Source Coding With Delayed Side Information

Osvaldo Simeone
CWCSPR, ECE Dept.
New Jersey Institute of Technology
Newark, NJ, 07102, USA
Email: osvaldo.simeone@njit.edu

Haim H. Permuter
ECE Dept.
Ben-Gurion University of the Negev
Beer-Sheva 84105, Israel
Email: haimp@bgu.ac.il

***Abstract*—For memoryless sources, delayed side information at the decoder does not improve the rate-distortion function. However, this is not the case for more general sources with memory, as demonstrated by a number of works focusing on the special case of (delayed) feedforward. In this paper, a setting is studied in which the side information is delayed and the encoder is informed about the side information sequence. Assuming a hidden Markov model for the sources, at first, a single-letter characterization is given for the set-up where the side information delay is arbitrary and known at the encoder, and the reconstruction at the destination is required to be (near) lossless. Then, with delay equal to zero or one source symbol, a single-letter characterization is given of the rate-distortion function for the case where side information may be delayed or not, unbeknownst to the encoder. Finally, an example for a binary source is provided.**

I. INTRODUCTION

Consider a sensor network in which a sensor measures a certain physical quantity Y_i over time $i = 1, 2, ...n$. The aim of the sensor is communicating a processed version $X^n = (X_1, ..., X_n)$ of the measured sequence $Y^n = (Y_1, ..., Y_n)$ to a receiver. As an example, each element X_i could be obtained by quantizing Y_i, for $i = 1, 2, ...n$. To this end, the sensor communicates a message M of nR bits to the receiver, based on the observation of X^n and Y^n (R is the message rate in bits per source symbol). The receiver is endowed with sensing capabilities and hence it can measure the physical quantity Y^n as well. However, due to the fact that the receiver is located further away from the physical source, such measure may come with a delay of d symbols. In other words, when estimating X_i, the receiver has available not only the message M received from the sensor, but also the sequence $Y^{i-d} = (Y_1, ..., Y^{i-d})$, so that the estimate Z_i is a function of M and Y^{i-d}. Delay d may or may not be known at the sensor[1]. The situation described above can be illustrated schematically as in Fig. 1 and in Fig. 2, where Fig. 1 models the case where the delay d is known to the sensor (i.e., the encoder), while 2 accounts for a setting where the side information at the decoder, unbeknownst to the encoder, *may* be delayed by d or *not* delayed.

[1]In order to ensure that the sensor can produce the message M based on the entire sequences X^n and Y^n, as in the example at hand, the delay at the receiver corresponding to the given decoding rule $(X_i(M, Y^{i-d}))$ can be seen to be more precisely $n + d$ and not d (see, e.g., [1]). Nevertheless, we will refer to the delay at the receiver as d for simplicity.

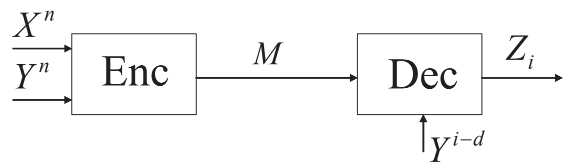

Figure 1. Lossy source coding with delayed side information at the decoder. The side information is fully available at the encoder.

Prior work: If sequences X^n and Y^n are *memoryless,* from available results [2][3], it can be inferred that: (*i*) for zero delay, i.e., $d = 0$, the performance of the systems in Fig. 1-2 would remain unchanged even if the decoder(s) had access to non-causal side information, in which case the decision about Z_{ji}, $j = 1, 2$, at each time i, could be based on the entire sequence Y^n, rather than only Y^i; and (*ii*) for strictly positive delay $d > 0$, delayed side information does not improve performance. However, these conclusions do not generally hold if the sources have memory.

