qquad (13)$$

From (13), the longest code length l_N^λ is

$$l_N^\lambda \le 2 - N\log_2 \frac{p_{x_{\min}}^{\frac{1}{1+\lambda}}}{\sum_{s\in\mathcal{X}} p_x(s)^{\frac{1}{1+\lambda}}} \qquad (14)$$

Where $p_{x_{\min}} = \min_{x\in\mathcal{X}} p_x(x)$. The constant 2 is insignificant compared to N.

This variable-length code is turned into a fixed rate R code as follows:

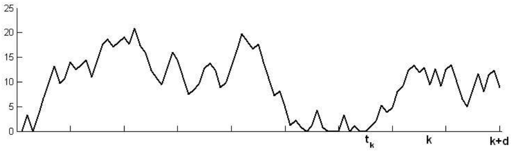

Fig. 4. Number of bits in the buffer: Δ_t

Definition 4: The sequential source coding scheme is illustrated in Figure 3. At time kN, $k = 1, 2, ...$ the encoder $\mathcal{E}$ uses the variable length code $\mathcal{C}_{N_\lambda}$ to encode the k'th source block $\vec{x}_k = x_{(k-1)N+1}^{kN}$ into a binary sequence $b(k)_1, b(k)_2, ...b(k)_{l(\vec{x}_k)}$.

This codeword is pushed into a FIFO queue with infinite buffer-size. The encoder drains a bit from the queue every $\frac{1}{R}$ seconds. If the queue is empty, the encoder simply sends 0's to the decoder until there are new bits pushed in the queue.

The decoder knows the variable length code book and the prefix-free nature[1] of the Shannon code guarantees that everything can be decoded correctly.

B. Sequential error events

The number of the bits Δ_k in the encoder buffer at time kN, is a random walk process with negative drift and a reflecting barrier. An example sample-path is illustrated in Figure 4. At time $(k+d)N$, the decoder can make an error in estimating $\vec{x}_k$ if and only if part of the variable length code for source block $\vec{x}_k$ is still in the encoder buffer.

Since the FIFO queue drains deterministically, it means that when the k-th block's codeword entered the queue, it was already doomed to miss its deadline of d. Formally, for an error to occur, the number of bits in the buffer $\Delta_k \geq \lfloor dNR \rfloor$. Thus, meeting a specific end-to-end latency constraint over a fixed-rate noiseless link is like the buffer overflow events analyzed in [8]. Define the random time $t_k N$ to be the last time before time kN when the queue was empty. A missed-deadline will occur only if $\sum_{i=t_k+1}^{k} l(\vec{x}_i) > (d+k-t_k)NR$.

For arbitrary $1 \leq t \leq k-1$, define the error event

$$P_N^{k,d}(t) = P(\sum_{i=t}^{k} l(\vec{x}_i) > (d+k-t)NR)$$

Using the following lemma, we will derive a tight, in the large deviation sense, upper bound on $P_N^{k,d}(t)$.

Lemma 2: Cramáer theorem [9] Consider iid random variables $Y_i \in \Sigma = \{\sigma_1, ...\sigma_{|\Sigma|}\}$, $Y_i \sim p_y$ and $f : \Sigma \to \mathcal{R}^+$, $X_i = f(Y_i)$. Write $S_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. For any closed set $F \subseteq \mathcal{R}^+$.

$$P(S_n \in F) \quad \leq \quad (n+1)^{|\Sigma|} 2^{-n \inf_{x \in F} I(x)}$$

Where the rate function is [9]:

$$I(x) \quad = \quad \inf_{\nu : \sum_{i=1}^{|\Sigma|} \nu_i f(\sigma_i) = x} D(\nu \| p_y)$$

ν is a distribution defined on Σ. It is shown in [9] that:

$$I(x) \quad = \quad \sup_{\rho \in \mathcal{R}} \{\rho x - \log_2(\sum_{i=1}^{|\Sigma|} p_{y_i} 2^{\rho f(\sigma_i)})\}$$

Write $I(x, \rho) = \rho x - \log_2(\sum_{i=1}^{|\Sigma|} p_{y_i} 2^{\rho f(\sigma_i)})$, then $\forall x > 0$, $\forall \rho < 0$, $I(x, \rho) < 0$. Obviously $I(x, 0) = 0$, which means that the ρ to maximize $I(x, \rho)$ is positive. This implies that $I(x)$ is monotonically increasing with x.

