Source Coding With Limited-Look-Ahead Side Information at the Decoder

Tsachy Weissman, *Member, IEEE*, and Abbas El Gamal, *Fellow, IEEE*

***Abstract*—We characterize the rate distortion function for the source coding with decoder side information setting when the ith reconstruction symbol is allowed to depend only on the first $i+\ell$ side information symbols, for some finite look-ahead ℓ, in addition to the index from the encoder. For the case of causal side information, i.e., $\ell = 0$, we find that the penalty of causality is the omission of the subtracted mutual information term in the Wyner–Ziv rate distortion function. For $\ell > 0$, we derive a computable "infinite-letter" expression for the rate distortion function. When specialized to the near-lossless case, our results characterize the best achievable rate for the Slepian–Wolf source coding problem with finite side information looka-head, and have some surprising implications. We find that side information is useless for any fixed ℓ when the joint probability mass function (PMF) of the source and side information satisfies the positivity condition $P(x,y) > 0$ for all (x,y). More generally, the optimal rate depends on the distribution of the pair X,Y only through the distribution of X and the bipartite graph whose edges represent the pairs x,y for which $P(x,y) > 0$. On the other hand, if side information look-ahead is allowed to grow faster than logarithmic in the block length, then $H(X \mid Y)$ is achievable. Finally, we apply our approach to derive a computable expression for channel capacity when state information is available at the encoder with limited look-ahead.**

***Index Terms*—Causal source codes, delay-constrained coding, Gel'fand–Pinsker channel, rate distortion function, Slepian–Wolf coding, Wyner–Ziv coding.**

I. Introduction

CONSIDER the problem of source coding with side information at the decoder, as depicted in Fig. 1. The source and side information are generated as independent drawings (X_i, Y_i) of the pair $(X,Y) \sim P(x,y)$. The encoder maps the sequence $X^n = (X_1, \ldots, X_n)$ into an index $T \in \{1 \ldots, \lfloor 2^{nR} \rfloor\}$, while the decoder is allowed to depend on the side-information sequence Y^n in addition to the index T to produce the estimate $\hat{X}^n$. Note that Y^n can be considered a noisy observation of X^n and, hence, the decoder can be thought of as a *denoiser* that is allowed to base its estimate $\hat{X}^n$ on information conveyed to it by the encoder, in addition to its noisy observation of the source. The problem is to find the *rate distortion function* $R(D)$, which is the smallest rate R needed to achieve expected per-symbol distortion D with respect to some distortion measure ρ.

In their seminal 1976 paper [42], Wyner and Ziv solved this problem when $\hat{X}^n$ can depend on Y^n in an arbitrary manner. In this paper, we characterize the rate distortion function for the Wyner–Ziv setting when the decoder is constrained to look at no more than $\ell \geq 0$ future side-information symbols to produce the ith reconstruction symbol, i.e., $\hat{X}_i(T, Y^{i+\ell})$.

Since the scenario where the encoder has access to the entire source sequence X^n while the decoder is restricted in its side information look-ahead may seem contrived, we first provide motivation for considering it. We first argue that it is well motivated by the more practical problem of delay-constrained sequential source coding with side information at the decoder. There are various ways of formalizing the notion of sequential zero-delay and delay-constrained source coding, e.g., [10], [11], [38], [34], [20], [24], [33], [35], [22]. Perhaps the most general formulation, encompassing all of these as special cases and accommodating the availability of side information at the decoder is the following: A source code is specified by an encoder and a decoder. The encoder is a sequence of mappings $\{E_i\}$, where E_i produces a code symbol U_i in some finite alphabet $\mathcal{U}_i$ based on observation of the source with some look-ahead s, i.e., $U_i = E_i(X^{i+s})$. The decoder is a sequence of mappings $\{D_i\}$, where D_i produces the ith reconstruction symbol based on some look-ahead m in the code symbols and some look-ahead ℓ in the side-information symbols, i.e., $\hat{X}_i = D_i(U^{i+m}, Y^{i+\ell})$. The instantaneous rate of the code is $\log |\mathcal{U}_i|$, so that the overall rate in encoding the first n source symbols is $R = \frac{1}{n}\sum_{i=1}^{n} \log |\mathcal{U}_i|$. It is easy to see that any source code with this structure fits within the setting we consider. More precisely, the performance of any code with look-ahead parameters (s,m,ℓ) can be attained arbitrarily closely by a block code in the form of Fig. 1, with a reconstruction of the form $\hat{X}_i(T, Y^{i+\ell})$, for some T taking values in a set of size 2^{nR}. To see this concretely, given a sequential code with look-ahead parameters (s,m,ℓ), construct a block code that emulates it as follows: Given X^n, the encoder produces the code symbols $U_1, U_2, \ldots, U_{n-s}$ just as a sequential encoder would. It then losslessly describes the sequence $U_1, U_2, \ldots, U_{n-s}$ to the decoder using an index in the set $\{1, \ldots, 2^{nR}\}$, which can be done since

$$\sum_{i=1}^{n-s} \log |\mathcal{U}_i| \leq \sum_{i=1}^{n} \log |\mathcal{U}_i| \leq nR.$$

The decoder, knowing $U_1, U_2, \ldots, U_{n-s}$ and having access to the side-information sequence produces $\hat{X}^{n-s-m}$ just as the sequential decoder would. Thus, the resulting block code uses

Manuscript received December 7, 2005; revised June 20, 2006. This work was supported in part by the National Science Foundation under Grant CCR-0311633 and a CAREER Grant. The material in this paper was presented at the IEEE International Symposium on Information Theory, Seattle, WA, July 2006.

The authors are with the Department of Electrical Engineering, Stanford University, Stanford, CA 94305 USA (e-mail: tsachy@stanford.edu; abbas@stanford.edu).

Communicated by Y. Steinberg, Associate Editor for Shannon Theory.

Digital Object Identifier 10.1109/TIT.2006.885500

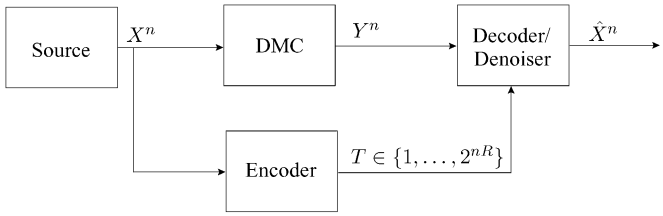

Fig. 1. Source coding with decoder side information.

the same rate for encoding the first n source symbols as the sequential code and incurs an overall cumulative distortion that exceeds that of the sequential code by no more than $(s+m) \times D_{\max}$. When normalized by the block length, this excess distortion is negligible in the (assumed) regime of s, m fixed and increasing n. Note that using a similar approach, one could alternatively construct a block code incurring exactly the same distortion on the first n source symbols as the sequential code with description rate that is higher than that of the sequential code by a diminishing amount. Consequently, our negative (converse) results immediately apply to all source coding scenarios that comply with this "delay-constrained sequential coding" formulation. On the positive side, it will be seen that, in some cases, optimum performance within the large family of schemes that we allow is attained by simple schemes from the more limited family of sequential codes.

Another motivation for our problem is denoising systems with side information. As mentioned, given the index from the encoder, the decoder can be viewed as a denoiser. The case $\ell = 0$ then corresponds to filtering (sequential denoising), while $\ell > 0$ to fixed-lag smoothing [8]. For example, if X_i denotes the location at time i of a moving target whose trajectory is known to an encoder, our problem can be viewed as characterizing the best performance of a (sequential) tracker of this target (based on its noisy trajectory) for a given rate at which the encoder can convey information about the target's (noiseless) trajectory to the tracker.

Our work is also motivated by an attempt to better understand the duality between channel coding and source coding. If the Wyner–Ziv problem [42] is the source coding analogue of the Gel'fand–Pinsker setting of channel coding with noncausal state information at the transmitter [12], [15], then the problem we consider here, for the case $l = 0$, is the source coding analogue of channel coding with causal state information at the transmitter, as considered by Shannon [31]. The relationship between the noncausal and causal versions of the source coding problem will be seen to have similarities to the relationship between the noncausal and causal channel coding problems. In the other direction, this duality will allow us to use our approach to characterize the rate distortion function for general $\ell > 0$ to obtain a computable expression for channel capacity when state information is available to the sender with a limited look-ahead.

Finally, our work complements the recent study in [36], which examined the effect of the introduction of side information into the causal source coding framework of Neuhoff and Gilbert [26] (cf. also [21]). As was pointed out, the Neuhoff–Gilbert notion of a causal source code is not well-matched to most scenarios involving side information at the decoder, and consequently, such scenarios were not considered in [36]. We believe that our setting with $\ell = 0$ is the closest in spirit to Neuhoff and Gilbert's causality notion in that the constraint is imposed only on the reconstruction, rather than directly on the delay introduced by the code.

Organization of Paper and Main Results

In Section II, we characterize the rate distortion function for source coding with causal side information, i.e., when $l = 0$. We find that the penalty of causality is the omission of the second (subtracted) mutual information term in the Wyner–Ziv rate distortion function. This implies that binning, which reduces the rate required to describe the auxiliary codeword in the direct proof of the Wyner–Ziv problem, is no longer viable when restricting the reconstruction to causal dependence on the side information. In Section III, we characterize the best achievable rate again for $\ell = 0$ when the reconstruction of the source is required to be perfect (with high probability). This characterization will be seen to have some surprising implications. For example, we find that side information is useless when the source and side information satisfy the positivity condition $P(x,y) > 0$ for all x, y. More generally, the best achievable rate will be seen to depend on the distribution of the pair X, Y only through the distribution of X and the bipartite graph whose edges represent the pairs x, y for which $P(x,y) > 0$.

Sections IV and V address the case of $0 < \ell < \infty$. Specifically, in Section IV, we obtain a characterization of the rate distortion function for any delay $\ell > 0$, which, although does not have a "single-letter" form, is computable in a sense that will be explained. Section V will give a computable characterization for the lossless case. Here too, the achievable rate in the general case depends on the distribution of the pair X, Y only through the distribution of X and the bipartite graph whose edges represent the pairs x, y for which $P(x,y) > 0$. Furthermore, this characterization will be shown to imply, under the positivity condition mentioned, that the side information is useless not only when $\ell = 0$, but for *any* value of $\ell < \infty$. This is in stark contrast to the case with no constraint on the look-ahead, where the conditional entropy $H(X \,|\, Y)$ is achievable [32]. Bridging these regimes, we will show that any rate R in the interval $(H(X \,|\, Y), H(X))$ is achievable when ℓ is allowed to grow with the block length as $C(R) \log n$, where $C(R)$ is an R-dependent constant. In particular, $H(X \,|\, Y)$ itself is achievable when ℓ_n increases faster than logarithmic. Thus, the Slepian–Wolf theorem applies when the decoder at time i is of the form $\hat{X}_i(T, Y^{i+\ell_n})$, where ℓ_n grows super-logarithmically.

In Section VI, we consider the channel coding dual to the source coding problem of Section IV. We obtain a computable characterization of the capacity of the memoryless channel in the presence of state information with limited look-ahead at the

encoder, which parallels the one obtained for the rate distortion problem in Section IV. We conclude in Section VII with a discussion of related open questions.

II. Rate Distortion Function With Causal Side Information

We assume throughout a source and side information that are generated as independent drawings (X_i, Y_i) of the generic pair (X, Y), where $\mathcal{X}, \mathcal{Y}$, and $\hat{\mathcal{X}}$ denote the alphabet of the source, the side information, and the reconstruction, respectively. The alphabets, unless otherwise indicated, are assumed to be finite. Without loss of generality, we assume that $P(x) > 0$ for all $x \in \mathcal{X}$, and $P(y) > 0$ for all $y \in \mathcal{Y}$. Note that we use P to denote probabilities in general, as well as specific PMFs. Thus, when A is an event, $P(A)$ will denote its probability, while, for $x \in \mathcal{X}$ and $y \in \mathcal{Y}$, $P(x), P(y)$, and $P(x, y)$ denote probabilities of the events $\{X = x\}, \{Y = y\}, \{X = x, Y = y\}$, respectively. Similarly, $P(y\,|\,x)$ will denote the probability of $\{Y = y\}$ conditioned on $\{X = x\}$.