For *sources with memory*, a number of works have focused on the scenario of Fig. 1 where $X_i = Y_i$, which entails that the decoder observes sequence X^n itself with a delay of d symbols. This setting is typically referred to as *source coding with feedforward*, as introduced in [5]. Reference [1] derives the rate-distortion function for this problem (i.e., Fig. 1 with $X_i = Y_i$) for ergodic and stationary sources in terms of multi-letter mutual informations[2]. This function is explicitly evaluated for some special cases in [2][4] (see also [6]), while an algorithm for its numerical calculation is also proposed in [4]. The more general case of Fig. 1 with $X_i \neq Y_i$ is studied in [7] assuming stationary and ergodic sources X^n and Y^n. The rate-distortion function is expressed in terms of multi-letter mutual informations, and no specific examples are provided for which the function is explicitly computable. Moreover, extensions of the characterization of achievable rate-distortion trade-offs to the setting of Fig. 2 for sources with memory has not, to the best of the authors' knowledge, been studied. We finally remark that for more complex networks than the ones studied here, strictly delayed side information may be useful

[2]Extensions are also given for arbitrary sources using information-spectrum methods.

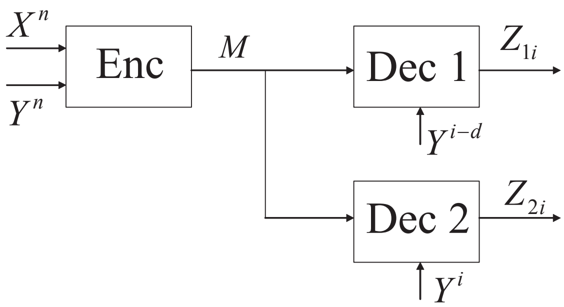

Figure 2. Lossy source coding where side information at the decoder may be delayed. The side information is fully available at the encoder.

also in the presence of memoryless sources. This is illustrated in [9] for a multiple description problem with feedforward.

Contributions: In this work, we assume that the source Y^n is a Markov chain, and X^n is such that X_i is obtained by passing Y_i through a memoryless channel $q(x|y)$ for $i = 1, ..., n$, i.e., X^n corresponds to a hidden Markov model. Note that the latter may model a symbol-by-symbol processing of source Y^n as per the initial example. We derive a single-letter characterization of the minimal rate (bits/source symbol) required for (near) lossless compression in the scenario of Fig. 1 for any delay $d \geq 0$ (Sec. III). Achievability is based on a novel scheme that consists of simple multiplexing/demultiplexing operations along with standard entropy coding techniques. Furthermore, we derive a single-letter characterization of the minimal rate (bits/source symbol) required for lossy compression in the scenarios of Fig. 1 and Fig. 2 for delays $d = 0$ and $d = 1$ (Sec. IV). Finally, we study the specific example of a binary-alphabet source with Hamming distortion (Sec. V).

Notation: For a, b integer with $a \geq b$, we define $[a, b]$ as the interval $[a, a+1, ..., b]$ and $x_a^b = (x_a, ..., x_b)$; if instead $a < b$ we set $[a, b] = \emptyset$ and $x_a^b = \emptyset$. We will also write x_1^b for x^b for simplicity of notation. Given a sequence $x^n = [x_1, ..., x_n]$ and a set $\mathcal{I} = \{i_1, ..., i_{|\mathcal{I}|}\} \subseteq [1, n]$, we define sequence $x^{\mathcal{I}}$ as $x^{\mathcal{I}} = [x_{i_1}, x_{i_2}, ..., x_{i_{|\mathcal{I}|}}]$ where $i_1 \leq ... \leq i_{|\mathcal{I}|}$.

II. System Model

We present the system model for the scenario of Fig. 2, as the scenarios of Fig. 1 follows as a special case. The random process $Y_i \in \mathcal{Y}$, $i \in \{..., -1, 0, 1, ...\}$, measured at the encoder, and, possibly with delay, at the decoders, is a stationary and ergodic Markov chain with transition probability $\Pr[Y_i = a|Y_{i-1} = b] = w_1(a|b)$. We define the probability $\Pr[Y_i = a] \triangleq \pi(a)$ and also the k-step transition probability $\Pr[Y_i = a_i|Y_{i-k} = b] \triangleq w_k(a|b)$, which are both independent of i by stationarity of Y_i. We also set, for notational convenience, $w_0(a|b) = \pi(a)$. Sequence $Y^n = (Y_1, ..., Y_n)$ is thus distributed as $p(y^n) = \pi(y_1) \prod_{i=2}^n w_1(y_i|y^{i-1})$ for any integer $n > 0$. The random process $X_i \in \mathcal{X}$, $i \in \{..., -1, 0, 1, ...\}$, measured only at the encoder, is such that vector $X^n = (X_1, ..., X_n) \in \mathcal{X}^n$, for any integer $n > 0$, is jointly distributed with Y^n as

$$\begin{aligned} p(x^n, y^n) &= \pi(y_1) q(x_1|y_1) \prod_{i=2}^n p(x_i, y_i|x^{i-1}, y^{i-1}) \\ &= \pi(y_1) q(x_1|y_1) \prod_{i=2}^n w_1(y_i|y_{i-1}) q(x_i|y_i). \end{aligned} \quad (1)$$