Using the definition of $l(\vec{x})$:

$$\begin{aligned}
&\log_2(\sum_{\vec{x} \in \mathcal{X}^N} p_x(\vec{x}) 2^{\lambda l(\vec{x})})\} \\
&\leq 2\lambda + N \log_2[(\sum_x p_x(x)^{1-\frac{\lambda}{1+\lambda}})(\sum_x p_x(x)^{\frac{1}{1+\lambda}})^\lambda] \\
&= 2\lambda + N(1+\lambda)\log_2[\sum_x p_x(x)^{\frac{1}{1+\lambda}}] \\
&= 2\lambda + NE_0(\lambda) \qquad (15)
\end{aligned}$$

The 2λ is insignificant and so as a simple corollary of the Cramáer theorem, we have an upper bound on $P_N^{k,d}(t)$.

$$\begin{aligned}
P_N^{k,d}(t) &= P(\sum_{i=t+1}^{k} l(\vec{x}_i) \geq (d+k-t)NR) \\
&= P(\frac{1}{k-t}\sum_{i=t+1}^{k} l(\vec{x}_i) \geq \frac{(d+k-t)NR}{k-t}) \\
&\leq (k-t)^{|\mathcal{X}|^N} 2^{-E_N(\mathcal{C}_{N_\lambda}, R, k-t, d)}
\end{aligned}$$

Where:

$$\begin{aligned}
&E_N(\mathcal{C}_{N_\lambda}, R, k-t, d) \\
\geq\ &(k-t)\sup_{\rho \in \mathcal{R}^+}\{\rho\frac{(d+k-t)NR}{k-t} - \log_2(\sum_{\vec{x} \in \mathcal{X}^N} p_x(\vec{x}) 2^{\rho l(\vec{x})})\} \\
\geq\ &(k-t)[\lambda\frac{(d+k-t)NR}{k-t} - \log_2(\sum_{\vec{x} \in \mathcal{X}^N} p_x(\vec{x}) 2^{\lambda l(\vec{x})})] \\
\geq\ &(k-t)N(\lambda\frac{(d+k-t)R}{k-t} - \frac{2\lambda}{N} - E_0(\lambda)) \\
=\ &dN\lambda R + (k-t)N[G_R(\lambda) - \frac{2\lambda}{N}] \qquad (16)
\end{aligned}$$

C. Lower bound on $E_s(R)$

Theorem 2: For any $\epsilon > 0$, by appropriate choice of N, λ, it is possible to achieve an error exponent with delay of $E_0(\rho^*) - \epsilon$ universally over large enough delays d for the ρ^* satisfying $R = \frac{E_0(\rho^*)}{\rho^*}$.

Proof: For the sequential source coding scheme of $\mathcal{C}_{N,\lambda}$, the decoding error for the k'th source block at time $(k+d)N$ is $P_N^{k,d}$. $t_k N$ is the last time before k when the buffer is empty.

[1] The initial 1 is not really required since the decoder knows the rate at which source-symbols are arriving at the encoder. Thus, it knows when the queue is empty and does not need to even interpret the 0s it receives.

$$\begin{aligned} P_N^{k,d} &= \sum_{t=0}^{k-1} P(t_k = t, \sum_{i=t+1}^{k} l(\vec{x}_i) \geq (d+k-t)NR) \\ &\leq \sum_{t=0}^{k-1} P(\sum_{i=t+1}^{k} l(\vec{x}_i) \geq (d+k-t)NR) \\ &\leq \sum_{t=0}^{k-1} (k-t+1)^{|\mathcal{X}|^N} 2^{-E_N(\mathcal{C}_{N_\lambda}, R, k-t, d)} \\ &= 2^{-dN\lambda R} \sum_{t=0}^{k-1} (k-t+1)^{|\mathcal{X}|^N} 2^{-(k-t)N[G_R(\lambda) - \frac{\lambda}{N}]} \end{aligned}$$

The above equality is true for all N, λ. Pick $\lambda = \rho^* - \frac{\epsilon}{R}$. From the discussion at the end of previous section, we know that $G_R(\lambda) > 0$. Choose $N > \frac{\lambda}{G_R(\lambda)}$. Then define

$$K(\epsilon, R, N) = \sum_{i=0}^{\infty} (i+1)^{|\mathcal{X}|^N} 2^{-iN[G_R(\lambda) - \frac{\lambda}{N}]} < \infty$$

which is guaranteed to be finite since dying exponentials dominate polynomials in sums. Thus:

$$P_N^{k,d} \leq K(\epsilon) 2^{-dN\lambda R} = K(\epsilon, R, N) 2^{-dN(\rho^* R - \epsilon)}$$

where the constant $K(\epsilon, R)$ does not depend on the delay in question. Since $E_0(\rho^*) = \rho^* R$ and the encoder also is not targeted to the delay dN, this scheme achieves the desired exponent delay-universally. □