A coding scheme for block length n and rate R consists of an encoder, which is a mapping $T : \mathcal{X}^n \to \{1, \ldots, \lfloor 2^{nR} \rfloor\}$, and a decoder, which is a sequence of mappings $\{\hat{X}_i\}_{i=1}^n$, where $\hat{X}_i : \{1, \ldots, \lfloor 2^{nR} \rfloor\} \times \mathcal{Y}^i \to \hat{\mathcal{X}}$. The scheme operates as follows: the encoder maps the source sequence $X^n = (X_1, \ldots X_n)$ into an index $T = T(X^n)$, and the decoder generates a reconstruction $\hat{X}^n = (\hat{X}_1, \ldots \hat{X}_n)$, where $\hat{X}_i = \hat{X}_i(T, Y^i)$ and $Y^i = (Y_1, \ldots, Y_i)$, i.e., the decoder can use the side information, but only *causally*. The performance of a scheme is measured by its expected per symbol distortion $\frac{1}{n}\sum_{i=1}^n E[\rho(X_i, \hat{X}_i)]$, where $\rho : \mathcal{X} \times \hat{\mathcal{X}} \to \mathbb{R}$ is a distortion measure. A rate distortion pair (R, D) is achievable if for every $\varepsilon > 0$ and sufficiently large n there exists a scheme for block length n and rate R with $\frac{1}{n}\sum_{i=1}^n E[\rho(X_i, \hat{X}_i)] \le D + \varepsilon$. The rate distortion function $R(D)$ is the infimum of rates R such that (R, D) is achievable. Letting $D_{\min} = E[\min_{\hat{x}} \rho(X, \hat{x})]$, it is clear that $D < D_{\min}$ is not achievable for any rate R. We are thus interested in the characterization of $R(D)$ for $D \ge D_{\min}$. The following theorem establishes $R(D)$.

Theorem 1: The rate distortion function for the case of causal side information at the decoder and $D \ge D_{\min}$ is given by

$$R(D) = \min I(X; W) \tag{1}$$

where the minimum is over all functions

$$f : \mathcal{W} \times \mathcal{Y} \to \hat{\mathcal{X}}, |\mathcal{W}| \le |\mathcal{X}| + 1$$

and $P(w\,|\,x)$ such that

$$E[\rho(X, f(W, Y))] \le D. \tag{2}$$

Proof of Theorem 1: The direct part follows from a standard random coding argument. Fix $P(w\,|\,x)$ and $f(w, y)$ that achieve the minimum in (1). Let $R = I(X; W) + \varepsilon$ and generate 2^{nR} independent and identically distributed (i.i.d.) codewords $W^n(s), s \in \{1, \ldots, 2^{nR}\}$, with components that are i.i.d. $\sim P(w)$. Given a source sequence X^n, the encoder looks for a codeword $W^n(s)$ such that $(X^n, W^n(s))$ are jointly typical, and sends s to the decoder (setting say $s = 1$ if no such codeword is found). The decoder creates its reconstruction $\hat{X}^n$ by letting $\hat{X}_i = f(W_i(s), Y_i)$. The Markov lemma [14, Lemma 14.8.1] guarantees that, for sufficiently large n, (X^n, W^n, Y^n) are jointly typical with probability arbitrarily close to one. The distortion between X^n and $\hat{X}^n$, for suffficiently large n, is thus arbitrarily close to $E[\rho(X, f(W, Y))]$ with probability arbitrarily close to one.

For the converse part, denote (until the end of the proof) the minimum in (1) by $\tilde{R}(D)$. We need to show for $D \ge D_{\min}$ that if (R, D) is an achievable pair, then $R \ge \tilde{R}(D)$. To this end, fix $\varepsilon > 0$ and a scheme for block length n and rate R satisfying

$$\frac{1}{n}\sum_{i=1}^n E[\rho(X_i, \hat{X}_i)] \le D + \varepsilon. \tag{3}$$

Then

$$\begin{aligned}
nR &\ge H(T) \\
&\ge I(X^n; T) \\
&= H(X^n) - H(X^n|T) \\
&= \sum_{i=1}^n H(X_i) - H(X_i|T, X^{i-1}) \\
&\overset{(a)}{\ge} \sum_{i=1}^n H(X_i) - H(X_i|T, Y^{i-1}) \\
&\overset{(b)}{=} \sum_{i=1}^n I(X_i; W_i) \\
&\overset{(c)}{\ge} \sum_{i=1}^n \tilde{R}(E[\rho(X_i, \hat{X}_i)]) \\
&\overset{(d)}{\ge} n\tilde{R}\left(\frac{1}{n}\sum_{i=1}^n E[\rho(X_i, \hat{X}_i)]\right) \\
&\overset{(e)}{\ge} n\tilde{R}(D + \varepsilon)
\end{aligned} \tag{4}$$

where

(a) follows from the Markovian structure $X_i \to (T, X^{i-1}) \to (T, Y^{i-1})$;

(b) follows by denoting $W_i = (T, Y^{i-1})$;

(c) follows by the fact that $\hat{X}_i = \hat{X}_i(T, Y^i) = \hat{X}_i(W_i, Y_i)$, the Markovian structure $W_i \to X_i \to Y_i$, the definition (as given in (1)) of $\tilde{R}(D)$, and the fact that the restriction $|\mathcal{W}| \le |\mathcal{X}| + 1$ does not change the minimum which follows from a standard application of Carathéodory's theorem (as stated in, e.g., [4, Theorem 14.3.4]);

(d) follows from the convexity of $\tilde{R}(D)$, which can be established in the same way as the convexity of the Wyner–Ziv rate distortion function, and given for completeness in the Appendix; and

(e) follows by (3) and the monotonicity of $\tilde{R}(D)$.

The proof is concluded by the arbitrariness of $\varepsilon > 0$ and the continuity from the right of $\tilde{R}(D)$ at all $D \ge D_{\min}$. This continuity follows, for $D > D_{\min}$, directly from the convexity of $\tilde{R}(D)$, while at $D = D_{\min}$, it follows from an argument similar to that in the proof of [5, Lemma 2.2]. □

Remarks:

1) The rate distortion function in Theorem 1 can be expressed as an ordinary rate distortion function with an appropriately defined distortion measure. Concretely, consider a

new alphabet $\mathcal{U}$ of size $|\hat{\mathcal{X}}|^{|\mathcal{Y}|}$ consisting of all mappings $u : \mathcal{Y} \to \hat{\mathcal{X}}$. Define a new distortion measure ρ' as

$$\rho'(x,u) = \sum_{y \in \mathcal{Y}} P(y \,|\, x)\rho(x, u(y)).$$

$R(D)$ of Theorem 1 is then seen to be the ordinary rate distortion function under distortion measure ρ'. Note that this is analogous to Shannon's capacity for memoryless channels with states known causally at the transmitter [31], which he expressed as the capacity of a discrete memorless channel whose channel input symbols are mappings from the set of channel states to the set of channel inputs (in the original channel).

2) Note that the only difference between the rate distortion function in Theorem 1 and the Wyner–Ziv rate distortion function

$$R_{\mathrm{WZ}}(D) = \min[I(X;W) - I(Y;W)] \tag{5}$$

which is minimized over *exactly* the same set as in (1), is the subtracted term $I(Y;W)$. This term arises in the achievability part of the Wyner–Ziv theorem as a consequence of binning, which allows the encoder to convey the index of the bin of W^n, rather than W^n itself. Theorem 1 implies that the noncausal availability of the side information is crucial for binning.

3) There is a certain similarity to channel coding with state information available at the encoder. The capacity when the state information is noncausally available was shown by Gel'fand and Pinsker [12] and Heegard and El Gamal [15] to be given by

$$C_{\mathrm{GP}} = \max_{P(u \,|\, s), P(x \,|\, u,s)} [I(U;Y) - I(U;S)], \tag{6}$$

while in [31], Shannon showed that when the state information is only causally available at the encoder, the capacity is given by

$$C_{\mathrm{CSI}} = \max_{P(u), P(x \,|\, u,s)} I(U;Y). \tag{7}$$

Here too, there is a subtracted mutual information term that disappears in the causal case. Unlike the rate distortion functions, however, here the maximization sets for the noncausal and causal settings are different, since in the latter case U and S are restricted to be independent. Since under this independence $I(U;S) = 0$ anyway, the penalty for causality in this setting can be regarded as due to the independence requirement. Here the analogy with our rate distortion problem breaks, since in general, the achieving distribution in (1) yields $I(Y;W) > 0$.

4) Note that the scheme in the achievability part of the proof is of the form $\hat{X}_i(T, Y_i)$, that is, the ith reconstruction uses only the ith side-information symbol and not the past symbols Y^{i-1}. This phenomenon where "if the future is not allowed to be looked into, the past is useless" occurs also in zero-delay [10], [11] and in causal [26] source coding.

5) In the spirit of the observations in [25], here too we observe that feedforward does not improve the rate–distortion tradeoff. In this context, feedforward means that $\hat{X}_i$ is allowed to depend not only on (T, Y^i), but also on X^{i-1}. The converse part of the proof of Theorem 1 would then carry through by adding X^{i-1} to the definition of W_i, i.e., setting $W_i = (T, Y^{i-1}, X^{i-1})$. The chain of equalities and inequalities leading to (4) remains valid by replacing the line of inequality (a) by $= \sum_{i=1}^{n} H(X_i) - H(X_i \,|\, T, Y^{i-1}, X^{i-1})$.

6) Consider the case where the encoder has access to the side information, i.e., the index conveyed to the decoder is of the form $T = T(X^n, Y^n)$. The reconstruction is restricted, as in our problem, to be of the form $\hat{X}_i(T, Y^i)$. We note that the conditional rate distortion function associated with the presence of side information at both encoder and decoder $R_{X \,|\, Y}(D)$ (cf., e.g., [2]) can be achieved independent of this restriction. Indeed, the original achievability argument carries over to this case. The sequence to be encoded is partitioned into subsequences according to the value of Y_i. The subsequence consisting of the indices i for which $Y_i = y$ is separately encoded using an optimal rate distortion code for the source $P_{X \,|\, Y=y}$. To produce the ith reconstruction symbol, it is clear that the decoder only needs to know the subsequence that Y_i belongs to, in addition to the index T, i.e., its reconstruction is of the form $\hat{X}_i(T, Y_i)$ (which, in particular, satisfies the causality constraint). Thus, when side information is available at the encoder too, restricting the reconstruction to depend causally on the side information entails no performance loss. This is in contrast to our problem setting with side information available only at the encoder, where, as Theorem 1 and the following examples show, the causality restriction can adversely impact the rate–distortion tradeoff.

Example: Doubly Symmetric Binary Source

Consider the case where X is the unbiased input to a BSC $(\delta), 0 \le \delta \le 1/2$, and Y is the corresponding output, and the distortion measure is the Hamming loss. The rate distortion function in this case is given by

$$R(D) = \begin{cases} 1 - h(D), & 0 \le D \le d_c \\ -h'(d_c)D + h'(d_c)\delta, & d_c < D \le \delta \end{cases} \tag{8}$$

where h is the binary entropy function, h' is its derivative, and d_c is the solution to the equation $(1 - h(d_c))/(d_c - \delta) = -h'(d_c)$. Thus, $R(D)$ coincides with the rate distortion function for $X, R_X(D)$, for $0 \le D \le d_c$, and is otherwise given by the straight-line tangent to the graph of $R_X(D)$ that passes through the point $(\delta, 0)$ for $d_c \le D \le \delta$. In other words, the optimum performance is achieved by time sharing between rate distortion coding with no side information and zero-rate decoding that uses only the side information. For small enough distortion, this performance is attained by simply ignoring the side information. The function $R(D)$ is plotted in Fig. 2 for $\delta = 1/4$.

