Sequential Coding of Gauss–Markov Sources over Packet-Erasure Channels with Feedback

Anatoly Khina, Ashish Khisti, and Babak Hassibi

Abstract—We consider the problem of sequential transmission of Gauss–Markov sources over packet-erasure channels with a possibly delayed output feedback. For the case of instantaneous feedback, we determine the optimal squared error distortions for given transmission rates for all time instants, and construct a scheme that achieves all of them simultaneously. This establishes the optimal rate–distortion region for sequential coding of Gauss–Markov sources without packet erasures, as a special case. For the case of delayed feedback, we connect the problem to that of compression with side information that is known at the encoder and may be known at the decoder — where the most recent packets serve as side information that may have been erased. We conclude the paper by demonstrating that the loss due to a delay by one time instant is rather small.

I. INTRODUCTION

Sequential coding of sources is increasingly finding applications, such as real-time video streaming, and cyberphysical and networked control. Such systems use compressed packet-based transmission and are prone to packet losses or “erasures”.

Without packet erasures, this setting was introduced and treated for the two-source case by Viswanathan and Berger [1] and for more users in [2]–[4]. For the special case of Gauss–Markov sources, an explicit expression for the achievable sum-rate for given distortions was derived in [2], [5] and extended for the (general) jointly Gaussian three-source case in [6].

The more intriguing case of sequential coding in the presence of packet erasures was treated for various erasure models. The case when only the first packet is prone to an erasure was considered in [7]. A more general approach which trades between the performance given all previously sent packets and the performance given only the last packet was proposed in [8]. For random independent identically distributed (i.i.d.) packet erasures, a hybridization between pulse-code modulation (PCM) and differential PCM (DPCM), termed leaky DPCM, was proposed in [9] and for the case of very low erasure probability in [10]. The scenario in which the erasures occur in bursts was considered in [11], [12]. Here a sequence of source vectors sampled from a Gauss–Markov process in the temporal dimension must be encoded sequentially and reconstructed with zero delay at the decoder. The channel introduces a burst of erasures of a certain maximum length and the decoder is not required to reconstruct the sequences that fall in the erasure period and a recovery window following it.

However, all of these works assume no feedback is available at the encoder, namely that the encoder does not know whether a transmitted packet successfully arrives to the decoder or is erased in the process.

In this work we consider sequential coding of Gauss–Markov sources over a channel with random i.i.d. packet erasures and a possibly delayed (ACK/NACK) output feedback.

We first treat the case of no packet drops, for which we propose a greedy scheme and prove that it is in fact globally optimal in the limit of large frames, in Section III. Interestingly, this technique allows to determine the whole rate–distortion region for sequential coding of Gauss–Markov sources, therby extending the result of Ma and Ishwar [2], [5] for this case. We further show that all of these results hold also for the case of packet drops with an instantaneous feedback, in Section IV.

For the case of delayed feedback, the encoder does not know whether the most recently transmitted packets arrived or not; we view these recent packets as side information (SI) available at the encoder and possibly at the decoder, which allows us to employ the results of Kaspi [13] for this scenario along with their specialization for the Gaussian case by Perron et al. [14].

We describe in detail the adaptation of the proposed scheme to the case of “delayed-by-one” feedback in Section V and demonstrate that its loss compared to the case of instantaneous feedback is rather small.

II. PROBLEM STATEMENT

We now describe the model of the source, channel, and the admissible encoder and decoder both of which are required to be causal in this work, where the former has access to (possibly delayed) output feedback.

We denote random variables by lower-case letters with temporal subscripts \( (a_t) \), and random vectors (“frames”) of length \( N \) by boldface possibly accented lower-case letters with temporal subscripts \( (\mathbf{a}, \mathbf{a}_t) \). \( \mathbb{N} \) denotes the set of natural numbers. All other notations represent deterministic scalars.

We assume that the communication spans the time interval \([1, T]\), where \( T \in \mathbb{N} \).

Source: Consider a Gauss–Markov source \( \{s_t\} \), whose entries are vectors of length \( N \) with i.i.d. samples along the

\[ a_k \]

This scenario can also be viewed as special case of the results of Heegard and Berger [15], where the SI is not available at the encoder, by adjusting the distortion measure and “augmenting” the source [16]. Interestingly, knowing the SI at the encoder allows to improve the optimal performance of this scenario in the Gaussian case; see Remark 6.
spatial dimension, that satisfy the temporal Markov relation (we assume $s_0 = 0$ for convenience)

$$s_t = \alpha_t s_{t-1} + w_t, \quad t = 1, \ldots, T,$$

where $\{\alpha_t\}$ are known process coefficients that satisfy $|\alpha_t| < 1$, and the entries of $\{w_t\}$ are i.i.d. across along the spatial dimension, are Gaussian and mutually independent across time, of zero mean and variances $\{W_t\}$.

Denote by $S_t$ the average power of an entry of $s_t$ (with $S_0 = 0$). Then, we obtain the following recursive relation:

$$S_t = \alpha_t^2 S_{t-1} + W_t, \quad t = 1, \ldots, T, \quad (2a)$$

$$S_0 = 0. \quad (2b)$$

If we specialize the source process into that of fixed parameters, namely,

$$\alpha_t \equiv \alpha, \quad W_t \equiv W, \quad t = 1, \ldots, T,$$

then its power converges to

$$S_\infty = \frac{W}{1 - \alpha^2}.$$

We shall refer to such a process as asymptotically stationary.