In other words, process $X_i \in \mathcal{X}$, $i \in \{..., -1, 0, 1, ...\}$ corresponds to a *hidden Markov model* with underlying Markov process given by Y^n.

An (d, n, R, D_1, D_2) code, with delay $d \geq 0$, is defined by: (*i*) An encoder function

$$\mathrm{f}\colon (\mathcal{X}^n \times \mathcal{Y}^n) \rightarrow [1, 2^{nR}], \quad (2)$$

which maps sequences X^n and Y^n into message $M \in [1, 2^{nR}]$; (*ii*) a sequence of decoding functions for decoder 1

$$\mathrm{g}_{1i}\colon [1, 2^{nR}] \times \mathcal{Y}^{i-d} \rightarrow \mathcal{Z}_1, \quad (3)$$

for $i \in [1, n]$, which, at each time i, map message M, or rate R [bits/source symbol], and the *delayed* side information Y^{i-d} into the estimate $Z_{1i} \in \mathcal{Z}_1$; (*iii*) a sequence of decoding function for decoder 2

$$\mathrm{g}_{2i}\colon [1, 2^{nR}] \times \mathcal{Y}^{i} \rightarrow \mathcal{Z}_2 \quad (4)$$

for $i \in [1, n]$, which, at each time i, map messages M and the *non-delayed* side information Y^i into the estimate $Z_{2i} \in \mathcal{Z}_2$. Encoding/decoding functions (2)-(4) must satisfy the distortion constraints

$$\frac{1}{n} \sum_{i=1}^n \mathrm{E}[\mathrm{d}_j(X_i, Y_i, Z_{ji})] \leq D_j, \text{ for } j = 1, 2, \quad (5)$$

where the distortion metrics $d_j(x, y, z_j)$: $\mathcal{X} \times \mathcal{Y} \times \mathcal{Z}_j \rightarrow [0, d_{\max}]$ are such that $0 \leq \mathrm{d}_j(x, y, z_j) \leq \mathrm{d}_{\max} < \infty$ for all $(x, y, z) \in \mathcal{X} \times \mathcal{Y} \times \mathcal{Z}_j$ for $j = 1, 2$. Note that these constraints are fairly general in that they allow to impose not only requirements on the lossy reconstruction of X_i or Y_i (obtained by setting $\mathrm{d}_j(x, y, z_j)$ independent of y or x, respectively), but also on some function of both X_i and Y_i (by setting $\mathrm{d}_j(x, y, z_j)$ to be dependent on such function of (x, y)).

Given a delay $d \geq 0$, for a distortion pair (D_1, D_2), we say that rate R is achievable if, for every $\epsilon > 0$ and sufficiently large n, there exists a $(d, n, R, D_1 + \epsilon, D_2 + \epsilon)$ code. We refer to the infimum of all achievable rates for a given distortion pair (D_1, D_2) and delay d as the *rate-distortion function* $R_d(D_1, D_2)$. For the setting of Fig. 1, we similarly define the rate-distortion function $R_d(D_1)$.

III. Lossless Source Coding with Delayed Side Information

Here we consider the setting of Fig. 1 and we characterize the rate-distortion function $R_d(D_1)$ for any delay $d \geq 0$ under the Hamming distortion metric (i.e., $\mathrm{d}_1(x, y, z_1) = 1(x \neq z_1)$, where $1(a) = 1$ if a is true and $1(a) = 0$ otherwise) for $D_1 = 0$. In other words, we impose that the sequence

X^n be recovered with vanishingly small average symbol error probability as $n \to \infty$ at the decoder. We refer to this scenario as (near) lossless. We have the following characterization of $R_d(0)$.