To arrive at (8) from Theorem 1, note first that time sharing between rate distortion coding with no side information and zero-rate decoding that uses only the side information, i.e., using the former for a fraction θ of the time and using the latter for the remainder of the time at distortion level β, achieves a rate $\theta[1 - h(\beta)]$ and distortion level $D = \theta\beta + (1 - \theta)\delta$. Thus, since the right-hand side of (8) corresponds to *optimum* time

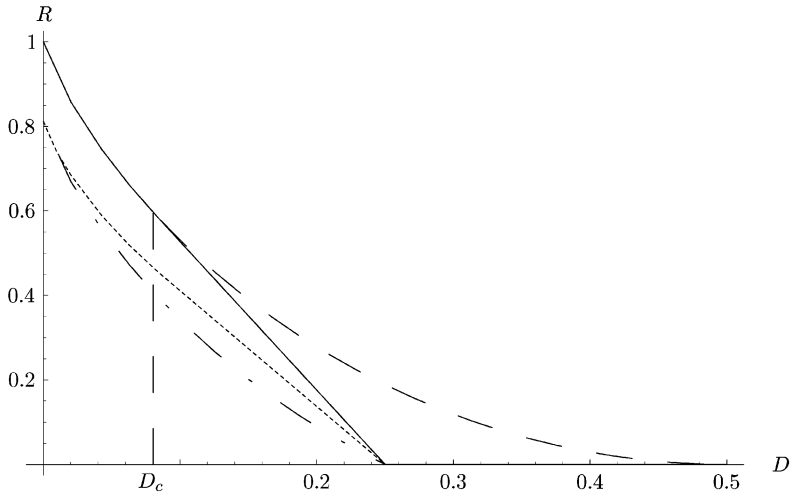

Fig. 2. Rate distortion curves for the doubly symmetric binary source and Hamming loss with $\delta = 1/4$. Curves are (from top to bottom): $R_X(D) = 1 - h(D)$ (dashed), $R(D)$ (solid), $R_{\mathrm{WZ}}(D)$ (dotted), $R_{X\,|\,Y}(D) = h(\delta) - h(D)$ (dot-dashed). We see that $R(D) = R_X(D)$ for $D \leq D_c$. For $D_c \leq D \leq \delta$, the graph of $R(D)$ is the straight line which is tangent to the graph of $R_X(D)$ and which passes through the point $(\delta, 0)$. The point of tangency is $(D_c, R_X(D_c))$. Thus, D_c is the solution to $\frac{1-h(D_c)}{D_c-\delta} = -h'(D_c)$. In this particular case of $\delta = 1/4$, D_c can be explicitly computed and is given by $D_c = (5 - 4(19 - 3\sqrt{33})^{-1/3} - (19 - 3\sqrt{33})^{1/3})/6 \approx 0.0803566$.

sharing between rate distortion coding with no side information and zero-rate decoding that uses only the side information, it can be expressed in the alternative form

$$R^*(D) = \inf_{\theta,\beta} \theta \cdot [1 - h(\beta)], \qquad \text{for } 0 \leq D \leq \delta \tag{9}$$

where the infimum is with respect to both the time sharing fraction $0 \leq \theta \leq 1$ and the distortion $0 \leq \beta < \delta$ (incurred in the fraction of the time where rate distortion with no side information is employed), such that

$$D = \theta\beta + (1 - \theta)\delta. \tag{10}$$

Thus, in this representation, θ corresponds to the fraction of time rate distortion coding with no side information is used, and β corresponds to the distortion level at which this coding operates. Assuming $W, X, Y, \hat{X} = f(Y, Z)$ achieve the minimum in (1) for distortion level D, it remains to show that $I(X; W) \geq R^*(D)$. To do so, define the sets

$$A = \{w : f(w, 0) = f(w, 1)\}$$

and

$$A^c = \{w : f(w, 0) \neq f(w, 1)\} \tag{11}$$

and note that by law of total expectation

$$E[\rho(X, \hat{X})] = P(W \in A)E[\rho(X, \hat{X}) \,|\, W \in A] + P(W \in A^c)E[\rho(X, \hat{X}) \,|\, W \in A^c] \leq D. \tag{12}$$

Also, the inequality

$$E[\rho(X, \hat{X}) \,|\, W \in A^c] \geq \delta \tag{13}$$

can be proved in the same way as equality (36) in [42]. Therefore, using

$$E[\rho(X, \hat{X}) \,|\, W \in A] = \sum_{w \in A} \frac{P(W = w)}{P(W \in A)} E[\rho(X, \hat{X}) \,|\, W = w]$$

(12) and (13) yield

$$d' \triangleq \theta \sum_{w \in A} \lambda_w d_w + (1 - \theta)\delta \leq D \tag{14}$$

where $\theta = P(W \in A), \lambda_w = \frac{P(W=w)}{P(W \in A)}$, and $d_w = E[\rho(X, \hat{X}) \,|\, W = w]$. Note that for $w \in A$, defining $\gamma(w) = f(w, 0) = f(w, 1)$

$$d_w = E[\rho(X, \hat{X}) \,|\, W = w] = P(X \neq \gamma(w) \,|\, W = w)$$

so that $H(X \,|\, W = w) = h(d_w)$. It follows that

$$\begin{aligned} I(X; W) &= 1 - H(X \,|\, W) \\ &\geq \sum_{w \in A} [1 - H(X \,|\, W = w)]P(W = w) \\ &= \theta \sum_{w \in A} [1 - h(d_w)]\lambda_w \\ &\geq \theta \left[1 - h\left(\sum_{w \in A} d_w \lambda_w\right)\right] \\ &\geq R^*(D) \end{aligned}$$

where the inequality before last follows from the concavity of $h(\,\cdot\,)$, and the last inequality follows from the definition of $R^*(D)$ in (9) and the fact that, by (14), $\lambda_w d_w \leq d' \leq D \leq \delta$.

Example: X, Y Jointly Gaussian

As in [42], [41], the results from the finite-alphabet setting and the form of the rate distortion function carry over to X, Y taking values in general alphabets. Consider the case of $X \sim \mathcal{N}(0, \sigma_X^2)$ and $Y = X + N$, where $N \sim \mathcal{N}(0, \sigma_N^2)$ and is

independent of X,[1] and ρ is the squared error distortion. We obtain an upper bound on $R(D)$ by taking $W = X + Z$, where $Z \sim \mathcal{N}(0, \sigma_Z^2)$ is independent of X, N.[2]

Assume first that $\sigma_X^2 = 1$. The optimal estimate of X based on W and Y is $\hat{X} = \alpha W + \beta Y$, where, by the orthogonality principle, α, β are the solutions to the linear equation

$$\begin{bmatrix} 1+\sigma_Z^2 & 1 \\ 1 & 1+\sigma_N^2 \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \tag{15}$$

and are given by

$$\begin{aligned} \alpha &= \frac{\sigma_N^2}{\sigma_N^2 + \sigma_N^2\sigma_Z^2 + \sigma_Z^2} \\ \beta &= \frac{\sigma_Z^2}{\sigma_N^2 + \sigma_N^2\sigma_Z^2 + \sigma_Z^2}. \end{aligned} \tag{16}$$

The resulting mean-square error (MSE) is

$$\begin{aligned} D &= E(X - \hat{X})^2 \\ &= [1 - (\alpha + \beta)]^2 + \alpha^2\sigma_Z^2 + \beta^2\sigma_N^2 \\ &= \frac{\sigma_N^2\sigma_Z^2}{\sigma_N^2 + \sigma_N^2\sigma_Z^2 + \sigma_Z^2}. \end{aligned} \tag{17}$$

Also

$$I(X; W) = \frac{1}{2}\log\left(1 + \frac{1}{\sigma_Z^2}\right). \tag{18}$$

This shows that the distortion in (17) is achievable at the rate in (18). In other words, if $R = \frac{1}{2}\log(1 + \frac{1}{\sigma_Z^2})$, or equivalently, $\sigma_Z^2 = 1/(2^{2R} - 1)$, an upper bound on the distortion rate function when $\sigma_X^2 = 1$ is

$$\begin{aligned} D_{\mathrm{ub}}(R) &= \frac{\sigma_N^2/(2^{2R}-1)}{\sigma_N^2 + \sigma_N^2/(2^{2R}-1) + 1/(2^{2R}-1)} \\ &= \frac{\sigma_N^2}{\sigma_N^2 \cdot 2^{2R} + 1}. \end{aligned} \tag{19}$$

For $\sigma_X^2 \neq 1$, the solution is obtained by considering the equivalent side information $Y/\sigma_X = X/\sigma_X + N/\sigma_X$. It is clear that the distortion rate function is given by multiplying the function for the case $\sigma_X^2 = 1$ by σ_X^2, and substituting in it σ_N^2/σ_X^2 for σ_N^2. We thus obtain

$$D_{\mathrm{ub}}(R) = \frac{\sigma_N^2\sigma_X^2}{\sigma_N^2 \cdot 2^{2R} + \sigma_X^2}. \tag{20}$$

Inverting the function, we obtain the equivalent bound on the rate distortion function

$$R_{\mathrm{ub}}(D) = \begin{cases} \frac{1}{2}\log\left[\sigma_X^2\left(\frac{1}{D} - \frac{1}{\sigma_N^2}\right)\right], & 0 < D \le \frac{\sigma_X^2\sigma_N^2}{\sigma_X^2+\sigma_N^2} \\ 0, & D \ge \frac{\sigma_X^2\sigma_N^2}{\sigma_X^2+\sigma_N^2}. \end{cases} \tag{21}$$

[1]Note that the seemingly more general situation of zero-mean jointly Gaussian random variables X, Y that satisfy the relation $Y = \alpha X + N$, where α is a deterministic constant, is subsumed by our setting by letting $X' = \alpha X$ and $\hat{X} = \alpha^{-1}\hat{X}'$, where $\hat{X}'$ is the reconstruction of X'.

[2]Note that there is no benefit in considering a more general W of the form $\gamma X + Z, \gamma \neq 0$, as this contains the same information as $X + Z/\gamma = X + Z'$, which is the W of the form we are considering (at a different value of $\sigma_{Z'}^2$).

Note that $R_{\mathrm{ub}}(D)$ is not necessarily convex, and hence this bound can be improved by convexification. Differentiating in the region $D < \sigma_X^2\sigma_N^2/(\sigma_X^2 + \sigma_N^2)$ gives

$$R'_{\mathrm{ub}}(D) = -\frac{1}{2D^2\left(\frac{1}{D} - \frac{1}{\sigma_N^2}\ln 2\right)} \tag{22}$$

and

$$R''_{\mathrm{ub}}(D) = \frac{\sigma_N^2\left(\sigma_N^2 - 2D\right)}{D^2\left(D - \sigma_N^2\right)^2 \cdot 2\ln 2}. \tag{23}$$

We see that $R''_{\mathrm{ub}}(D)$ is positive or negative depending on whether $D < \sigma_N^2/2$ or $D > \sigma_N^2/2$, provided of course that $\sigma_N^2/2$ lies in the relevant distortion region $[0, \sigma_X^2\sigma_N^2/(\sigma_X^2 + \sigma_N^2)]$, i.e., that $\sigma_N^2 < \sigma_X^2$. We thus have the following.

Lemma 1:

1) For $\sigma_N^2 \ge \sigma_X^2$, $R_{\mathrm{ub}}(D)$, as given in (21), is convex.
2) For $\sigma_N^2 < \sigma_X^2$, $R_{\mathrm{ub}}(D)$ has an inflection point at $D = \sigma_N^2/2$: It is convex for $D < \sigma_N^2/2$ and concave for $\sigma_N^2/2 < D \le \sigma_X^2\sigma_N^2/(\sigma_X^2 + \sigma_N^2)$.

For $\sigma_N^2 < \sigma_X^2$, we can improve on $R_{\mathrm{ub}}(D)$ of (21) by taking its lower convex envelope $\underline{R}_{\mathrm{ub}}(D)$. More explicitly, the envelope $\underline{R}_{\mathrm{ub}}(D)$ coincides with $R_{\mathrm{ub}}(D)$ for $0 \le D \le D_c$, where D_c is the solution of the equation

$$\frac{R_{\mathrm{ub}}(D_c)}{D_c - \frac{\sigma_X^2\sigma_N^2}{\sigma_X^2+\sigma_N^2}} = R'_{\mathrm{ub}}(D_c). \tag{24}$$

For $D_c \le D \le \sigma_X^2\sigma_N^2/(\sigma_X^2 + \sigma_N^2)$, the graph of $\underline{R}_{\mathrm{ub}}(D)$ is the straight-line tangent to $R_{\mathrm{ub}}(D)$ that connects the point $(D_c, R_{\mathrm{ub}}(D_c))$ to the point $(\sigma_X^2\sigma_N^2/(\sigma_X^2 + \sigma_N^2), 0)$. Fig. 3 shows examples of these curves. Whether or not $R(D) = \underline{R}_{\mathrm{ub}}(D)$ remains to be determined.

For comparison, we recall the Wyner–Ziv rate distortion function, which in the Gaussian case coincides with the conditional rate distortion function

$$R_{X\,|\,Y}(D) = \begin{cases} \frac{1}{2}\log\left[\frac{\sigma_X^2\sigma_N^2}{\sigma_X^2+\sigma_N^2}\frac{1}{D}\right], & 0 < D \le \frac{\sigma_X^2\sigma_N^2}{\sigma_X^2+\sigma_N^2} \\ 0, & D \ge \frac{\sigma_X^2\sigma_N^2}{\sigma_X^2+\sigma_N^2}. \end{cases} \tag{25}$$

In particular

$$\begin{aligned} R_{\mathrm{ub}}(D) - R_{X\,|\,Y}(D) &= \frac{1}{2}\log\left[\frac{\left(\sigma_N^2 - D\right)\left(\sigma_X^2 + \sigma_N^2\right)}{\sigma_N^4}\right] \\ &\le \frac{1}{2}\log\left[\frac{\sigma_X^2 + \sigma_N^2}{\sigma_N^2}\right]. \end{aligned} \tag{26}$$

So, for example, when $\sigma_X^2 \le \sigma_N^2$, the penalty of causal dependence of the reconstruction on the side information is no more than $1/2$ a bit per source symbol, independent of the distortion level.