**Channel:** At time $t$, the packet $f_t$ is sent over a packet-erasure channel; it arrives with probability $\beta$ to the decoder, and is lost (erased) with probability $(1 - \beta)$. We denote by $b_t \sim \text{Ber}(\beta)$ the binary event of successful packet arrival: $b_t = 1$ if the packet at time $t$ arrives successfully, and $b_t = 0$ otherwise. Then, the channel can be written as

$$g_t = \begin{cases} f_t & b_t = 1 \\ \oplus & b_t = 0 \end{cases}$$

where $\oplus$ denotes an erasure.

**Causal encoder with feedback:** Observes $s_t$ and $g_{t-\ell}$ at time $t$ and generates a packet $f_t \in \{1, 2, \ldots, 2^{NR}\}$ via a causal function $F_t$ of the source sequence and the delayed by $\ell$ feedback:

$$f_t = F_t(s_1, \ldots, s_t, g_1, g_2, \ldots, g_{t-\ell}),$$

where $\ell \in \mathbb{N}$ is the feedback arrival latency.

**Causal decoder:** Reconstructs an estimate $\hat{s}_t$ of $s_t$ at time $t$, using the received packets until time $t$ via a causal function $G_t$:

$$\hat{s}_t = G_t(s_1, g_1, g_2, \ldots, g_t).$$

**Distortion:** Define the distortion at time $t$, conditioned on all previous packet erasures, by

$$D_t \triangleq \frac{1}{N} \mathbb{E} \left[ \|s_t - \hat{s}_t\|^2 \right]_{b_1, \ldots, b_t},$$

where $\|\cdot\|$ denotes the Euclidean norm.

Define further the average distortions (without conditioning) by

$$\bar{D} \triangleq \frac{1}{N} \mathbb{E} \left[ \|s_t - \hat{s}_t\|^2 \right].$$

For asymptotically stationary source processes, we further define the time-averaged (steady-state) distortion by the limit of the Cesàro means of $\{D_t\}$:

$$\bar{D}_T \triangleq \lim_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} D_t.$$

**Goal:** Minimize $\{\bar{D}_T\}$ for a given $R$.

**Definition (Distortion-rate region):** The distortion-rate region of rate $R$ is the closure of all achievable distortions, for any $N$, however large; its inverse is the rate–distortion region.

We shall see in the sequel, that if an instantaneous feedback is available ($\ell = 1$), simultaneous optimality of $\{D_t\}$ for all $t \in [1, T]$ is possible. We shall further construct a scheme based on the work of [14] for the case of $\ell > 1$.

### III. Sequential Coding without Packet Erasures

The optimal achievable distortions $\{D_t\}$ for the model of Section II without packet erasures, namely, for the case of $\beta = 1$ (which guarantees $f_t \equiv g_t, \forall t \in [1, T]$), are stated in the following theorem.

**Theorem 1** (No erasures: rate–distortion region). The distortion–rate region of sequential coding without packet erasures is given by the distortions $\{D_t\}$ that satisfy

$$D_t = (\alpha^2 D_{t-1} + W_t) 2^{-2R}, \quad t = 1, \ldots, T,$$

$$D_0 = 0.$$  (7b)

**Remark 1.** Theorem 1 establishes the optimal rate–distortion region for the “causal encoder–causal decoder” setting of Ma and Ishwar [2] for the case of Gauss–Markov sources. We note that Ma and Ishwar [2] provide an explicit result only for the sum-rate for the Gauss–Markov case [5]. Torbatian and Yang [6] extend the sum-rate result to the case of three jointly Gaussian sources (which do not necessarily constitute a Markov chain). Our work, on the other hand, allows to fully characterize the rate–distortion region for the case of Gauss–Markov sources, and can be easily extended to non-equal rates.

**Remark 2.** The results and proof (provided in the sequel) of Theorem 1 imply that optimal greedy quantization at every step — which is achieved via Gaussian backward [17, Ch. 10.3] or forward [17, pp. 338–339] channels — becomes optimal when $N$ is large. Moreover, it achieves the optimum for all $t \in [1, T]$ simultaneously, meaning that there is no tension between minimizing the current distortion and future distortions.

To prove this theorem we first construct the optimal greedy scheme and determine its performance in Section III-A. We then show that it is in fact (globally) optimal when $N$ goes to infinity, by constructing an outer bound for this scenario, in Section III-B.
A. Achievable

We construct an inner bound using the optimal greedy scheme to be described next.

In this scheme all the quantizers are assumed to be minimum mean square error (MMSE) quantizers. We note that the quantized values of such quantizers are orthogonal to the resulting quantization errors.

**Scheme** (No erasures).

**Encoder:** At time $t$:
- Generates the prediction error
  \[
  \hat{s}_t = s_t - \alpha_t \hat{s}_{t-1},
  \]
  where $\hat{s}_0 = 0$ and \{${\hat{s}}_t | t = 1, \ldots, T$\} are the previous source reconstructions at the decoder which will be determined in the sequel in [9] [recall further [5][6]].
- Generates $f_t$ by quantizing the prediction error $\hat{s}_t$ using the optimal MMSE quantizer of rate $R$ and frame length $N$; denote the quantized reconstruction by $\hat{s}_t$.
- Sends $f_t$ over the channel.

**Decoder:** At time $t$:
- Receives $g_t \equiv f_t$.
- Generates a reconstruction $\hat{s}_t$ of the prediction error $\hat{s}_t$.
- Generates an estimate $\hat{s}_t$ of $s_t$:
  \[
  \hat{s}_t = \alpha \hat{s}_{t-1} + \hat{s}_t.
  \]

The optimal achievable distortions \{${D}_t$\} of this scheme for long frame lengths $N$, are as follows.