Proposition 1. *For any delay $d \geq 0$, the rate-distortion function for the set-up in Fig. 1 under Hamming distortion is given at $D_1 = 0$ by*

$$R_d(0) = H(X_{d+1}|X_2^d, Y_1), \tag{6}$$

where the conditional entropy is calculated with respect to the distribution

$$p(y_1, x_1) = \pi(y_1) q(x_1|y_1) \text{ for } d = 0, \tag{7}$$

$$\text{and } p(y_1, x_2, ..., x_{d+1}) = \pi(y_1) \tag{8}$$

$$\cdot \sum_{\substack{y_i \in \mathcal{Y} \\ i \in [2, d+1]}} \prod_{i=2}^{d+1} w_1(y_i|y_{i-1}) q(x_i|y_i),$$

for $d \geq 1$.

The proof of achievability is sketched below. Details can be found in [10], along with the proof of the converse.

Remark 2. Proposition 1 provides a "single-letter" characterization of $R_d(0)$ for the setting of Fig. 1, since it only involves a finite number of variables. This contrasts with the general characterization for stationary ergodic processes of $R_d(D)$ (in the general lossy case $D \geq 0$) given in [7], which is a "multi-letter" expression, whose computation can generally only attempted numerically using approaches such as the ones proposed in [4]. Note that a multi-letter expression is also given in [2] to characterize $R_d(D)$ for *negative* delays $d < 0$.

Remark 3. By setting $d = 0$ in (6) we obtain $R_0(0) = H(X_1|Y_1)$. This result generalizes [2, Remark 3, p. 5227] from i.i.d. sources (X^n, Y^n) to the hidden Markov model (1) considered here. Note that, for $d = 1$, we instead obtain $R_1(0) = H(X_2|Y_1)$. As another notable special case, if side information is absent, or equivalently if $d \to \infty$, in accordance to well-known results, we obtain that $R_\infty(0)$ equals the entropy rate $H(\mathcal{X})$.

Remark 4. Is *delayed* side information useful (when known also at the encoder)? That this is generally the case follows from the inequality $R_d(0) = H(X_{d+1}|X_2^d, Y_1) \leq R_\infty(0) = H(\mathcal{X})$, since $R_\infty(0)$ is the required rate without side information. However, the inequality above may not be strict, and thus side information may not be useful. This is the case for instance if X_i is an i.i.d. process or in the setting of source coding with feedforward [5], [1], i.e., $X_i = Y_i$, with a Markov source X^n. We will see below that the conclusion that feedforward is not useful for Markov sources need not hold for lossy compression (i.e., for $D_1 > 0$).

A. Proof of Achievability for Proposition 1

Proof: (Achievability) Here we propose a coding scheme that achieves rate (6). The basic idea is a non-trivial extension of the approach discussed in [2, Remark 3, p. 5227] and is described as follows. A block diagram is shown in Fig. 3 for encoder (Fig. 3-(a)) and decoder (Fig. 3-(b)). We first describe the *encoder,* which is illustrated in Fig. 3-(a). To encode sequences $(x^n, y^n) \in (\mathcal{X}^n \times \mathcal{Y}^n)$, we first partition the interval $[1, n]$ into $|\mathcal{X}|^{d-1}|\mathcal{Y}|$ subintervals, which we denote as $\mathcal{I}(\tilde{x}^{d-1}, \tilde{y}) \subseteq [1, n]$, for all $\tilde{x}^{d-1} \in \mathcal{X}^{d-1}$ and $\tilde{y} \in \mathcal{Y}$. Every such subinterval $\mathcal{I}(\tilde{x}^{d-1}, \tilde{y})$ is defined as

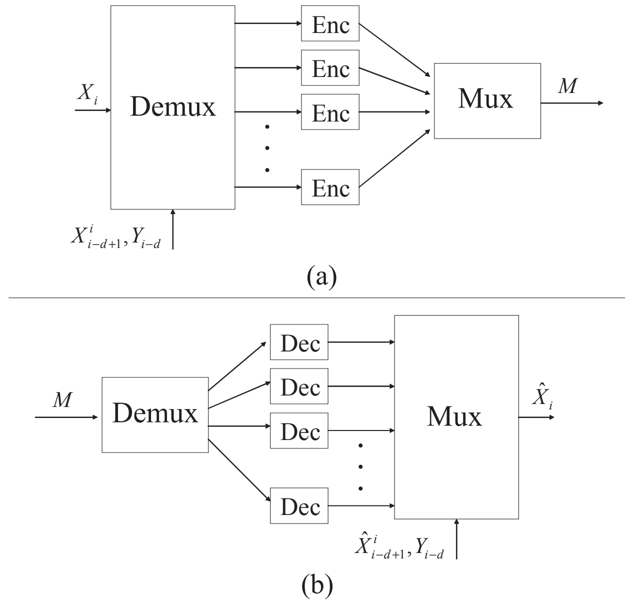

Figure 3. A block diagram for encoder (a) and decoder (b) used in the proof of achievability of Proposition 1.