Alternative Characterization of $R(D)$

Denote by G_X the undirected graph with vertex set $\mathcal{X}$, where an edge (x, x') exists if and only if there exists $y \in \mathcal{Y}$ such

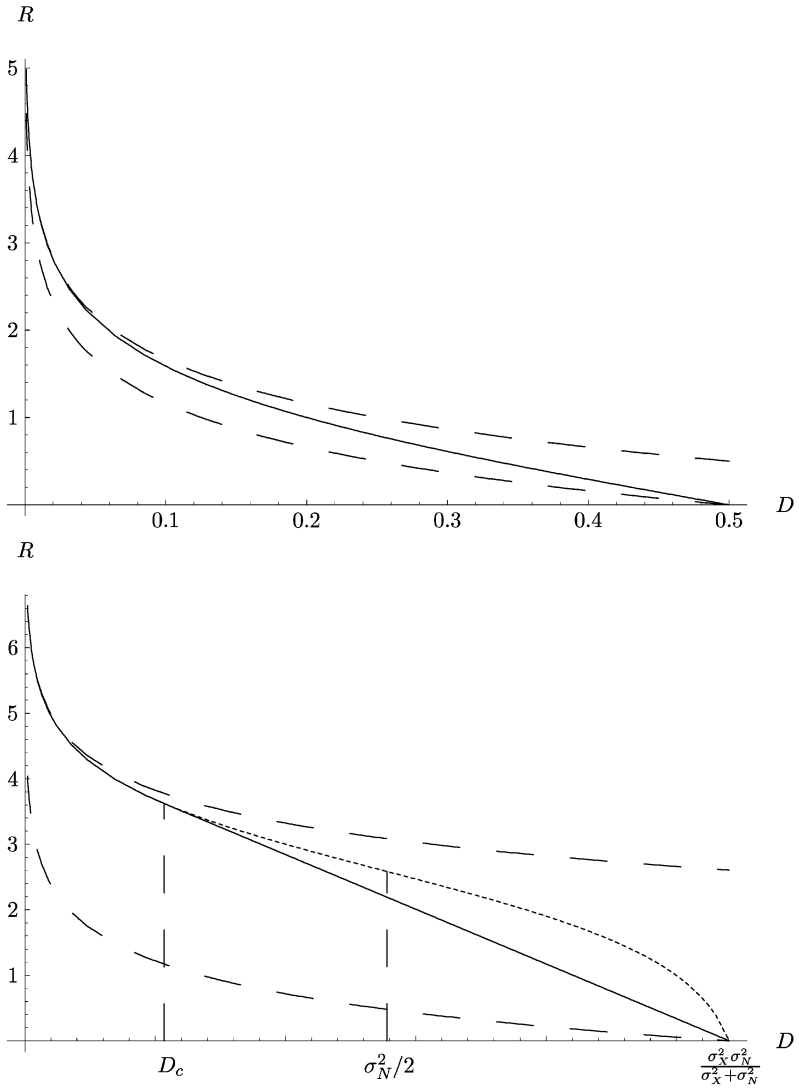

Fig. 3. Plots of $R_X(D)$ (upper dashed), $R_{\rm ub}(D)$ (dotted), its lower convex envelope $\underline{R}_{\rm ub}(D)$ (solid), and $R_{X\,|\,Y}(D)$ (lower dashed). Top graph shows curves for the case $\sigma_X = \sigma_N = 1$, where by Lemma 1, $R_{\rm ub}(D) = \underline{R}_{\rm ub}(D)$. Bottom graph shows curves for $\sigma_X = 1, \sigma_N = 1/6$. Lemma 1 implies that $R_{\rm ub}(D)$ is not convex (and is therefore improved by $\underline{R}_{\rm ub}(D)$). The solution to (24) in this case is $D_c = 5.352215 \times 10^{-3}$. The inflection point of $R_{\rm ub}(D)$, as stated in Lemma 1, is at $D = \sigma_N^2/2$.

that both $P(x,y) > 0$ and $P(x',y) > 0$.[3] Let $\{A_j\}_{j=1}^{M}$ $(M \leq |\mathcal{X}|)$ be a partition of $\mathcal{X}$ into the vertex sets associated with maximally connected components of G_X, and let $i(x)$ denote the index of the set j for which $x \in A_j$. As in [39], we define the *common information* random variable Z as

$$Z = i(X). \tag{27}$$

Note that for every $y \in \mathcal{Y}$ such that both $P(x,y) > 0$ and $P(x',y) > 0, i(x) = i(x')$. Thus, Z is a deterministic function both of X and of Y. Furthermore, any other random variable Z' for which there exist (deterministic) functions f and g such that $Z' = f(X) = g(Y)$ almost surely (a.s.) is a deterministic function of Z. To see this, consider any (x,x') such that $i(x) = i(x')$, and y for which both $P(x,y) > 0$ and $P(x',y) > 0$ (the fact that $i(x) = i(x')$ guarantees the existence of at least one such y). Then we must have both $g(y) = f(x)$ and $g(y) = f(x')$, since otherwise $P(g(Y) \neq f(X)) > 0$. Clearly, $i(x) = i(x')$ implies that $f(x) = f(x')$ or, in other words, $f(x)$ depends on x only through $i(x)$. But $Z = i(X)$ and $Z' = f(X)$, so Z' is a deterministic function of Z. Thus, of all random variables Z' with the property that there exist f, g such that $Z' = f(X) = g(Y)$ a.s., the common information Z is the most informative one. Fig. 4 gives an example of G_X and its associated Z. An equivalent characterization of the rate distortion function that explicitly includes Z is as follows.

Proposition 1: The rate distortion function of Theorem 1 is equivalently given by

$$R(D) = \min I(X; W \,|\, Z), \qquad D \geq D_{\min} \tag{28}$$

where Z is the common information between X and Y, and the minimum is over the same set as the minimum in (1), namely, over all functions $f : \mathcal{W} \times \mathcal{Y} \to \hat{\mathcal{X}}, |\mathcal{W}| \leq |\mathcal{X}|+1$, and $P(w\,|\,x)$ such that

$$E[\rho(X, f(W,Y))] \leq D. \tag{29}$$

[3]When convenient, we identify G_X with its set of edges.

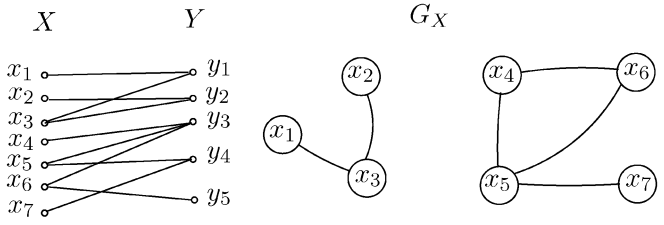

Fig. 4. Example for $|\mathcal{X}| = 7$ and $|\mathcal{Y}| = 5$. The bipartite graph associated with the jointly distributed X, Y has an edge between x_i and y_j if and only if $P(X = x_i, Y = y_j) > 0$. G_X is the graph whose nodes are the elements of $\mathcal{X}$ with an edge between x_i and x_j if and only if there is a y such that both (x_i, y) and (x_j, y) are edges in the bipartite graph. The graph G_X in this example consists of two maximally connected components, so Z is binary, assuming one value on $\{x_1, x_2, x_3\}$ and another value on $\{x_4, x_5, x_6, x_7\}$.

The characterization in (28) is useful when G_X has more than one maximally connected component. In such cases, the minimum in (28) decomposes to a computation of the rate distortion functions of the simpler sources $\{X|Z = z\}_z$ (corresponding to the maximally connected components of G_X). Such decomposition will be exploited in some computations in Section III.

Proof of Proposition 1: Denote the right-hand side of (28) by $R'(D)$. We need to show that $R'(D) = R(D)$. The achievability argument we outline is similar to that in the proof of Theorem 1: The sequence to be encoded can be partitioned into subsequences according to the value of Z_i (which both encoder and decoder know). The subsequence consisting of the indices i for which $Z_i = z$ will appear $\approx nP(z)$ times, and encoding the corresponding subsequence of W_is will take $\approx nP(z)I(X; W \,|\, Z = z)$ bits. Hence, the overall number of bits used will be

$$\approx n \sum_z P(z) I(X; W \,|\, Z = z) = nI(X; W \,|\, Z).$$

Reconstructing according to $\hat{X}_i = f(W_i, Y_i)$ will yield cumulative distortion $\approx nP(z)E[\rho(X, f(W, Y))|Z = z]$ on the subsequence corresponding to z, and thus overall cumulative distortion

$$\approx n \sum_z P(z) E[\rho(X, f(W,Y))|Z=z] = nE[\rho(X, f(W,Y))].$$

Thus, $R(D) \le R'(D)$. On the other hand, under any distribution in the feasible set (over which the minimum in (1) and (28) is taken)

$$\begin{aligned} I(X; W) &= H(W) - H(W \,|\, X) && (30)\\ &\ge H(W \,|\, Z) - H(W|X, Z) && (31)\\ &= I(X; W \,|\, Z) && (32) \end{aligned}$$

where the inequality follows since $H(W) \ge H(W \,|\, Z)$ and $H(W \,|\, X) = H(W|X, Z)$ (as Z is a deterministic function of X). Thus, $R(D) \ge R'(D)$. □

Recapping the above proof: $I(X; W) \ge I(X; W \,|\, Z)$ implies that the right-hand side of (1) is lower-bounded by that of (28), while the reverse inequality is implied by the achievability of $I(X; W \,|\, Z)$ and the converse part of Theorem 1. It is useful to verify that the right-hand side of (1) equals that of (28) without appealing to Theorem 1. This is done in the following alternative proof.

Alternative Proof of Proposition 1: Denote the minimum in (1) by $R^T(D)$ and that in (28) by $R^P(D)$. As argued, $R^T(D) \ge R^P(D)$ is immediate since the minimized expression in $R^T(D)$ is lower-bounded by that in $R^P(D)$. To prove that $R^T(D) \le R^P(D)$, assume that W, X, Y, Z achieve the minimum in (28) and let $\{1, \ldots, M_W\}$ be the alphabet for W. Note that W, X, Y, Z satisfy the Markov relation $(W, Z) \to X \to Y$. Now construct $\tilde{W}$ such that $\tilde{W} \to (W, Z) \to X \to Y$ according to

$$\begin{aligned} \tilde{W}|\{W = i, Z = z\} \sim \mathrm{Unif}[&P(W \in \{1, \ldots i-1\}|Z = z),\\ &P(W \in \{1, \ldots i\}|Z = z)], \quad 1 \le i \le M_W \end{aligned} \tag{33}$$

where $\mathrm{Unif}[a, b]$ denotes the uniform distribution on $[a, b]$. Note the following.

1) By construction, $\tilde{W} \,|\, \{Z = z\} \sim \mathrm{Unif}[0, 1]$ for all z. Thus, $\tilde{W}$ is independent of Z.
2) There exists a (deterministic) function g such that $g(\tilde{W}, Z) = W$ a.s. Specifically, the following g is readily seen to have this property:

$$\begin{aligned} g(\tilde{w}, z) = \sum_{1=1}^{M_W} i \cdot 1(\tilde{w} \in [&P(W \in \{1, \ldots i-1\} \,|\, Z = z),\\ &P(W \in \{1, \ldots i\} \,|\, Z = z))) \end{aligned}$$

where $1(\cdot)$ denotes the indicator function.

Hence,

$$\begin{aligned} I(X; \tilde{W}) &\overset{(a)}{=} I(X; \tilde{W} \,|\, Z)\\ &= H(X \,|\, Z) - H(X \,|\, \tilde{W}, Z)\\ &\overset{(b)}{=} H(X \,|\, Z) - H(X|\tilde{W}, W, Z)\\ &\overset{(c)}{=} H(X \,|\, Z) - H(X \,|\, W, Z)\\ &= I(X; W \,|\, Z)\\ &= R^P(D) \end{aligned} \tag{34}$$

where (a) follows from the independence of $\tilde{W}$ and Z, (b) from the fact that W is determined by $(\tilde{W}, Z)$, and (c) follows from the Markov relation $\tilde{W} \to (W, Z) \to X$.

Letting α denote the function satisfying $Z = \alpha(Y)$ a.s. and defining $\tilde{f}(\tilde{w}, y) = f(g(\tilde{w}, \alpha(y)), y)$, we have

$$\begin{aligned} E[\rho(X, \tilde{f}(\tilde{W}, Y))] &= E[\rho(X, f(g(\tilde{W}, \alpha(Y)), Y))]\\ &= E[\rho(X, f(g(\tilde{W}, Z), Y))]\\ &= E[\rho(X, f(W, Y))] \le D. \end{aligned}$$

The Markov relation $\tilde{W} \to X \to Y$ holds (due to the relation $\tilde{W} \to (W, Z) \to X \to Y$).

These arguments imply that the minimum in (1), when the cardinality of W is not restricted, is upper-bounded by $R^P(D)$. But, as already argued in the proof of Theorem 1, this restriction does not affect the minimum. □

Remarks:

1) Note that the independence of $\tilde{W}$ and Z in the above proof implies that the minima in (1) and (28) are not affected by imposing the additional condition of independence of W and Z when the constraint on the cardinality of $\mathcal{W}$ is not imposed.