**Assertion 1** (No erasures: inner bound). Let $\epsilon > 0$, however small. Then, the expected distortion of the scheme at time $t \in [1, T]$ satisfies the recursion
\[
D_t \leq \left( \alpha_t^2 D_{t-1} + W_t \right) 2^{-2R} + \epsilon,
\]
where $D_0 = 0$, \begin{equation}
(10a)
\end{equation}
for a large enough $N$.

**Proof:** First note that the error between $s_t$ and $\hat{s}_t$, denoted by $e_t$, is equal to
\[
e_t \equiv s_t - \hat{s}_t = (\hat{s}_t + \alpha \hat{s}_{t-1}) - (\alpha \hat{s}_{t-1} + \hat{s}_t)
\]
\begin{equation}
(11a)
\end{equation}
and is of average power $D_t$, where \begin{equation}
(11b)
\end{equation}
gives rise to
\[
\hat{s}_t = \alpha \hat{s}_{t-1} - \hat{s}_t
\]
\begin{equation}
(11c)
\end{equation}
which is fed to the quantizer at time $t$, to obtain $\hat{s}_t$.

Since $w_t$ is independent of $e_{t-1}$, the average power of the entries of $\hat{s}_t$ is equal to
\[
\hat{s}_t = \alpha_t^2 D_{t-1} + W_t.
\]

Using the property that the rate–distortion function of an i.i.d. source with zero mean, given power and desired squared-error distortion is upper bounded by that of a white Gaussian source with the same power (see, e.g., [17 pp. 338–339]), we arrive at the following recursion
\begin{equation}
D_t \leq \left( \alpha_t^2 D_{t-1} + W_t \right) 2^{-2R} + \epsilon,
\end{equation}
and hence \begin{equation}
(7)
\end{equation}
is achievable within an arbitrarily small $\epsilon > 0$, for a sufficiently large $N$.

\[\Box\]

B. Converse

We shall now construct an outer bound that coincides with the inner bound of Assertion [1] for large frame lengths $N$.

**Assertion 2** (No erasures: outer bound). Consider the setting of Section [7] with $\beta = 1$. Then, the average achievable distortion $D_t$ at time $t \in [1, T]$ is bounded from below by
\[
D_t \geq \frac{D_t^*}{1 - \epsilon},
\]
where $D_t^*$ satisfies \begin{equation}
(12a)
\end{equation}
by induction, where \{${D}_t$\} denotes the sequence that satisfies \begin{equation}
(7)
\end{equation}
and the expected entropy-power of $s_t$ given the (specific) past packets \{${f}_1, \ldots, f_{T-1}$\} is with respect to \{${f}_1, \ldots, f_{T-1}$\}.

**Basic step** ($t = 1$). First note that, since $s_0 = 0$ and \begin{equation}
(12b)
\end{equation}
comprises i.i.d. Gaussian entries of variance $1$, \begin{equation}
(12b)
\end{equation}
is satisfied with equality. To prove \begin{equation}
(12b)
\end{equation}
we use the fact that the optimal achievable distortion $D_1$ for a Gaussian source ($s_1 = \alpha_w$) with i.i.d. entries of power $W_1$ and rate $R$ is dictated by its rate–distortion function [17 Ch. 10.3.2]:
\[
D_1 \geq W_1 2^{-2R}.
\]

**Inductive step.** Let $k \geq 2$ and suppose \begin{equation}
(12)
\end{equation}
is true for all $t \leq k - 1$. We shall now prove that it holds also for $t = k$.

To that end, re-write the distortion $D_{k}$ at time $k$ as
\[
D_k = \frac{1}{N} \mathbb{E} \left[ \|s_k - \hat{s}_k\|^2 \right]
\]
\[
= \frac{1}{N} \mathbb{E} \left[ \|s_k - \hat{s}_k\|^2 \left| f_1, \ldots, f_{k-1} \right| \right].
\]

We shall next bound the inner conditional expectation.

\[\Box\]
By bounding it from below by the rate–distortion function and applying to it the Shannon lower bound [17] Ch. 10, we have

\[
\frac{1}{N} \mathbb{E} \left[ \| s_k - \hat{s}_k \|^2 \Big| f_1 = F_1, \ldots, f_{k-1} = F_{k-1} \right] \geq \frac{1}{2e} 2^{-2 \frac{1}{2} h(\mathbb{E}[s_k|f_1=F_1,\ldots,f_{k-1}=F_{k-1}]-NR)}
\]

By taking the expectation over \( f_1, \ldots, f_{k-1} \) we attain (12a).

The following set of inequalities establishes (12b):

\[
\begin{align*}
E_{F_1, \ldots, F_{k-1}} \left[ \frac{1}{2e} 2^{-2 \frac{1}{2} h(\mathbb{E}[s_k|f_1=F_1,\ldots,f_{k-1}=F_{k-1}]-NR)} \right] \\
= 2^{-2R} E_{F_1, \ldots, F_{k-1}} \left[ \frac{1}{2e} \mathbb{E}[h(s_k|F_1,\ldots,F_{k-1}=F_{k-1})] \right] \\
\geq 2^{-2R} \frac{2}{2e} \mathbb{E}[h(s_k|F_1,\ldots,F_{k-1}=F_{k-1})] \\
\geq 2^{-2R} \mathbb{E}[h(s_k|F_1,\ldots,F_{k-1}=F_{k-1})] + \mathbb{E}[W_k] \\
\geq 2^{-2R} \mathbb{E}[h(s_k|F_1,\ldots,F_{k-1}=F_{k-1})] + W_k \\
\geq 2^{-2R} \mathbb{E}[h(s_k|F_1,\ldots,F_{k-1}=F_{k-1})] + W_k \\
= D_k + W_k
\end{align*}
\]