$$\mathcal{I}(\tilde{x}^{d-1}, \tilde{y}) = \{i\colon i \in [1, n] \text{ and } y_{i-d} = \tilde{y},\ x_{i-d+1}^{i-1} = \tilde{x}^{d-1}\}. \tag{9}$$

In words, the subinterval $\mathcal{I}(\tilde{x}^{d-1}, \tilde{y})$ contains all symbol indices i such that the corresponding delayed side information available at the decoder is $y_{i-d} = \tilde{y}$ and the previous $d-1$ samples in x^n are $x_{i-d+1}^{i-1} = \tilde{x}^{d-1}$. For the out-of-range indices $i \in [-d+1, 0]$, one can assume arbitrary values for $x_i \in \mathcal{X}$ and $y_i \in \mathcal{Y}$, which are also shared with the decoder once and for all. Note that $\bigcup_{\tilde{x}^{d-1} \in \mathcal{X}^{d-1},\ \tilde{y} \in \mathcal{Y}} \mathcal{I}(\tilde{x}^{d-1}, \tilde{y}) = [1, n]$. Fig. 4 illustrates the definitions at hand for $d = 2$.

As a result of the partition described above, the encoder "demultiplexes" sequence x^n into $|\mathcal{X}|^{d-1}|\mathcal{Y}|$ sequences $x^{\mathcal{I}(\tilde{x}^{d-1}, \tilde{y})}$, one for each tuple $(\tilde{x}^{d-1}, \tilde{y}) \in \mathcal{X}^{d-1} \times \mathcal{Y}$. This demultiplexing operation, which is controlled by the previous values of source and side information, is performed in Fig. 3-(a) by the block labelled as "Demux", and an example of its operation is shown in Fig. 4. By the ergodicity of process X_i and Y_i, for every $\epsilon > 0$ and all sufficiently large n, the length of any sequence $x^{\mathcal{I}(\tilde{x}^{d-1}, \tilde{y}_1)}$ is guaranteed to be less than $n p_{Y_1 X_2, ..., X_d}(\tilde{y}, \tilde{x}^{d-1}) + \epsilon$ symbols with abitrarily large probability.

The entropy encoder can be implemented in different ways, e.g., using typicality or Huffman coding. Here we consider a typicality-based encoder. Note that the entries X_i of each sequence $X^{\mathcal{I}(\tilde{x}^{d-1}, \tilde{y})}$ are i.i.d. with distribution

$\tilde{x}$	$\tilde{y}$	$I(\tilde{x},\tilde{y})$	$x^{I(\tilde{x},\tilde{y})}$
0	0	{1, 2, 3, 9}	[0,0,1,1]
0	1	{5,7}	[1,1]
1	0	{6,10}	[0,1]
1	1	{4.8}	[0,0]

Figure 4. An example that illustrates the operations of the "Demux" block of the encoder used for the achievability proof of Proposition 1, as shown in Fig. 3, for sequences $x^n = (0,0,1,0,1,0,1,0,1,1)$ and $y^n = (0,1,1,0,1,1,0,0,1,1)$, $n = 10$ and $d = 2$ (symbols corresponding to out-of-range indices are set to zero).

$p_{X_{d+1}|Y_1X_2,...,X_d}(\cdot|\tilde{y},\tilde{x}^{d-1})$, since conditioning on the event $\{y_{i-d} = \tilde{y},\ x_{i-d+1}^{i-1} = \tilde{x}^{d-1}\}$ makes the random variables X_i independent. Therefore, a rate in bits per source symbol of $H(X_{d+1}|X_2^d = \tilde{x}^{d-1}, Y_1 = \tilde{y}) + \epsilon$ is sufficient for the entropy encoder to label all ϵ-typical sequences.