2) Note that under the Markov relation $W \to X \to Y$

$$I(X;W) - I(Y;W) = I(X;W|Y) = I(X;W|Y,Z)$$

where the second equality follows since Z is a deterministic function of Y. That is, unlike in our case above, conditioning on Z is inconsequential for the Wyner–Ziv setting (i.e., for simplifying the minimization in (5)).

III. Lossless Source Coding With Causal Side Information

Consider the lossless source coding version of our problem. As in the previous section, the encoder maps the sequence X^n into $T \in \{1, \ldots, 2^{nR}\}$, and reconstruction is of the form $\hat{X}_i(T, Y^i)$. A rate R is achievable if there exists a sequence of schemes of rate R for which $P(X^n \neq \hat{X}^n) \longrightarrow 0$. Let R_{LSCSI} (subscript standing for "lossless source coding with causal side information") denote the infimum over achievable rates for this problem.

Clearly, $H(X\,|\,Y) \leq R_{\mathrm{LSCSI}} \leq H(X)$, where the lower bound is known to be achievable when causal dependence on the side information is not imposed [32], and the upper bound can be achieved without side information. So, where does R_{LSCSI} lie in the interval $[H(X\,|\,Y), H(X)]$? First consider the following three cases:

1) $X = Y$ a.s.: $R_{\mathrm{LSCSI}} = H(X\,|\,Y)(= 0)$;
2) X and Y independent: $R_{\mathrm{LSCSI}} = H(X\,|\,Y)(= H(X))$;
3) U and Y independent, and $X = (U,Y)$: $R_{\mathrm{LSCSI}} = H(X\,|\,Y)(= H(U))$.

In these cases, it is trivial to see that $R_{\mathrm{LSCSI}} = H(X\,|\,Y)$. This, however, turned out to be the exception. The following result shows that, under weak conditions on $P(x,y)$, causal side information is useless.

Theorem 2:

$$R_{\mathrm{LSCSI}} = \min I(X;W) = \min I(X;W\,|\,Z) \qquad (35)$$

where (in both minima) the minimization is over all $P(w\,|\,x)$, $|\mathcal{W}| \leq |\mathcal{X}| + 1$, such that $H(X|W,Y) = 0$.

Theorem 2 is not surprising when noting that the expression in the right-hand side of (35) is the value of $\lim_{D \to 0} R(D)$, where $R(D)$ is the rate–distortion function in Theorem 1 under Hamming loss. Showing that this limit is a lower bound on R_{LSCSI} is trivial. Proving that it is achievable is not entirely immediate because the definition of an achievable rate in the lossless setting requires diminishing block error probability, while in Theorem 1 achievability requires only diminishing *symbol* error probability. In the proof that follows, we use the method of types to prove achievability.

We will use $R_{\mathrm{LSCSI}}(P_{X,Y})$ when we wish to make the dependence of R_{LSCSI} on the distribution of (X,Y) explicit. Note that Theorem 2 implies the relationship

$$R_{\mathrm{LSCSI}}(P_{X,Y}) = \sum_z P(z) R_{\mathrm{LSCSI}}(P_{X,Y\,|\,Z=z}) \qquad (36)$$

where $P_{X,Y\,|\,Z=z}$ denotes the joint pmf of X,Y conditioned on $Z = z$. Thus, when G_X has more than one maximally connected component, finding $R_{\mathrm{LSCSI}}(P_{X,Y})$ reduces to a computation of the R_{LSCSI} associated with each component.

Proof of Theorem 2: Consider the setting of Section II with $\mathcal{X} = \hat{\mathcal{X}}$ and ρ being the Hamming loss. Note that the two minima in (35) are those in (1) and in (28) evaluated at $D = D_{\min} = 0$, respectively. Thus, by Proposition 1, they are equal, and by the converse part of Theorem 1, they lower-bound R_{LSCSI}. Letting R^*_{LSCSI} denote these minima, it remains to show that $R_{\mathrm{LSCSI}} \leq R^*_{\mathrm{LSCSI}}$. Note that this does not directly follow from the direct part of Theorem 1, since the setting of Theorem 1 (and, hence, the argument used in proving its achievability part) under Hamming loss at $D = 0$ corresponds to vanishing *symbol* error rate rather than the *block* error probability under which R_{LSCSI} is defined. To account for this, assume that X, Y, W achieve the (first) minimum in (35), as well as the associated f that satisfies $E[\rho(X, f(W,Y))] = 0$, or equivalently

$$f(w,y) = x, \quad \text{for all } x,y,w : P(x,y) > 0 \text{ and } P(w\,|\,x) > 0. \qquad (37)$$

Proposition 1 of [7] implies the existence of a mapping $g_n : T^n_{[X]} \to \mathcal{W}^n$ satisfying

$$g_n(x^n) \in T^n_{[W|X]}(x^n) \quad \text{for all } x^n \in T^n_{[X]} \qquad (38)$$

and

$$|\left\{g_n(x^n) : x^n \in T^n_{[X]}\right\}| \leq 2^{n(I(X;W)+\varepsilon_n)} \qquad (39)$$

where we here use the notation and conventions of [5] for typical and jointly typical sets, and $\varepsilon_n \to 0$. Now consider a coding scheme where the encoder conveys the index of $g_n(x^n)$ when $x^n \in T^n_{[X]}$ (according to some arbitrary indexing of the set on the left side of (39)), otherwise, it outputs an arbitrary member of $\{1, \ldots, 2^{n(I(X;W)+\varepsilon_n)}\}$. The decoder then produces

$$\hat{x}_i = f(g_{n,i}(x^n), y_i) \qquad (40)$$

where $g_{n,i}(x^n)$ denotes the ith component of the n-tuple $g_n(x^n)$. Note that, by the definition of $T^n_{[W\,|\,X]}$ (which, in particular, implies that if $w^n \in T^n_{[W\,|\,X]}(x^n)$ then $P(W = w_i\,|\,X = x_i) > 0$ for $1 \leq i \leq n$), for all sufficiently large n, every $1 \leq i \leq n$ and $x^n \in T^n_{[X]}$

$$P(W = g_{n,i}(x^n)\,|\,X = x_i) > 0. \qquad (41)$$

The combination of (37), (40), and (41) implies that

$$P(\hat{X}^n = X^n | X^n \in T^n_{[X]}) = 1 \qquad (42)$$

and consequently

$$P(\hat{X}^n \neq X^n) \leq P\left(X^n \notin T^n_{[X]}\right) \longrightarrow 0. \qquad (43)$$

We have thus constructed a code of rate $I(X;W)+\varepsilon_n$ with vanishing probability of block reconstruction error. This concludes the proof of the direct part. □

The $P(w\,|\,x)$ that achieves the minimum in (35) must satisfy the following constraints.

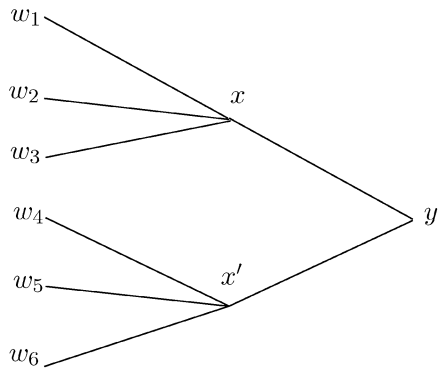

Fig. 5. Illustration of the condition in Lemma 2: $(x, x') \in G_X$ implies that x and x' are connected by some y for which $P(x, y)$ and $P(x', y)$ are positive. Then Y cannot completely distinguish between x and x'. Therefore, W will need to have this property, i.e., to satisfy $N_W(x) \cap N_W(x') = \emptyset$. In the illustration, $N_W(x) = \{w_1, w_2, w_3\}$ and $N_W(x') = \{w_1, w_2, w_3\}$, and the condition is that no w symbol be connected to both x and x'.

Lemma 2: Let W, X, Y be discrete random variables with the Markov relation $W \to X \to Y$. For each x, define

$$N_W(x) = \{w : P(w \,|\, x) > 0\}. \tag{44}$$

Then $H(X|W, Y) = 0$ if and only if

$$N_W(x) \cap N_W(x') = \emptyset \quad \text{whenever } (x, x') \in G_X. \tag{45}$$

The idea behind the proof of Lemma 2 is illustrated in Fig. 5 and the proof is given in Appendix II. When combined with Theorem 2, Lemma 2 implies that $R_{\mathrm{LSCSI}} = \min I(X; W)$, where the minimum is over $P(w \,|\, x)$ satisfying (45). In particular, we observe that R_{LSCSI} depends only on the distribution of X and on G_X. The dependence on $P(y \,|\, x)$ is only through its effect on G_X. For example, when G_X is complete, the condition (45), which becomes $N_W(x) \cap N_W(x') = \emptyset$ for all $x \neq x'$ implies that $H(X \,|\, W) = 0$, and when combined with Theorem 2 gives the following.

Corollary 1: $R_{\mathrm{LSCSI}} = H(X)$ whenever G_X is complete.

Remarks:

1) It is interesting to note that R_{LSCSI}, as characterized in Theorem 2, coincides with Körner's "graph entropy" [17], [18] (also known as "Körner's entropy," cf. [1, Sec. III]), which arises as the exponential rate of growth of the "ε-chromatic" number of the nth power of the graph G_X with the product distribution induced by P_X. Indeed, Corollary 1 and Corollary 2 (below) can also be seen to follow from properties of this graph entropy that were proved in [18].
2) Note that G_X is complete whenever X, Y satisfy the positivity condition $P(x, y) > 0$ for all $x \in \mathcal{X}, y \in \mathcal{Y}$. Corollary 1 implies that in such cases, the restriction to causality in the side information not only precludes achievability of the conditional entropy $H(X \,|\, Y)$ of [32], but in fact renders the side information useless. On the other hand, as examples 1 and 3 preceding Theorem 2 show, when this positivity condition does not hold, R_{LSCSI} can be strictly smaller than $H(X)$ and, in fact, can be as small as $H(X \,|\, Y)$. Clearly, Corollary 1 and the aforementioned examples show that $R_{\mathrm{LSCSI}}(P_{X,Y}) = H(X)$ in the interior of the simplex of distributions on $\mathcal{X} \times \mathcal{Y}$ with a discontinuity at its boundary. While discontinuities of this type are well known to arise in problems such as zero-error channel coding [30] and the zero-error Slepian–Wolf problem [40], [1], [19], it is interesting to see it arise in our setting, which assumes the standard "near-lossless" formulation. This discontinuity was observed also in [18], whose "ε-chromatic" number is also defined through a near-lossless source coding problem.
3) Consider our lossless source coding when the encoder also has access to the side information, i.e., the index from the encoder is allowed to be of the form $T(X^n, Y^n)$, while the reconstruction is still restricted to be of the form $\hat{X}_i(T, Y^i)$. An argument similar to that of partitioning to subsequences indexed by the $\mathcal{Y}$-alphabet given in the last of the remarks following the proof of Theorem 1 shows that $H(X \,|\, Y)$ is achievable. Furthermore, if variable-rate encoding is allowed, the same argument implies that $H(X \,|\, Y)$ is achievable even under the requirement for strictly lossless (as opposed to near-lossless) reconstruction, regardless of the restriction to causal decoder side information. Therefore, in the presence of encoder side information, the restriction to causal side information at the decoder entails no loss. This is in stark contrast to the case when side information is not available at the encoder where, as discussed in the previous item, the causal dependence requirement entails a severe penalty. Viewed alternatively, whereas availability of encoder side information does not improve on the compression limit in the absence of a causality requirement on decoder side information, when such causality is imposed, availability of encoder side information can mean the difference between a minimum rate of $H(X)$ and a minimum rate of $H(X \,|\, Y)$.

Theorem 2 and Lemma 2 also imply the following monotonicity of R_{LSCSI}.

Corollary 2: Let X, Y and X', Y' be pairs of jointly distributed random variables with $X \overset{d}{=} X'$. Let $R_{\mathrm{LSCSI}}, R'_{\mathrm{LSCSI}}$ denote the respective optimal rates. If $G_X \subseteq G_{X'}$ then $R_{\mathrm{LSCSI}} \leq R'_{\mathrm{LSCSI}}$.