where (14a) follows from (1), (14b) follows from the entropy-power inequality [17] Ch. 17, (14c) holds since \( W_k \) is Gaussian, the scaling property of differential entropies and Jensen’s inequality:

\[
\begin{align*}
E_{F_{k-1}} \left[ 2^{-2R} \mathbb{E}[h(s_k|F_1,\ldots,F_{k-1}=F_{k-1})] \Big| f_1 = F_1, \ldots, f_{k-2} = F_{k-2} \right] \\
\geq 2^{-2R} \mathbb{E}[h(s_k|F_1,\ldots,F_{k-1}=F_{k-1})] \\
\equiv 2^{-2R} \mathbb{E}[h(s_k|F_1,\ldots,F_{k-1}=F_{k-1})] \\
\geq 2^{-2R} \mathbb{E}[h(s_k|f_1=F_1,\ldots,f_{k-2}=F_{k-2})] \\
= D_k
\end{align*}
\]

(14d) follows from the following standard set of inequalities:

\[
NR \geq H(f_{k-1}|f_1 = F_1, \ldots, f_{k-2} = F_{k-2}) \\
\geq I(s_k-1; f_{k-1}|f_1 = F_1, \ldots, f_{k-2} = F_{k-2}) \\
= h(s_k|f_1 = F_1, \ldots, f_{k-2} = F_{k-2}) \\
- h(s_k|f_1 = F_1, \ldots, f_{k-2} = F_{k-2})
\]

for (14c) we use the induction hypothesis, and (14d) holds true by the definition of \( \{D_k^*\} \) as the sequence that satisfies (17).

This concludes the proof of (12b) as desired.

C. Steady State of Asymptotically Stationary Sources

Assume here the special case of an asymptotically stationary source (3). For this case the steady-state average distortion is as follows.

Corollary 1 (No drops: steady state). Let \( \epsilon > 0 \), however small. Then, the expected time-averaged distortion \( \bar{D}_\infty \) of the scheme is equal to

\[
\bar{D}_\infty = W 2^{-2R} + \epsilon,
\]

for large enough \( N \).

Proof: Note that (16) is a fixed point of (17) (up to \( \epsilon \)). Now since \( \alpha < 1 \) and \( 2^{-2R} < 1 \), \( D_t \) converges to \( \bar{D}_\infty \). This can be easily proved as follows. Assume \( D_{t-1} \neq \bar{D}_\infty \) (otherwise we are already at the fixed point).

\[
D_t - \bar{D}_\infty = [(\alpha^2 D_{t-1} + W_t) 2^{-2R}] - [(\alpha^2 \bar{D}_\infty + W_t) 2^{-2R}]
\]

or equivalently

\[
\frac{D_t - \bar{D}_\infty}{D_{t-1} - \bar{D}_\infty} = \alpha^2 2^{-2R} < 1.
\]

Hence, if \( 0 \leq D_{t-1} - \bar{D}_\infty \), then \( 0 \leq D_t - \bar{D}_\infty \leq D_{t-1} - \bar{D}_\infty \)

and converges (exponentially fast) to \( \bar{D}_\infty \).

Remark 3. As is evident from the proof, the result of Corollary 1 remains true for any initial value \( D_0 \).

IV. SEQUENTIAL CODING WITH INSTANTANEOUS FEEDBACK

In this section we generalize the results of Section III by allowing packet erasures.

The optimal achievable distortions \( \{D_t\} \) for the model of Section II with instantaneous feedback, namely for the case of \( \ell = 1 \), are stated in the following theorem.

Theorem 2. The distortion–rate region of sequential coding with instantaneous feedback \( (\ell = 1) \), conditioned on \( \{b_1, \ldots, b_t\} \), is given by the \( \{D_t\} \) that satisfy

\[
D_t = (\alpha^2 D_{t-1} + W_t) 2^{-b_t 2R}, \quad t = 1, \ldots, T,
\]

\[
D_0 = 0.
\]

This theorem gives rise to the following distortion–rate region.

Corollary 2 (Instantaneous feedback: distortion–rate region). The distortion–rate region of sequential coding with instantaneous feedback \( (\ell = 1) \), is given by the average distortions \( \{\bar{D}_t\} \) that satisfy

\[
\bar{D}_t = (\alpha^2 \bar{D}_{t-1} + W_t) B, \quad t = 1, \ldots, T,
\]

\[
\bar{D}_0 = 0.
\]

where

\[
B \triangleq 1 - \beta (1 - 2^{-2R})
\]

Proof of Corollary 2. By taking the expectation over both sides of (17a), we have for every \( t = [1, T] \):

\[
\bar{D}_t = \mathbb{E} \left[ (\alpha^2 D_{t-1} + W_t) 2^{-b_t 2R} \right]
\]

\[
= (\alpha^2 \bar{D}_{t-1} + W_t) \mathbb{E} \left[ 2^{-b_t 2R} \right]
\]

\[
= (\alpha^2 \bar{D}_{t-1} + W_t) \left[ \beta + (1 - \beta)2^{-2R} \right],
\]

Given \( \{b_1, \ldots, b_t\} \), \( \{D_k|k = 1, \ldots, t\} \) are deterministic.
where (19b) holds since the random variable \( b_t \) is independent of \((s_1, \ldots , s_T, w_1, (dofs, w_T, s_1, \ldots , s_{t-1})\), and \( w_t \) is independent of \((s_1, \ldots , s_{t-1}, s_1, \ldots , s_{t-1})\). ■

To prove Theorem 2 we extend the scheme of Section III-A for the case of packet erasures and instantaneous feedback in Section IV-A. By extending the result of Section III-B, we establish the optimality of this scheme in the limit of large \( N \), in Section IV-B.