We now describe the *decoder,* which is illustrated in Fig. 3-(b). By undoing the multiplexing operation just described, the decoder, from the message M, can recover the individual sequences $x^{\mathcal{I}(\tilde{x}^{d-1},\tilde{y})}$ through a simple demultiplexing operation for all $\tilde{x}^{d-1} \in \mathcal{X}^{d-1}$ and $\tilde{x}^{d-1} \in \mathcal{X}^{d-1}$. This operation is represented by block "Demux" in Fig. 3-(b). However, while the individual sequences $x^{\mathcal{I}(\tilde{x}^{d-1},\tilde{y})}$ can be recovered through the discussed demultiplexing operation, this does not imply that the decoder is also able to reorder the symbols in the sequences so as to obtain the original sequence x^n. However, note that at time i, the decoder knows Y_{i-d} and the previously decoded X^{i-1} and can thus identify the subinterval $\mathcal{I}(\tilde{x}^{d-1},\tilde{y})$ to which the current symbol X_i belongs. This symbol can be then immediately read as the next yet-to-be-read symbol from the corresponding sequence $x^{\mathcal{I}(\tilde{x}^{d-1},\tilde{y})}$. Note that for the first d symbols, the decoder uses the values for x_i and y_i at the out-of-range indices i that were agreed upon with the encoder (see above). A more detailed description, including the analysis of the impact of errors, can be found in [10]. ■

IV. Lossy Source Coding Where Side Information May Be Delayed

In this section, we consider the general problem of lossy compression for the set-up of Fig. 2, and we obtain an achievable rate $R_d^{(a)}(D_1,D_2) \geq R_d(D_1,D_2)$ for all delays $d \geq 0$ and prove that such rate equals the rate-distortion function, i.e., $R_d^{(a)}(D_1,D_2) = R_d(D_1,D_2)$, for $d = 0$ and $d = 1$.

Proposition 5. *For any delay $d \geq 0$ and distortion pair (D_1,D_2), the following rate is achievable for the setting of Fig. 2*

$$R_d^{(a)}(D_1,D_2) = \min I(XY;Z_1|Y_d) + I(X;Z_2|YY_dZ_1) \tag{10}$$

$$= \min I(Y;Z_1|Y_d) + I(X;Z_1Z_2|YY_d), \tag{11}$$

with mutual informations evaluated with respect to the joint distribution

$$p(x,y,y_d,z_1,z_2) = \pi(y_d)w_d(y|y_d)q(x|y)p(z_1,z_2|x,y,y_d), \tag{12}$$

and where minimization is done over all conditional distributions $p(z_1,z_2|x,y,y_d)$ such that

$$\mathrm{E}[d_j(X,Y,Z_j)] \leq D_j,\ \textit{for}\ j = 1,2. \tag{13}$$

Moreover, rate (10)-(11) is the rate-distortion function, i.e., $R_d^{(a)}(D_1,D_2) = R_d(D_1,D_2)$, for $d = 0$ and $d = 1$.

Remark 6. Rate (10) can be easily interpreted in terms of achievability. To this end, we remark that variable Y_d plays the role of the delayed side information Y^{i-d} at decoder 1. The coding scheme achieving rate (10) operates in two successive phases. In the first phase, the encoder encodes the reconstruction sequence Z_1^n for decoder 1. Since decoder 1 has available delayed side information, using a strategy similar to the one discussed in Sec. III-A, this operation requires $I(XY;Z_1|Y_d)$ bits per source sample. Note that decoder 2 is able to recover Z_1^n as well, since decoder 2 has available side information Y^i, and thus also the delayed side information Y^{i-d}. In the second phase, the reconstruction sequence Z_2^n for decoder 2 is encoded. Given the side information available at decoder 2, this operation requires rate $I(X;Z_2|YY_dZ_1)$, using again an approach similar to the one discussed in Sec. III-A. Details can be found in [10], along with the converse proof.

Remark 7. For memoryless sources X^n and Y^n, by comparison with the results in [3], it can be concluded that *delayed side information is not useful for memoryless sources*. This conclusion generalizes the result of [2], which applies for the setting of Fig. 1 in the special case of feedforward (i.e., $X_i = Y_i$).