Proof: By the remark following Lemma 2, R_{LSCSI} and R'_{LSCSI} are both given as a minimum of the same mutual information but, since $G_X \subseteq G_{X'}$, the set over which the minimum defining R_{LSCSI} is taken is less constrained (larger) than that associated with R'_{LSCSI}. □

Theorem 2 characterizes R_{LSCSI} as a solution to a simple finite-dimensional optimization problem. The conditioning on Z in (35), Lemma 2, Corollary 1, and Corollary 2 can be used to simplify the optimization. Also often useful for these computations is the representation of the rate distortion function as an ordinary rate distortion function, as detailed in the first remark following Theorem 1. The computation of R_{LSCSI}, which is the limit of that function as $D \to 0$, is then equivalent to the minimization of the Lagrangian

$$I(X; U) + \beta E\rho'(X, U)$$

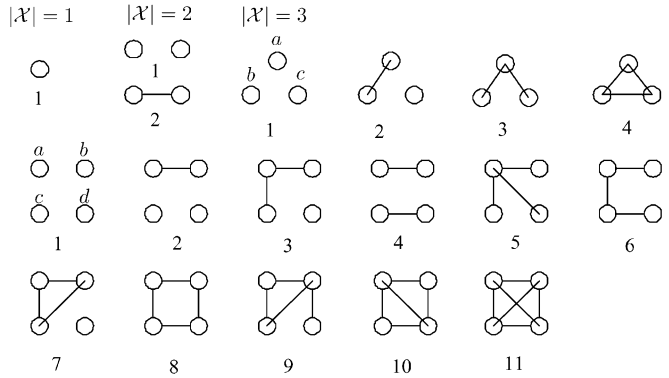

Fig. 6. Possible forms of G_X for $|\mathcal{X}| = 1, 2, 3, 4$.

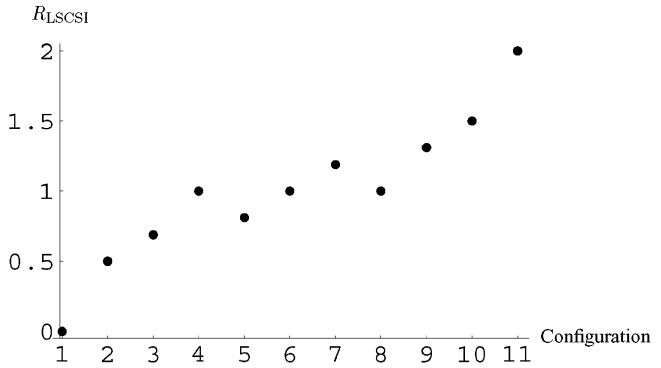

Fig. 7. R_{LSCSI} for the uniform quaternary source. The x-axis corresponds to the category that G_X belongs to, as enumerated in Fig. 6. The two points in the sixth category correspond to the upper and lower bounds given in Table I (which follow from Corollary 2). Corollary 2 implies that the graph is monotone on subsequences of increasing configurations.

as $\beta \to \infty$, where ρ' is induced by the Hamming distortion measure ρ. The explicit expressions of R_{LSCSI} that such optimizations yield, for the various possible forms of the graph G_X, as enumerated in Fig. 6 for alphabet sizes $|\mathcal{X}| = 1, 2, 3, 4$, are given in Table I. A plot of the values of R_{LSCSI} corresponding to the various configurations when X is a uniform quaternary source is given in Fig. 7.

IV. Rate Distortion With Limited Side Information Look-Ahead

Consider the setting of Section II when the causality requirement is relaxed to a look-ahead of $\ell > 0$. Specifically, the ith reconstruction this time is of the form $\hat{X}_i(T, Y^{i+\ell})$. Let $R_\ell(D)$ denote the associated rate distortion function. Note that $R_0(D) = R(D)$, which is the rate distortion function of Section II. For any integer $k \geq 1$, define

$$R_{k,\ell}(D) = \frac{1}{k} \min I(X^k; W) \tag{46}$$

where the minimum is over all functions $f_i : \mathcal{W} \times \mathcal{Y}^{i+\ell} \to \hat{\mathcal{X}}$, $1 \leq i \leq k-\ell$, $|\mathcal{W}| \leq |\mathcal{X}|^k + 1$, and $P(w|x^k)$ such that

$$\frac{1}{k-\ell} \sum_{i=1}^{k-\ell} E[\rho(X_i, f_i(W, Y^{i+\ell}))] \leq D. \tag{47}$$

For $\ell \geq 1$, define $\tilde{R}_\ell(D)$ as

$$\tilde{R}_\ell(D) = \frac{1}{\ell} \min I(X^\ell; W) \tag{48}$$

where the minimum is over all functions $f_i : \mathcal{W} \times \mathcal{Y}^\ell \to \hat{\mathcal{X}}$, $1 \leq i \leq d$, $|\mathcal{W}| \leq |\mathcal{X}|^\ell + 1$, and $P(w|x^\ell)$ such that

$$\frac{1}{\ell} \sum_{i=1}^{\ell} E[\rho(X_i, f_i(W, Y^\ell))] \leq D. \tag{49}$$

The main result of this section is as follows.

Theorem 3: The rate distortion function for ℓ look-ahead $R_\ell(D)$ is bounded for any $k \geq 1, 0 < \ell < \infty$ and $D \geq D_{\min}$, as follows:

$$R_{k,\ell}(D) \leq R_\ell(D) \leq R_{k,\ell}(D) + \frac{\ell}{k} \tilde{R}_\ell(D). \tag{50}$$

Consequently

$$R_\ell(D) = \lim_{k \to \infty} R_{k,\ell}(D). \tag{51}$$

Remarks:

1) Note that $\tilde{R}_\ell(D) \leq H(X)$, thus, (50) implies that
$$R_{k,\ell}(D) \leq R_\ell(D) \leq R_{k,\ell}(D) + \frac{\ell}{k} H(X) \tag{52}$$
which only slightly increases the upper bound when k is moderately larger than ℓ and there is no need to evaluate $\tilde{R}_\ell(D)$.
2) Although the exact form of $R_\ell(D)$ is given by the limit in (51), the more significant point of Theorem 3 is the bounds in (50), or the implied bounds in (52), which enable the computation of $R_\ell(D)$ to any desired accuracy. Specifically, given any $\varepsilon > 0$, $R_{k,\ell}(D)$ is guaranteed to be within ε of $R_\ell(D)$ provided k is taken to satisfy $\ell H(X) \leq k\varepsilon$. $R_{k,\ell}(D)$ for a fixed value of D is then the solution to an optimization problem over $|\mathcal{X}|^{2k}$ real variables, since for each value of x^k, the conditional distribution of W is an element of the $|\mathcal{X}|^k$-dimensional simplex. This optimization problem can be further simplified by searching only for an ε-optimal solution, yielding an overall approximation guaranteed of being within 2ε from $R_\ell(D)$. Note that the minimization in (46) that defines $R_{k,\ell}(D)$, though presented as involving a minimization over the set of possible functions f_i, in addition to $P(w|x^k)$, is in effect a minimization only with respect to $P(w \,|\, x^k)$. The reason is that the optimum f_i associated with a given $P(w \,|\, x^k)$ is readily found as
$$f_i^{\mathrm{opt}}(w, y^{i+\ell}) = \arg\min_{\hat{x}} E[\rho(X_i, \hat{x}) \,|\, w, y^{i+\ell}].$$
This "computability" is in contrast to other instances of both source and channel coding settings where characterizations that are given as limits of finite-dimensional optimization problems offer little guidance on the rate at which the solutions to these problems approach their limits and, hence, do not provide computable approximation procedures of the type implied by Theorem 3.

TABLE I
R_{LSCSI} FOR QUARTERNARY X

$\lvert\mathcal{X}\rvert$	G_X	R_{LSCSI}
1	1	$H(X)=0$
2	1	0
2	2	$H(X)$
3	1	0
3	2	$[P(a)+P(b)]h\left(\frac{P(a)}{P(a)+P(b)}\right)$
3	3	$h\left(P(a)\right)$
3	4	$H(X)$
4	1	0
4	2	$[P(a)+P(b)]\cdot h\left(\frac{P(a)}{P(a)+P(b)}\right)$
4	3	$[1-P(d)]h\left(\frac{P(a)}{1-P(d)}\right)$
4	4	$[P(a)+P(b)]h\left(\frac{P(a)}{P(a)+P(b)}\right)+[P(c)+P(d)]h\left(\frac{P(c)}{P(c)+P(d)}\right)$
4	5	$h\left(P(a)\right)$
4	6	$(P(a)+P(b))h\left(\frac{P(a)}{P(a)+P(b)}\right)+(P(c)+P(d))h\left(\frac{P(c)}{P(c)+P(d)}\right)$ if $P(a)P(c)\le P(b)P(d)$ $h\left(P(b)+P(c)\right)$ if $P(a)P(c)\ge P(b)P(d)$
4	7	$[1-P(d)]H\left(X\lvert X\in\{a,b,c\}\right)$ $=P(a)\log\frac{1-P(d)}{P(a)}+P(b)\log\frac{1-P(d)}{P(b)}+P(c)\log\frac{1-P(d)}{P(c)}$
4	8	$h\left(P(a)+P(d)\right)$
4	9	$I_{\alpha^*}(X;W')$, where $\alpha^*=P(a)/[P(a)+P(c)]$ for $P(w'\lvert x)$ having $\mathcal{W}'=\{w_0,w_1,w_2\}$ and $P(W'=w_0\lvert X=b)=P(W'=w_1\lvert X=a)=P(W'=w_2\lvert X=c)=1$, $P(W'=w_1\lvert X=d)=1-P(W'=w_2\lvert X=d)=\alpha^*$
4	10	$-P(a)\log P(a)-P(d)\log P(d)-[P(b)+P(c)]\log[P(b)+P(c)]$
4	11	$H(X)$

Proof of Theorem 3: We simplify the exposition by assuming $\ell=1$. The proof extends to general ℓ with obvious modifications. Thus, we prove the bounds

$$R_{k,1}(D)\le R_1(D)\le R_{k,1}(D)+\frac{R(D)}{k} \tag{53}$$

where $R(D)$ is the rate distortion function of Section II and

$$R_{k,1}(D)=\frac{1}{k}\min I(X^k;W) \tag{54}$$

where the minimum is over all functions $f_i:\mathcal{W}\times\mathcal{Y}^{i+1}\to\hat{\mathcal{X}}$, $1\le i\le k-1$, $|\mathcal{W}|\le|\mathcal{X}|^k+1$, and $P(w\,|\,x^k)$ such that

$$\frac{1}{k-1}\sum_{i=1}^{k-1}E[\rho(X_i,f_i(W,Y^{i+1}))]\le D. \tag{55}$$

The upper bound $R_1(D)\le R_{k,1}(D)+R(D)/k$ is established using a scheme that encodes successive k blocks. Viewing a k-block as a symbol of the k-super-alphabet, an achievability argument similar to that in the proof of Theorem 1 yields a scheme with look-ahead $\ell=1$, expected distortion upper-bounded by D per symbol on the first $k-1$ symbols of each k block, and rate of $kR_{k,1}(D)$ per k-block. The subsequence consisting of the last symbol in each block is encoded separately at distortion D and rate $R(D)$ (per k-block) using the optimal scheme for the causal case. The outcome is a *bona fide* scheme for look-ahead $\ell=1$ that achieves distortion D at rate $R_{k,1}(D)+R(D)/k$. To prove the lower bound, fix any scheme with block length n and rate R satisfying

$$\frac{1}{n}\sum_{i=1}^{n}E[\rho(X_i,\hat{X}_i)]\le D. \tag{56}$$

Then

$$\begin{aligned}
knR &\ge kH(T)\\
&\ge kI(X^n;T)\\
&= kH(X^n)-k\sum_{i=1}^{n}H(X_i\,|\,T,X^{i-1})\\
&= kH(X^n)-\sum_{j=0}^{k-1}\sum_{i=1-j}^{n-j}H(X_{i+j}\,|\,T,X^{i+j-1})\\
&\overset{(o)}{=} kH(X^n)-\sum_{j=0}^{k-1}\sum_{i=1}^{n}H(X_{i+j}\,|\,T,X^{i+j-1})\\
&\quad+\sum_{j=0}^{k-1}\left[\sum_{i=n-j+1}^{n}H(X_{i+j}\,|\,T,X^{i+j-1})\right.\\
&\quad\left.-\sum_{i=1-j}^{0}H(X_{i+j}\,|\,T,X^{i+j-1})\right]
\end{aligned}$$

$$
\begin{aligned}
&\overset{(a)}{=} kH(X^n) - \sum_{j=0}^{k-1}\sum_{i=1}^{n} H(X_{i+j}\,|\,T, X^{i+j-1}) \\
&\quad + \sum_{j=0}^{k-1}\left[\sum_{i=n-j+1}^{n} H(X_{i+j}) \right. \\
&\quad \left. - \sum_{i=1-j}^{0} H(X_{i+j}\,|\,T, X^{i+j-1})\right] \\
&= kH(X^n) - \sum_{j=0}^{k-1}\sum_{i=1}^{n} H(X_{i+j}\,|\,T, X^{i+j-1}) \\
&\quad + \sum_{j=0}^{k-1}\sum_{i=1-j}^{0}\left[H(X_{i+j}) - H(X_{i+j}\,|\,T, X^{i+j-1})\right] \\
&= kH(X^n) - \sum_{j=0}^{k-1}\sum_{i=1}^{n} H(X_{i+j}\,|\,T, X^{i+j-1}) \\
&\quad + \sum_{j=0}^{k-1}\sum_{i=1-j}^{0} I(X_{i+j}; T, X^{i+j-1}) \\
&\geq \sum_{i=1}^{n} H\left(X_i^{i+k-1}\right) - \sum_{i=1}^{n}\sum_{j=0}^{k-1} H(X_{i+j}\,|\,T, X^{i+j-1}) \\
&= \sum_{i=1}^{n} H\left(X_i^{i+k-1}\right) - \sum_{i=1}^{n} H\left(X_i^{i+k-1}|T, X^{i-1}\right) \\
&= \sum_{i=1}^{n} I\left(X_i^{i+k-1}; T, X^{i-1}\right) \\
&\overset{(b)}{\geq} \sum_{i=1}^{n} I\left(X_i^{i+k-1}; W_i\right) \\
&\overset{(c)}{\geq} \sum_{i=1}^{n} kR_{k,1}\left(\frac{1}{k-1}\sum_{j=0}^{k-2} E[\rho(X_{i+j}, \hat{X}_{i+j})]\right) \\
&\overset{(d)}{\geq} knR_{k,1}\left(\frac{1}{n}\sum_{i=1}^{n}\frac{1}{k-1}\sum_{j=0}^{k-2} E[\rho(X_{i+j}, \hat{X}_{i+j})]\right) \\
&\overset{(e)}{\geq} knR_{k,1}\left(\frac{kD_{\max}}{n} + \frac{1}{n}\sum_{i=1}^{n} E[\rho(X_i, \hat{X}_i)]\right) \\
&\overset{(f)}{\geq} knR_{k,1}\left(D + \frac{kD_{\max}}{n}\right)
\end{aligned}
$$