A. Achievable

We construct an inner bound by extending the greedy scheme of Section III-A.

Again, all quantizers are assumed to be minimum mean square error (MMSE) quantizers. We note that the quantized values of such quantizers are orthogonal to the resulting quantization errors.

**Scheme (Instantaneous feedback).**

**Encoder.** At time \( t \):
- Generates the prediction error
  \[ \hat{s}_t = s_t - \alpha_t \hat{s}_{t-1}, \]
  where \( \hat{s}_0 = 0 \) and \( \{s_t|t = 1, \ldots , T\} \) will be determined in the sequel [recall further (2)].
- Generates \( f_t \) by quantizing the prediction error \( \hat{s}_t \) using the optimal MMSE quantizer of rate \( R \) and frame length \( N \); denote the reconstruction from \( f_t \) and \( (g_1, \ldots , g_{t-1}) \) by \( Q_t(\hat{s}_t) \).
- Sends \( f_t \) over the channel.

**Decoder.** At time \( t \):
- Receives \( g_t \).
- Generates a reconstruction \( \hat{s}_t \) of the prediction error \( \hat{s}_t \):
  \[ \hat{s}_t = \left\{ \begin{array}{ll} Q_t(\hat{s}_t) & b_t = 1 \\ 0 & b_t = 0 \end{array} \right. \]
- Generates an estimate \( \hat{s}_t \) of \( s_t \):
  \[ \hat{s}_t = \alpha_t \hat{s}_{t-1} + \hat{s}_t. \]

The optimal achievable distortions \( \{D_t\} \) of this scheme for long frame lengths \( N \), are as follows. We show that these are the optimal achievable distortions in Section IV-B.

**Assertion 3 (Instantaneous feedback: inner bound).** Let \( \epsilon > 0 \), however small. Then, the expected distortion of the scheme at time \( t \in [1, T] \) given \( (b_1, \ldots , b_t) \) satisfies the recursion
\[
D_t \leq (\alpha_t^2 D_{t-1} + W_t) 2^{-b_t R} + \epsilon, \quad t = 1, \ldots , T,
\]
\[
D_0 = 0,
\]
for a large enough \( N \).

The proof of this assertion is identical to that in Section III-A.

\[ \hat{s}_t = 1 \] and \( \alpha_t \hat{s}_{t-1} = 1 \) are the MMSE estimators of \( s_{t-1} \) and \( s_t \), respectively, given all the past channel outputs.

**B. Converse**

We shall utilize the proof in Assertion III-B for the construction of a tight outer bound for the case where an instantaneous feedback is available.

**Assertion 4 (Instantaneous feedback: outer bound).** Consider the setting of Section I with \( \ell = 1 \). Then, the achievable distortion \( D_t \) at time \( t \in [1, T] \) given \( (b_1, \ldots , b_t) \) is bounded from below by \( D_t \geq D_t^* \), where \( D_t^* \) satisfies (17) with equality.

**Remark 4.** Clearly, the outer bound of Assertion 3 remains valid if the feedback is available with a delay \( \ell > 1 \), though not tight for this case.

**Remark 5.** The lower bound in Theorem 2 is valid even if \( \{b_t|t = 1, \ldots , T\} \) is revealed to the encoder and the decoder before transmission begins. This fact is used to prove Assertion 3.

**Proof:** To construct an outer bound, we first reveal all the erasure events \( \{b_1, \ldots , b_t\} \) to both the encoder and the decoder, before transmission begins; this can only improve the achievable distortions, and therefore is valid for our purposes.

Now note that this reduces this scenario to that of no erasures and a lower total transmission duration \( T \). This holds true due to the following two simple observations.

- The source at time \( t_1 \) can be represented as
  \[
  s_{t_1} = s_{t_0} \prod_{t=t_0+1}^{t_1} \alpha_t + \sum_{t=t_0+1}^{t_1} w_t \prod_{k=t+1}^{t_1} \alpha_k
  \]
  \[
  \triangleq \alpha_{1; t_0:t_1} s_{t_0} + w_{1; t_0:t_1}^\text{eff}
  \]
  \[
  \alpha_{1; t_0:t_1} \triangleq \prod_{t=t_0+1}^{t_1} \alpha_t
  \]
  \[
  w^\text{eff}_{1; t_0:t_1} \triangleq \sum_{t=t_0+1}^{t_1} w_t \prod_{k=t+1}^{t_1} \alpha_k
  \]
  where \( \prod_{i=1}^{j} \alpha_i \) is defined as equal to 1 if \( i > j \). Thus, a source sample \( s_{t_1} \) at time \( t_1 \) can be re-written in the form of Section I with \( s_{t_0} \), \( \alpha_{1; t_0:t_1} \) and \( w_{1; t_0:t_1}^\text{eff} \) taking the roles of \( s_{t_1} \), \( \alpha_t \) and \( w_t \), respectively, with \( 0 < t_0 < t_1 \).