Note that, by setting $D_2 = d_{\max}$ in $R_d^{(a)}(D_1,D_2)$, we obtain an achievable rate $R_d^{(a)}(D_1)$ for the setting of Fig. 1 (see [10] for details).

V. Example: Binary Hidden Markov Model

In this section, we assume that Y_i is a binary Markov chain with symmetric transition probabilities $w_1(1|0) = w_1(0|1) \triangleq \varepsilon$ and we assume that $X_i = Y_i \oplus N_i$, with "$\oplus$" being the modulo-2 sum and N_i being i.i.d. binary variables, independent of Y^n, with $p_{N_i}(1) \triangleq q$, $q \leq 1/2$. Note that we have the k-step transition probabilities $w_k(1|0) = w_k(0|1) \triangleq \varepsilon^{(k)}$, which can be obtained recursively as $\varepsilon^{(1)} = \varepsilon$ and $\varepsilon^{(k)} = 2\varepsilon^{(k-1)}(1-\varepsilon^{(k-1)})$ for $k \geq 2$. We adopt the Hamming distortion $d_1(x,z_1) = x \oplus z_1$.

We start by showing in Fig. 5 the rate $R_d(0)$ obtained from Proposition 1 corresponding to zero distortion ($D_1 = 0$) versus the delay d for different values of ε and for $q = 0.1$. For $d = 0$, we have $R_0(0) = H(X_1|Y_1) = H_b(q) = 0.589$, irrespective of the value of ε, where we have defined the binary entropy function $H_b(a) = -a\log_2 a - (1-a)\log_2(1-a)$. Instead, for d increasingly large, the rate $R_d(0)$ tends to the entropy rate

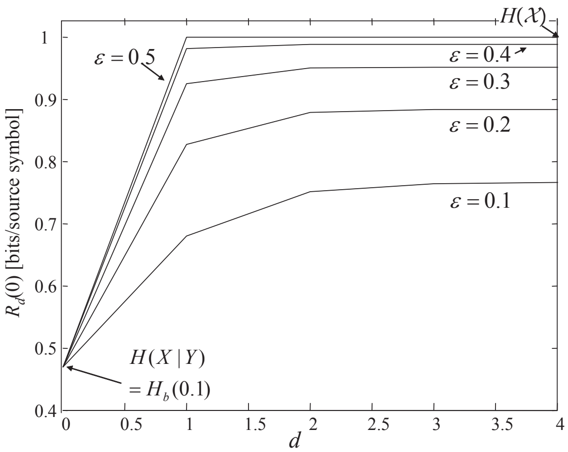

Figure 5. Minimum required rate $R_d(0)$ for lossless reconstruction for the set-up of Fig. 1 with binary sources versus delay d ($q = 0.1$).

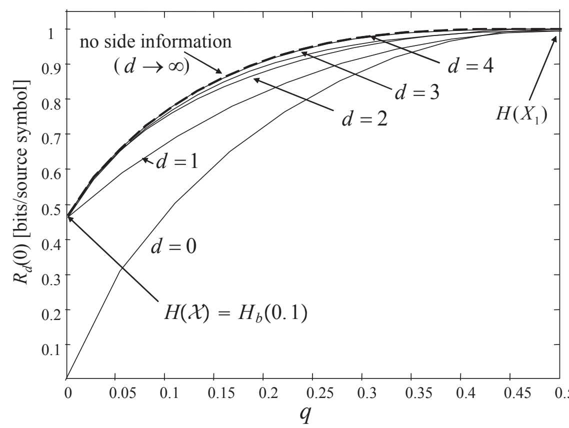

Figure 6. Minimum required rate $R_d(0)$ for lossless reconstruction for the set-up of Fig. 1 with binary sources versus parameter q ($\varepsilon = 0.1$).

$R_\infty(0) = H(\mathcal{X})$. Note that a larger memory, i.e., a smaller ε, leads to smaller required rate $R_d(0)$ for all values of d.