This theorem effectively proves that $E_s^*(R) = E_0(\rho^*)$ is asymptotically achievable. For every realization of $\vec{x}$, the code-lengths of this code differ from the optimal universal code by at most $O(\log(N))$, which is insignificant compared to N. Thus, this delay-exponent can also be attained universally over both discrete-memoryless sources and delays d.

IV. An example of suboptimal coding

Large N are not needed to do substantially better than block-coding in the fixed delay context. The advantages of variable-length encoding are so great that even very simple VL-codes will outperform the best possible block-code. Consider the following ternary example from [10].

Suppose $p_x(a) = 0.9$, $p_x(b) = 0.05$ and $p_x(c) = 0.05$. Consider rate $R = \frac{3}{2}$. The best possible block coding error exponent is 1.474. Consider the following sub-optimal variable length sequential coding scheme with $N = 2$. Map (a, a) to 00, and other source symbol pairs are mapped to $1000, 1001, ...1111$. Simple birth-death Markov chain analysis reveals that this achieves an error exponent of 6.27 with fixed-delay despite being suboptimal even as a VL-code!

V. Conclusions and Future Work

The lossless source-coding version of the uncertainty focusing bound was developed and used to show that fixed-to-variable length coding is optimal from an end-to-end latency point of view, even when the deadlines are specified in terms of a fixed latency. In a very precise sense, lossless source-coding is like channel coding *with feedback* and using block-codes results in a substantial loss in the error exponents. While the parametric form of the focusing bound is the same in the channel-coding and source-coding cases, the interpretation is a bit different. In source-coding, the dominant error events always involve only the past. As the rate varies, the only change is how much of the past is involved. For channel coding with feedback, *both* the past and the future are involved since the past is now known while the future remains only partially controllable. In channel-coding without feedback, only the future is involved since the past is entirely unknown.

A similar story should hold for the error exponents of point-to-point *lossy* source coding. Once again, the source-coding context guarantees that only the past behavior will matter in the dominant error event. The results here hint that VQ followed by variable-rate entropy-coding may also be optimal in the fixed-delay setting. A more interesting direction is in extending our understanding of sequential distributed coding for correlated sources (Slepian-Wolf source coding). So far, we have only achievable exponents in general [7] and a good upper bound only for symmetric cases with side-information at the receiver [11]. For the case considered in [11], it turns out that only the future behavior of the source matters. We suspect that depending on the rate point and the nature of the source, different combinations of the past and future will be involved in the dominant error event and the optimal codes will in general have a mixed nature involving both queuing and binning.

Acknowledgments

The authors thank Sanjoy Mitter, Tunc Simsek, Stark Draper, and Sandeep Pradhan for helpful discussions along the way and acknowledge the support of NSF ITR Award: CNS - 0326503 in making this work possible.

References

[1] 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.

[2] 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.

[3] A. Sahai, “Why block length and delay are not the same thing,” *IEEE Trans. Inform. Theory*, submitted. [Online]. Available: http://www.eecs.berkeley.edu/˜sahai/Papers/FocusingBound.pdf

[4] ——, “How to beat the sphere-packing bound with feedback,” submitted to ISIT, 2006.

[5] I. Csisz*á*r and J. Körner, *Information Theory*. Budapest: Akad*é*miai Kiad*ó*, 1986.

[6] R. G. Gallager, *Information Theory and Reliable Communication*. New York, NY: John Wiley, 1971.

[7] C. Chang, S. Draper, and A. Sahai, “Sequential random binning for streaming distributed source coding,” *In preparation*.

[8] F. Jelinek, “Buffer overflow in variable length coding of fixed rate sources,” *IEEE Transactions on Information Theory*, vol. 14, pp. 490–501, 1968.

[9] A. Dembo and O. Zeitouni, *Large Deviations Techniques and Applications*. Springer, 1998.

[10] S. Draper, C. Chang, and A. Sahai, “Sequential random binning for streaming distributed source coding,” *ISIT*, 2005.

[11] C. Chang and A. Sahai, “Upper bound on error exponents with delay for lossless source coding with side-information,” submitted to ISIT, 2006.