Assume that
\[
\begin{align*}
  b_{t_0} &= 1, \\
  b_{t_0+1} &= b_{t_0+2} = \cdots = b_{t_0+k} = 0, \\
  b_{t_0+(k+1)} &= 1.
\end{align*}
\]

Using (21c), the MMSE estimator of \( s_{t_0} \) given \( \{f_{t_0}, f_{t_1}, \ldots , f_{t_{k-1}}\} \) is equal to
\[
\hat{s}_{t_0} (f_{t_0}, f_{t_1}, \ldots , f_{t_{k-1}}) = \alpha_{1; t_0:t_1} \hat{s}_{t_0} \left( f_{t_0}, f_{t_1}, \ldots , f_{t_{k-1}} \right)
\]
where \( \hat{s}_{t_0} \) is the MMSE estimator of \( s_{t_0} \) from \( \{f_{t_0}, f_{t_1}, \ldots , f_{t_{k-1}}\} \).

Noting that (17a) reduces to \( D_t = \alpha_t^2 D_{t-1} + W_t \) in case of an erasure and since these two observations mean that an erasure can be viewed as one sample less with appropriate adjustments of the parameters of the source, concludes the proof. ■
C. Steady State of Asymptotically Stationary Sources

For the special case of an asymptotically stationary source \( S \), the steady-state average distortion is given as follows.

**Corollary 3** (Instantaneous feedback: steady state). Let \( \epsilon > 0 \), however small. Then, the expected time-averaged distortion \( \hat{D}_\infty \) of the scheme is equal to

\[
\hat{D}_\infty = \frac{BW}{1 - \alpha^2 B} + \epsilon
\]

for large enough \( N \), where

\[
B = 1 - \beta \left(1 - 2^{-2R}\right).
\]

The proof is a straightforward extension of the proof of Corollary 1 and is therefore relegated to the appendix.

V. SEQUENTIAL CODING WITH DELAYED FEEDBACK

In this section we consider the case of a delayed-by-one feedback, i.e., \( \ell = 2 \) in (1).

To that end, we recall the following result by Perron et al. [14] Theorem 2], which is a specialization of the rate-distortion region established by Kaspi [13] Theorem 1] (and can be viewed also as a special case of [15] with some adjustments; see [16]) to the jointly Gaussian case, of the two-sided SI setting where the SI may or may not be available at the decoder.

**Theorem 3 (Kaspi-based).** Let \( S \) be an i.i.d. zero-mean Gaussian source of power \( S \), which is jointly Gaussian with SI \( y \), which is available at the encoder and satisfies \( s = y + z \) where \( z \) is an i.i.d. Gaussian noise of power \( Z \) that is independent of \( y \). Denote by \( \tilde{s}^+ \) and \( \tilde{s}^- \) the reconstructions of \( s \) with and without the SI \( y \), and by \( D^+ \) and \( D^- \) their mean squared error distortion requirements, respectively. Then, the smallest rate required to achieve these distortions is given by

\[
R_{\text{Kaspi}}(S, Z, D^-, D^+) = \begin{cases} 
0, & \text{if } D^- \geq S \text{ and } D^+ \geq Z \\
\frac{1}{2} \log \left( \frac{S}{D^-} \right), & \text{if } D^- < S \text{ and } D^+ \geq D^- \geq Z \\
\frac{1}{2} \log \left( \frac{Z}{D^+} \right), & \text{if } D^+ < Z \text{ and } D^- \geq D^+ \geq S - Z \\
\frac{1}{2} \log \left( \frac{S}{D^+} \right), & \text{if } D^+ \geq S \text{ and } D^- \geq D^+ \geq Z \text{ and } \left( D^+ \geq S - Z \text{ and } D^- \geq D^+ \geq S - Z \right)
\end{cases}
\]

where \( a \parallel b \triangleq \frac{ab}{a+b} \) denotes the harmonic mean of \( a \) and \( b \), and

\[
\Delta \triangleq \sqrt{(S-Z)(S-D^+)} - \sqrt{(Z-D^+)(D^- - D^+)}. 
\]

Remark 6. Surprisingly, as observed by Perron et al. [14], if the side-information signal \( y \) is not available at the encoder — corresponding to the case considered in [15] and [13] Theorem 2] — the required rate can be strictly higher than that in Theorem 1. This is in stark contrast to the case where the side-information is never available at the encoder and the case where the side-information is always available at the decoder studied by Wyner and Ziv [18], [19]. Knowing the SI at the encoder allows to (anti-)correlate the noise \( z \) with the quantization error — a thing that is not possible when the SI is not available at the encoder, as the two noises must be independent in that case. This allows for some improvement, though a modest one, as implied by the results for the dual channel problem [20] Proposition 1], [21].

In our case the previous packet will serve as the SI. Note that it is always available to the encoder; the decoder may or may not have access to it, depending whether the previous packet arrived or not — a thing unknown at the encoder during the transmission of the current packet.

The optimal tradeoff between \( D^+ \) and \( D^- \) for a given rate \( R \) will be determined by the probability of a successful packet arrival \( \beta \).

**Scheme** (Kaspi-based).

**Encoder.** At time \( t \):

- Generates the prediction error

\[
\tilde{s}_t \triangleq s_t - \alpha_t^2 \tilde{s}_{t-2}. 
\]

- Generates \( f_t \) by quantizing the prediction error \( \tilde{s}_t \) as in Theorem 3 where \( f_{t-1} \) is available as SI at the encoder and possibly at the decoder (depending on \( b_{t-1} \)) using the optimal quantizer of rate \( R \) and frame length \( N \) that minimizes the averaged over \( b_{t-1} \) distortion:

\[
D_t^{\text{Weighted}} = \beta D_t^+ + (1 - \beta) D_t^-; \tag{23}
\]

more precisely, since the encoder does not know \((b_{t-1}, b_t)\) at time \( t \):

- Denote the reconstruction of \( \tilde{s}_t \) at the decoder from \( f_t \) and \((g_1, \ldots, g_{t-1}) \) — namely given that \( b_t = 0 \) — by \( \bar{Q}(\tilde{s}_t) \), and the corresponding distortion by \( D_t^{\text{Weighted}} \).