(o) follows when interpreting a summation running from $n-j+1$ to n as a null summation whenever $n-j+1 > n$, where or $j < 1$. Similarly, when $j < 1$, the second summation from $1-j$ to 0 is to be understood as 0;

(a) follows since $H(X_t|T, X^{t-1}) = H(X_t)$ for $t > n$;

(b) follows by setting $W_i = (T, Y^{i-1})$, as in the proof of Theorem 1, and using the Markov relation $X_i^{i+k-1} \to (T, X^{i-1}) \to (T, Y^{i-1})$;

(c) follows from i) $\hat{X}_{i+j} = \hat{X}_{i+j}(W_i, Y_i^{i+j+1})$, for $1 \leq i \leq n$ and $0 \leq j \leq k-2$ (where $\hat{X}_{n+1}^{n+k} = (\hat{X}_{n+1}, \ldots, \hat{X}^{n+k})$ arbitrarily), ii) the Markov relation $W_i \to X_i^{i+k-1} \to Y_i^{i+k-1}$, (iii) the definition of $R_{k,1}(D)$, and (iv) the fact that, by Carathéodory's theorem, the restriction to $|\mathcal{W}| \leq |\mathcal{X}|^k + 1$ does not affect the minimum in (54);

(d) follows from the convexity of $R_{k,1}(D)$, which can be proved in the same way as the proof of convexity of $R(D)$ given in the Appendix;

(e) follows from the monotonicity of $R_{k,1}(D)$ by denoting $D_{\max} = E[\max_{\hat{x}} \rho(X, \hat{x})]$;

(f) follows from the monotonicity of $R_{k,1}(D)$ and (56).

Thus, we have $R \geq R_{k,1}(D + kD_{\max}/n)$ for all n, which implies that $R \geq R_{k,1}(D)$ by the continuity of $R_{k,1}(D)$ (which follows by a similar argument to that for $R(D)$ in Section II). □

A. Process Characterization of $R_\ell(D)$

For jointly stationary processes $\boldsymbol{X} = \{X_i\}$ and $\boldsymbol{W} = \{W_i\}$ let $\bar{I}(\boldsymbol{X}; \boldsymbol{W})$ denote the mutual information rate

$$\bar{I}(\boldsymbol{X}; \boldsymbol{W}) = \lim_{n\to\infty} \frac{1}{n} I(X^n; W^n). \tag{57}$$

The following theorem gives a "process characterization" for $R_\ell(D)$.

Theorem 4:

$$R_\ell(D) = \inf\left\{\bar{I}(\boldsymbol{X}; \boldsymbol{W}) : E\left[\rho\left(X_0, \hat{X}_0^{\text{opt}}\left(\boldsymbol{W}, Y_{-\infty}^{\ell}\right)\right)\right] D\right\} \tag{58}$$

where the infimum is over jointly stationary $\boldsymbol{W}, \boldsymbol{X}, \boldsymbol{Y}$ with the Markov relation $\boldsymbol{W} \to \boldsymbol{X} \to \boldsymbol{Y}$ (consistent with the given distribution of $\boldsymbol{X}, \boldsymbol{Y}$), and $\hat{X}_0^{\text{opt}}(\boldsymbol{W}, Y_{-\infty}^{\ell})$ is the optimum estimate of X_0 based on $\boldsymbol{W}, Y_{-\infty}^{\ell}$, namely

$$\hat{X}_0^{\text{opt}}\left(\boldsymbol{w}, y_{-\infty}^{\ell}\right) = \arg\min_{\hat{x}} E\left[\rho(X_0, \hat{x})\,|\,\boldsymbol{W} = \boldsymbol{w}, Y_{-\infty}^{\ell} = y_{-\infty}^{\ell}\right]. \tag{59}$$

It will be clear from the proof, given in Appendix III, that the alphabet of the components of the process $\boldsymbol{W}$ can be restricted to $|\mathcal{W}| \leq |\mathcal{X}| + 1$ without affecting the infimum in (58). Our proof will apply to any $0 \leq \ell < \infty$. Also, if we let $R_\infty(D)$ be the function defined in (58) with ℓ replaced by ∞, we obtain

$$R_\infty(D) = R_{\text{WZ}}(D). \tag{60}$$

This result can be proved by a simple generalization of the arguments used to prove the analogous characterization for the classical rate distortion function [13], [14].

Gaussian $\boldsymbol{W}$ is Ineffectual for Gaussian X, Y

Though the expression in (58) is no more explicit than the limit in (51), it may be useful for deriving upper bounds on $R_\ell(D)$ by considering specific processes $\boldsymbol{W}, \boldsymbol{X}, \boldsymbol{Y}$ from the feasible set for the infimum in (58). In particular, consider the case where $\boldsymbol{X}, \boldsymbol{Y}$ are i.i.d. drawings of the Gaussian pair X, Y and ρ is the squared error distortion, as described in Section II. Motivated by Theorem 4 (and the fact that it can be shown to hold for general alphabets via the tools of [41]), consider the following upper bound on $R_\ell(D)$:

$$R_\ell^G(D) = \inf\left\{\bar{I}(\boldsymbol{X}; \boldsymbol{W}) : E\left(X_0 - \hat{X}_0^{\text{opt}}\left(\boldsymbol{W}, Y_{-\infty}^{\ell}\right)\right)^2 \leq D\right\} \tag{61}$$

where the infimum is over jointly stationary and *Gaussian* $\boldsymbol{W}, \boldsymbol{X}, \boldsymbol{Y}$ with the Markov relation $\boldsymbol{W} \to \boldsymbol{X} \to \boldsymbol{Y}$, and

$\hat{X}_0^{\rm opt}(\boldsymbol{W}, Y_{-\infty}^{\ell})$ is the optimum (linear) estimate of X_0 based on $\boldsymbol{W}, Y_{-\infty}^{\ell}$. Perhaps somewhat surprisingly, $\underline{R}_{\ell}^{G}(D)$ is trivial in the following sense.

Theorem 5: For every $\ell \geq 0$

$$\underline{R}_{\ell}^{G}(D) = \underline{R}_{\rm ub}(D) \tag{62}$$

where $\underline{R}_{\rm ub}(D)$ is the lower convex envelope of $R_{\rm ub}(D)$, the upper bound on $R(D)$ derived in Section II.

Thus, for $\ell > 0$, in order to obtain upper bounds on $R_{\ell}(D)$ that are better (smaller) than our upper bound for the case $\ell = 0$, one must search outside the set of jointly stationary and Gaussian $\boldsymbol{W}, \boldsymbol{X}, \boldsymbol{Y}$. The proof of Theorem 5 is given in Appendix IV.

V. Lossless Source Coding With Side Information Look-Ahead

Consider the same case as in the previous section when encoding is performed by mapping the sequence X^n into $T \in \{1, \ldots, 2^{nR}\}$, and reconstruction is of the form $\hat{X}_i(T, Y^{i+\ell})$. A rate R is achievable if there exists a sequence of schemes of rate R for which $P(X^n \neq \hat{X}^n) \longrightarrow 0$. Let $R_{\rm LSCSI}^{\ell}$ be the infimum over achievable rates for this problem. Thus, $R_{\rm LSCSI}^{0} = R_{\rm LSCSI}$, which we characterized in Section III.

As we have seen in Section III, under the positivity condition, $R_{\rm LSCSI}^{0} = H(X)$ (and not $H(X \,|\, Y)$), which is achievable when no delay constraint is imposed. Thus, it is perhaps natural to expect $R_{\rm LSCSI}^{\ell}$ to decrease with increasing ℓ toward $H(X \,|\, Y)$. This will turn out to be false. In fact, we will show, under the positivity condition, that not only is $\lim_{\ell \to \infty} R_{\rm LSCSI}^{\ell} > H(X \,|\, Y)$ in general, but it is equal to $H(X)$ for all $0 \leq \ell < \infty$. That is, side information for any $\ell < \infty$ is useless.

We first give a general characterization of $R_{\rm LSCSI}^{\ell}$. Let

$$R_{\rm LSCSI}^{k,\ell} = \frac{1}{k} \min I(X^k; W) \tag{63}$$

where the minimum is over all $P(w \,|\, x^k), |\mathcal{W}| \leq |\mathcal{X}|^k + 1$, such that

$$H(X_i \,|\, W, Y^{i+\ell}) = 0, \qquad \text{for all } 1 \leq i \leq k - \ell. \tag{64}$$

Theorem 6: For every k and ℓ

$$R_{\rm LSCSI}^{k,\ell} \leq R_{\rm LSCSI}^{\ell} \leq R_{\rm LSCSI}^{k,\ell} + \frac{\ell}{k} R_{\rm LSCSI}. \tag{65}$$

So, in particular

$$R_{\rm LSCSI}^{\ell} = \lim_{k \to \infty} R_{\rm LSCSI}^{k,\ell}. \tag{66}$$

Proof of Theorem 6: Considering $R_{k,\ell}(D)$ of the previous section with Hamming loss, it is clear that $R_{\rm LSCSI}^{k,\ell}$, as defined in (63), is nothing but $R_{k,\ell}(0)$. The fact that

$$R_{\rm LSCSI}^{k,\ell} = R_{k,\ell}(0) \leq R_{\rm LSCSI}^{\ell}$$

follows by Theorem 3. To prove the right-hand side inequality, define

$$\tilde{R}_{\rm LSCSI}^{k,\ell} = \frac{1}{k} \min I(X^k; W) \tag{67}$$

where the minimum is over all $P(w|x^k), |\mathcal{W}| \leq |\mathcal{X}|^k + 1$, such that

$$H(X_i \,|\, W, Y^{i+\ell}) = 0, \qquad \text{for all } 1 \leq i \leq k - \ell \tag{68}$$

and

$$H\left(X_{k-\ell+1}^{k} \,|\, W, Y_{k-\ell+1}^{k}\right) = 0. \tag{69}$$

Achievability of $\tilde{R}_{\rm LSCSI}^{k,\ell}$ follows via a typical sequence argument similar to that used in the proof of Theorem 2 applied to k-block super symbols. Thus, it only remains to establish that $\tilde{R}_{\rm LSCSI}^{k,\ell} \leq R_{\rm LSCSI}^{k,\ell} + \frac{\ell}{k} R_{\rm LSCSI}$, which is proved in Appendix V. □

In Appendix VI we prove the following simple fact.

Lemma 3: Let W, X, Y, Y' be discrete random variables with the Markov relation $W \to X \to (Y, Y')$, and assume that the pairs X, Y and X, Y' have the same bipartite graphs, that is, for every $x, y, P(X = x, Y = y) = 0$ if and only if $P(X = x, Y' = y) = 0$. Then, for any (deterministic) function $f, H(f(X) \,|\, W, Y) = 0$ if and only if $H(f(X) \,|\, W, Y') = 0$.

Lemma 3 combined with Theorem 6 imply the following.

Proposition 2: $R_{\rm LSCSI}^{\ell}$ depends on the distribution of the pair X, Y only through the distribution of X and the bipartite graph whose edges are the pairs (x, y) for which $P(x, y) > 0$.