Fig. 6 shows the rate $R_d(0)$ for $\varepsilon = 0.1$ versus q for different values of delay d. For reference, we also show the performance with no side information, i.e., $R_\infty(0) = H(\mathcal{X})$. For $q = 1/2$, the source X^n is i.i.d. and delayed side information is useless in the sense that $R_d(0) = R_\infty(0) = H(X_1) = 1$ (Remark 4). Moreover, for $q = 0$, we have $X_i = Y_i$, so that X_i is a Markov chain and the problem becomes one of lossless source coding with feedforward. From Remark 4, we know that delayed side information is useless also in this case, as $R_d(0) = R_\infty(0) = H(\mathcal{X}) = H_b(\varepsilon) = 0.469$.[3] For intermediate values of q, side information is generally useful, unless the delay d is too large.

Finally, we evaluate the achievable rate of Proposition 2 for a general non-zero distortion D_1 (see details in [10]), obtaining

$$R_d^{(a)}(D_1) = H_b(\varepsilon^{(d)} * q) - H_b(D_1) \tag{14}$$

for $0 \leq D_1 \leq \min\{\varepsilon^{(d)} * q, 1 - \varepsilon^{(d)} * q\}$ and $R_d^{(a)}(D_1) = 0$ otherwise, where $p * q \triangleq p(1-q) + (1-p)q$. This result with $q = 0$ (i.e., with feedforward) and $d = 1$ recovers the calculation in [5, Example 2] (see also [4]). We remark that the rate-distortion function of a Markov source X^n without feedforward, i.e., $R_\infty(D_1)$, is equal to $H_b(\varepsilon) - H(D_1)$ only for D_1 smaller than a critical value, but is otherwise larger [8]. This demonstrates that feedforward, unlike in the lossless setting discussed above, can be useful in the lossy case for distortion levels D_1 sufficiently large.

VI. Concluding Remarks

A general information-theoretic characterization of the trade-off between rate and distortion for the problem of compressing information sources in the presence of delayed side information can be generally given in terms of multi-letter expressions, as done in [7]. In this work, we have instead focused on a specific class of sources, which evolve according to hidden Markov models, and derived single-letter characterizations of the rate-distortion trade-off. Such characterizations are established based on simple achievable scheme that are based on standard "off-the-shelf" compression techniques. Moreover, we have extended the analysis to a more general set-up in which side information may or may not be delayed.

[3] We use the conventional definition of the binary entropy as $H(x) \triangleq -x\log_2 x - (1-x)\log_2(1-x)$.

VII. Acknowledgement

The work of O. Simeone was supported in part by the U.S. National Science Foundation under Grant No. 0914899. H. H. Permuter was supported in part by the Marie Curie Reintegration fellowship.

References

[1] R. Venkataramanan and S. S. Pradhan, "Source coding with feed-forward: Rate-distortion theorems and error exponents for a general source," *IEEE Trans. Inform. Theory*, vol. 53, no. 6, pp. 2154-2179, Jun. 2007.

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

[3] A. H. Kaspi, "Rate-distortion function when side-information may be present at the decoder," *IEEE Trans. Inform. Theory*, vol. 40, no. 6, pp. 2031-2034, Nov 1994.

[4] I. Naiss and H. Permuter, "Computable bounds for rate distortion with feed-forward for stationary and ergodic sources," arXiv:1106.0895v1.

[5] T. Weissman and N. Merhav, "On competitive prediction and its relation to rate-distortion theory," *IEEE Trans. Inform. Theory*, vol. 49, no. 12, pp. 3185- 3194, Dec. 2003.

[6] R. Venkataramanan and S. S. Pradhan, "On computing the feedback capacity of channels and the feed-forward rate-distortion function of sources," *IEEE Trans. Commun*., vol. 58, no. 7, pp. 1889–1896, July 2010.

[7] R. Venkataramanan and S. S. Pradhan, "Directed information for communication problems with side-information and feedback/feed-forward," in *Proc. of the 43rd Annual Allerton Conference*, Monticello, IL, 2005.

[8] R. Gray, "Information rates of autoregressive processes," *IEEE Trans. Inform. Theory*, vol. 16, no. 4, pp. 412- 421, Jul. 1970.

[9] R. Venkataramanan and S. S. Pradhan, "Achievable rates for multiple descriptions with feed-forward," Information Theory, *IEEE Trans. Inform. Theory*, vol. 57, no. 4, pp. 2270-2277, Apr. 2011.

[10] O. Simeone and H. H. Permuter, "Source coding when the side information may be delayed," arXiv:1109.1293.