- Denote the reconstruction from \( f_t \) and \((g_1, \ldots, g_{t-2}) \) — namely given that \( b_t = 1 \) and \( b_{t-1} = 0 \) — by \( \bar{Q}_t^+(\tilde{s}_t) \), and the corresponding distortion by \( D_t^+ \).

- Denote the reconstruction from \((f_{t-1}, f_t)\) and \((g_1, \ldots, g_{t-2}) \) — namely given that \( b_t = 1 \) and \( b_{t-1} = 1 \) — by \( \bar{Q}_t^+(\tilde{s}_t) \), and the corresponding distortion by \( D_t^+ \).

Then, the encoder sets \( \alpha_t Q^+(\tilde{s}_{t-1}) \) as possible SI available at the decoder to minimize \( D_t^{\text{Weighted}} \) as in (23).

**Decoder.** At time \( t \):

- Sends \( f_t \) over the channel.

- Receives \( g_t \).
Generates a reconstruction \( \hat{s}_t \) of the prediction error \( s_t \):

\[
\hat{s}_t = \begin{cases} 
Q^+_t(\hat{s}), & b_t = 1, b_{t-1} = 1 \\
Q^-_t(\hat{s}), & b_t = 1, b_{t-1} = 0 \\
0, & b_t = 0 
\end{cases} \tag{24}
\]

Generates an estimate \( \hat{\tilde{s}}_t \) of \( s_t \):

\[
\hat{\tilde{s}}_t = \alpha_t \hat{s}_{t-1} + \hat{s}_t .
\]

This scheme is the optimal greedy scheme whose performance is stated next, in the limit of large \( N \).

**Theorem 4.** Let \( \epsilon > 0 \), however small. Then, for a large enough \( N \), the expected distortion of the scheme at time \( t \in [2, T] \) given \( (b_1, \ldots, b_t) \) satisfies the recursion

\[
D_t = \begin{cases} 
D^+_t + \epsilon, & b_t = 1, b_{t-1} = 1 \\
D^-_t + \epsilon, & b_t = 1, b_{t-1} = 0 \\
\alpha^2 D_{t-1} + W + \epsilon, & b_t = 0 
\end{cases}
\]

\[
D_1 = D^+_1 = D^-_1 = W t 2^{-b_t 2R} + \epsilon
\]

where \( D^+_t \) and \( D^-_t \) are the distortions that minimize

\[
D_t^\text{Weighted} = \beta D^+_t + (1 - \beta) D^-_t ,
\]

such that the rate of Theorem 3 satisfies

\[
R^{\text{Kaspi}}(\alpha_t D^-_{t-1} + W, \alpha_t D^+_{t-1} + W, D^-_t, D^+_t) = R .
\]

The proof is again the same as that of Theorems 1 and 2 with \( \hat{s}_t \) generated as in (24).

**Remark 7.** Here, in contrast to the case of \( \ell = 1 \), evaluating the average distortions \{\( D_t \)\} in explicit form (recall Corollary 2) is much more challenging. We do it numerically, instead.

Somewhat surprisingly, the loss in performance of the Kaspi-based scheme due to the feedback delay (\( \ell = 2 \)) is rather small compared to the case where the feedback is available instantaneously (\( \ell = 1 \)) of Sections V-A and V-B for all values of \( \beta \). This is demonstrated in Fig. 1 where the performances of these schemes are compared along with the performances of the following three simple schemes for \( \alpha = 0.7, W = 1, \beta = 0.5, R = 2 \) (we derive their performance for the special case of an asymptotically stationary source):

- **No prediction:** A scheme that does not use prediction at all, as if the source samples were independent. This scheme achieves a distortion of

\[
D_t = \beta S_t 2^{-2R} + (1 - \beta) S_t, \quad t = 1, \ldots, T ,
\]

where \( S_t \) is the power of the entries of \( s_t \) as given in (2).

- **Assumes worst case (WC):** Since at time \( t \) the encoder does not know \( b_{t-1} \), a “safe” way would be to work as if \( b_{t-1} = 0 \). This achieves a distortion of

\[
D_t = \beta \left[ \alpha^2 D_{t-1} + W + \epsilon \right] + (1 - \beta) \left[ \alpha^2 D_{t-1} + W, t = 2, \ldots, T ,
\]

\[
D_0 = 0, \quad D_1 = W 2^{-2R} .
\]

- **Assumes best case (BC):** The optimistic counterpart of the previous scheme is that which always works as if \( b_{t-1} = 1 \). This scheme achieves a distortion of

\[
D_t = \beta \left[ \alpha^2 D_{t-1} + W \right] + (1 - \beta) \left[ \alpha^2 D_{t-1} + W, t = 2, \ldots, T ,
\]

\[
D_{t-1} W 2^{-2R} .
\]

VI. DISCUSSION: FEEDBACK WITH \( \ell > 2 \) DELAY

To extend the scheme of Section V for larger delays, a generalization of Theorem 3 is needed. Unfortunately, the optimal rate-distortion region for more than two decoders remains an open problem and is only known for the case when the source and the possible SIs form a Markov chain (“degraded”). Nonetheless, achievable regions for multiple decoders have been proposed in [15], which can be used for the construction of a scheme for \( \ell > 2 \).

**APPENDIX**

**Proof of Corollary 3**

Proof: Note that (22) is a fixed point of (18).

Since \( \alpha < 1 \) and \( B < 1 \), \( \mathbb{E} [D_t] \) converges to \( D_\infty \). This can be easily proved as follows. Assume \( D_{t-1} \neq D_\infty \) (otherwise we are already at the fixed point). Then,

\[
\tilde{D}_t - D_\infty = \left( \alpha^2 \tilde{D}_{t-1} + W \right) B - \left( \alpha^2 D_\infty + W \right) B
\]

or equivalently

\[
\frac{D_t - D_\infty}{D_{t-1} - D_\infty} = \alpha^2 B < 1 .
\]

For \( \beta \) values close to 0 or 1, the loss becomes even smaller as in these cases using the scheme of Section V-B that assumes that the previous packet arrived or was erased, respectively, becomes optimal.
Hence, if $0 \leq \bar{D}_{t-1} - \bar{D}_\infty$, then

$$0 \leq \bar{D}_t - \bar{D}_\infty \leq \bar{D}_{t-1} - \bar{D}_\infty$$

and converges exponentially fast to $\bar{D}_\infty$.

**ACKNOWLEDGMENT**

The authors thank Y. Su from Caltech for valuable discussions.

**REFERENCES**

[1] H. Viswanathan and T. Berger, “Sequential coding of correlated sources,” *IEEE Trans. Inf. Theory*, vol. 46, no. 1, pp. 236–246, Jan. 2000.

[2] N. Ma and P. Ishwar, “On delayed sequential coding of correlated sources,” *IEEE Trans. Inf. Theory*, vol. 57, no. 6, pp. 3763–3782, 2011.

[3] E.-H. Yang, L. Zheng, and D.-K. He, “Rate distortion theory for causal video coding: Characterization, computation algorithm, and comparison,” *IEEE Trans. Inf. Theory*, vol. 57, no. 8, pp. 5258–5280, 2011.

[4] ——, “On the information theoretic performance comparison of causal video coding and predictive video coding,” *IEEE Trans. Inf. Theory*, vol. 60, no. 3, pp. 1428–1446, Mar. 2014.

[5] N. Ma and P. Ishwar, “Erratum to “on delayed sequential coding of correlated sources”,” *IEEE Trans. Inf. Theory*, vol. 58, no. 6, p. 4074, June 2012.

[6] M. Torbatian and E.-H. Yang, “Causal coding of multiple jointly Gaussian sources,” in *Proc. Annual Allerton Conf. on Comm., Control, and Comput.*, Monticello, IL, USA, Oct. 2012, pp. 2060–2067.

[7] M. Eslamifar, “On causal video coding with possible loss of the first encoded frame,” Master’s thesis, University of Waterloo, Waterloo, Ontario, Canada, 2013.

[8] L. Song, J. Chen, J. Wang, and T. Liu, “Gaussian robust sequential and predictive coding,” *IEEE Trans. Inf. Theory*, vol. 59, no. 6, pp. 3635–3652, June 2013.

[9] H. C. Huang, W. H. Peng, and T. Chiang, “Advances in the scalable amendment of H.264/AVC,” *IEEE Comm. Magazine*, vol. 45, no. 1, pp. 68–76, Jan. 2007.

[10] Y.-Z. Huang, Y. Kochman, and G. W. Wornell, “Causal transmission of colored source frames over packet erasure channel,” in *Proc. Data Comp. Conf. (DCC)*, Snowbird, UT, USA, Mar. 2010, pp. 129–138.

[11] F. Etezadi, A. Khisti, and M. Trott, “Zero-delay sequential transmission of Markov sources over burst erasure channels,” *IEEE Trans. Inf. Theory*, vol. 60, no. 8, pp. 4584–4613, Aug. 2014.

[12] F. Etezadi, A. Khisti, and J. Chen, “A truncated prediction framework for streaming over erasure channels,” *IEEE Trans. Inf. Theory*, Submitted Jul. 2014. Revised Oct. 2016.

[13] A. H. Kaspi, “Rate–distortion when side-information may be present at the decoder,” *IEEE Trans. Inf. Theory*, vol. 40, no. 6, pp. 2031–2034, Nov. 1994.

[14] E. Perron, S. Diggavi, and I. E. Telatar, “On the role of encoder side-information in source coding for multiple decoders,” in *Proc. IEEE Int. Symp. on Inf. Theory (ISIT)*, Seattle, WA, USA, July 2006, pp. 351–355.

[15] C. Heegard and T. Berger, “Rate–distortion when side information may be absent,” *IEEE Trans. Inf. Theory*, vol. 31, pp. 727–734, Nov. 1985.

[16] A. Khina and U. Erez, “Source coding with composite side information at the decoder,” in *Proc. IEEE Conv. Electrical and Electron. Engineers in Israel (IEEEI)*, Eilat, Israel, Nov. 2012.

[17] T. M. Cover and J. A. Thomas, *Elements of Information Theory, Second Edition*. New York: Wiley, 2006.

[18] A. D. Wyner and J. Ziv, “The rate–distortion function for source coding with side information at the decoder,” *IEEE Trans. Inf. Theory*, vol. 22, pp. 1–10, Jan. 1976.

[19] A. D. Wyner, “The rate–distortion function for source coding with side information at the decoder — II: General sources,” *Information and Control*, vol. 38, pp. 60–80, 1978.

[20] R. Zamir and U. Erez, “A Gaussian input is not too bad,” *IEEE Trans. Inf. Theory*, vol. 50, no. 6, pp. 1362–1367, Jun. 2004.

[21] T. Philosof and R. Zamir, “The cost of uncorrelation and noncooperation in MIMO channels,” *IEEE Trans. Inf. Theory*, vol. 53, no. 11, pp. 3904–3920, Nov. 2007.