Proof: Consider $R_{\rm LSCSI}^{k,\ell}$ as defined in (63). Applying Lemma 3 for each $1 \leq i \leq k - \ell$ after replacing X by X^k, Y by $Y^{i+\ell}$, and $f(X)$ by X_i, implies that whether or not the constraint (64) is satisfied depends solely on the distribution of X and on the bipartite graphs associated with the pairs $(X^k, Y^{i+\ell})$ and $1 \leq i \leq k - \ell$. The latter graphs, however, are determined by the original bipartite graph of X, Y. Thus, for every k, $R_{\rm LSCSI}^{k,\ell}$ depends on the distribution of the pair X, Y only through the distribution of X and the latter graph, which when combined with Theorem 6 completes the proof. □

Note that although Theorem 3 yields a computable characterization of $R_{\rm LSCSI}^{\ell}$, it does not lead to a "single-letter" characterization for the general case. However, under the positivity condition and when combined with Proposition 2, it reduces to such a characterization.

Corollary 3: Assume that X, Y satisfy the positivity condition, i.e., $P(x, y) > 0$ for all $x \in \mathcal{X}, y \in \mathcal{Y}$. Then, for any $0 \leq \ell < \infty$, $R_{\rm LSCSI}^{\ell} = H(X)$.

Proof: Assume that X_i, Y_i, instead of being generated i.i.d. according to X, Y, are generated according to X, Y', where Y' is independent of X and satisfies $Y' \stackrel{d}{=} Y$. Clearly, $R_{\rm LSCSI}^{\ell}(P_{X,Y'}) = H(X)$. On the other hand, by Proposition 2, $R_{\rm LSCSI}^{\ell}(P_{X,Y}) = R_{\rm LSCSI}^{\ell}(P_{X,Y'})$, since $P_{X,Y}$ and $P_{X,Y'}$ have

the same X-marginal and both satisfy the positivity condition (so, in particular, induce the same bipartite graph). □

Block-Length-Dependent Look-Ahead

For a look-ahead sequence $\{\ell_n\}$, let $R_{\mathrm{LSCSI}}^{\{\ell_n\}}$ denote the infimum over achievable rates when $\ell = \ell_n$, i.e., allowing the look-ahead to depend on the block length n. As we have seen, under the positivity condition, if $\ell_n = \ell$ for all n, $R_{\mathrm{LSCSI}}^{\{\ell_n\}} = H(X)$. On the other hand, $\ell_n = n$ corresponds to the setting of [32], so in this case $R_{\mathrm{LSCSI}}^{\{\ell_n\}} = H(X\,|\,Y)$. As it turns out, $R_{\mathrm{LSCSI}}^{\{\ell_n\}} = H(X\,|\,Y)$ remains true so long as the rate of increase of ℓ_n with n is faster than logarithmic. Furthermore, any rate $R > H(X\,|\,Y)$ is achievable so long as $\ell_n = C(R)\log n$, for a sufficiently large constant $C(R)$. To this end, define

$$E(R) = \min_{Q_{X,Y}} [D(Q_{X,Y} \,\|\, P_{X,Y}) + \max\{0, R - H_Q(X\,|\,Y)\}] \tag{70}$$

where $P_{X,Y}$ denotes the distribution of X, Y and $H_Q(X\,|\,Y)$ is the conditional entropy under $Q_{X,Y}$. As is implied by [5, Example 3.1.5, p. 264], this can be considered a "random coding error exponent" for the Slepian–Wolf problem.

Theorem 7: For every $R > H(X\,|\,Y)$, $R_{\mathrm{LSCSI}}^{\{\ell_n\}} \le R$ provided $\ell_n = C\cdot\log n$ and $C > 1/E(R)$. In particular, $R_{\mathrm{LSCSI}}^{\{\ell_n\}} = H(X\,|\,Y)$ whenever the increase of ℓ_n is faster than logarithmic.

Proof: From [5, Example 3.1.5, p. 264], it follows that when $R > H(X\,|\,Y)$, not only does there exist a sequence of schemes (for the Slepian–Wolf problem with no delay constraint on the side information) with $P_e \to 0$, but also $P_e \le \exp\{-n(E(R) + o(1))\}$, where $E(R) > 0$ is given by (70). In particular, for any $\varepsilon > 0$ and sufficiently large n

$$P_e \le \exp\{-n(E(R) - \varepsilon)\}. \tag{71}$$

Consider now the probability of error of the rate R scheme with look-ahead ℓ for block length n formed by concatenating n/ℓ Slepian–Wolf schemes[4] of block length ℓ. Choosing a good scheme for ℓ-blocks, the probability of error for each of the n/ℓ subblocks of length ℓ, by (71), can be upper-bounded by $P_e \le \exp(-\ell(E(R) - \varepsilon))$. Hence, the probability of error of the resulting n-block scheme is upper-bounded by $\frac{n}{\ell}\exp(-\ell(E(R) - \varepsilon))$, which converges to zero for $\ell = \ell_n = C\log n$ when, for example, $C = \varepsilon + 1/(E(R) - \varepsilon)$. This completes the proof by the arbitrariness of $\varepsilon > 0$. □

Remarks:

- From Theorem 7, for any R where $E(R)$ is strictly increasing, $R_{\mathrm{LSCSI}}^{\{\ell_n\}} \le R$ for $\ell_n = \log n/E(R)$, that is, $C = C(R)$ can be taken to be $1/E(R)$. This follows from the fact that, for any $\varepsilon > 0$, $E(R) < E(R+\varepsilon)$. Thus, Theorem 7 implies that $R_{\mathrm{LSCSI}}^{\{\ell_n\}} \le R+\varepsilon$ for $\ell_n = \log n/E(R)$, which implies that $R_{\mathrm{LSCSI}}^{\{\ell_n\}} \le R$ by the arbitrariness of ε.
- Theorem 7 and its proof remain valid when $E(R)$ is the true error exponent (which is at least as large as the random coding one) rather than the random coding error exponent of (70). The former is unknown, however.
- Theorem 7 only addresses achievability. Whether or not a logarithmic growth rate of ℓ_n is also *necessary* to achieve rates $R \in (H(X\,|\,Y), H(X))$ (say under the positivity condition) remains to be determined.

Bridging the Gap Between Lossless and Lossy Results

Let $R_\ell(D)$ denote the rate distortion function for look-ahead ℓ from Section IV, when ρ is Hamming loss. Assuming X, Y satisfy the positivity condition, we have seen in this section (in proof of Theorem 6) that

$$R_\ell(0) = H(X), \qquad \text{for all } 0 \le \ell < \infty \tag{72}$$

whereas, as observed in [42]

$$R_{\mathrm{WZ}}(0) = H(X\,|\,Y). \tag{73}$$

On the other hand

$$\lim_{\ell\to\infty} R_\ell(D) = R_{\mathrm{WZ}}(D), \qquad \text{for all } D > 0. \tag{74}$$

To see this formally, consider the characterization in Theorem 4. Specifically, combining (58) and (60) with the fact that, by the martingale convergence theorem (cf., in particular, [3, Theorem 5.21])

$$E\left[\rho\left(X_0, \hat{X}_0^{\mathrm{opt}}(\boldsymbol{W}, Y_{-\infty}^{\ell})\right)\right] \le E\left[\rho\left(X_0, \hat{X}_0^{\mathrm{opt}}(\boldsymbol{W}, Y_{-\infty}^{\infty})\right)\right] + \varepsilon$$

for all $\varepsilon > 0$ and sufficiently large ℓ, we obtain for every $D \ge 0$ and $\varepsilon > 0$

$$\lim_{\ell\to\infty} R_\ell(D+\varepsilon) \le R_{\mathrm{WZ}}(D). \tag{75}$$

This implies (74) by the arbitrariness of $\varepsilon > 0$ and the continuity of $R_{\mathrm{WZ}}(D)$ for $D > 0$ (which is guaranteed by its convexity). Note, in particular, that this implies a sensitivity to the order of the limits

$$H(X\,|\,Y) = \lim_{D\downarrow 0}\lim_{\ell\to\infty} R_\ell(D) < \lim_{\ell\to\infty}\lim_{D\downarrow 0} R_\ell(D) = H(X). \tag{76}$$

A sketch of this situation is given in Fig. 8.

VI. Channel Coding With Limited Look-Ahead State Information at the Encoder

Consider the problem of coding for a memoryless channel with state information known at the encoder, as considered by Shannon [31] for the case where the states are only causally known, and by Gel'fand and Pinsker [12] and Heegard and

[4]Neglecting the inconsequential edge effect when n/ℓ is not an integer.

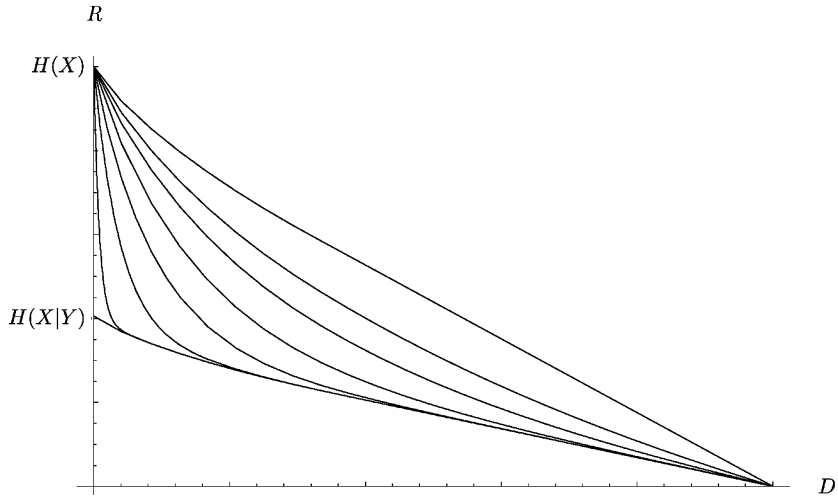

Fig. 8. Sketch of the form of the rate distortion functions under Hamming distortion. The upper curve represents $R(D) = R_0(D)$. The lower curve represents $R_{\mathrm{WZ}}(D)$. The curves in between represent $R_\ell(D)$ (approaching $R_{\mathrm{WZ}}(D)$ with increasing ℓ). Though $R_\ell(0) = H(X)$ for all $0 \le \ell < \infty$ and $R_{\mathrm{WZ}}(0) = H(X\,|\,Y)$, $\lim_{\ell\to\infty} R_\ell(D) = R_{\mathrm{WZ}}(D)$ for all $D > 0$. In particular, $H(X\,|\,Y) = \lim_{D\downarrow 0}\lim_{\ell\to\infty} R_\ell(D) < \lim_{\ell\to\infty}\lim_{D\downarrow 0} R_\ell(D) = H(X)$.

El Gamal [15] for the case where all the states are known in advance. In this section, we consider the same problem when the restriction to causality in the setting of [31] is relaxed to some positive look-ahead (cf. [6], [9] and references therein for the significance of this problem). We establish a computable characterization of capacity in this case, which is analogous to the characterization of $R_\ell(D)$ given in Section IV.

Consider a memoryless channel with transition probability $P(y|x,s)$, where X is the input and S is the state. To simplify the derivation below, we assume that the alphabets of the channel input $\mathcal{X}$, channel output $\mathcal{Y}$, and state $\mathcal{S}$ are finite. For a message index $W \in \{1,\ldots,2^{nR}\}$, the ith-channel input is of the form $X_i(W, S^{i+\ell})$, where $S_1, S_2, \ldots, S_i$ are i.i.d. with PMF $P(s)$. Decoding is based only on the channel output Y^n. The following theorem provides a computable characterization for the capacity of this channel.

Theorem 8: Let

$$C_{\ell,k} = \frac{1}{k}\max I(U;Y^k) \tag{77}$$

where the max is over all $P(u\,|\,s^\ell)$

$$|\mathcal{U}| \le \min\{|\mathcal{X}|^k, |\mathcal{Y}|^k\} + |\mathcal{S}|^k - 1$$

and f_i, $1 \le i \le k$, such that $X_i = f_i(U, S^{i+\ell})$. Then

$$\frac{k}{k-\ell}C_{\ell,k} - \frac{\ell\log|\mathcal{Y}|}{k-\ell} \le C_\ell \le C_{\ell,k}. \tag{78}$$

In particular

$$C_\ell = \lim_{k\to\infty} C_{\ell,k}. \tag{79}$$

Remarks:

1) The significant point here, as in Theorem 3, is not the limit in (79), but rather the bounds in (78), which enable the computation of C_ℓ to any desired accuracy.
2) The lower bound can be refined to obtain an expression involving the capacity of the Shannon channel (with causal state information) associated with a k super-symbol, in lieu of the $\log|\mathcal{Y}|$ term appearing in (78). Such a bound would be useful for channels where $|\mathcal{Y}|$ is large or infinite.

Proof of Theorem 8: To simplify the exposition, we only give the proof for $\ell = 1$. The proof extends to general ℓ with obvious modifications. To recap, we let

$$C_{1,k} = \frac{1}{k}\max I(U;Y^k) \tag{80}$$

where the max is over all $P(u\,|\,s_1)$

$$|\mathcal{U}| \le \min\{|\mathcal{X}|^k, |\mathcal{Y}|^k\} + |\mathcal{S}|^k - 1$$

and f_i, $1 \le i \le k$, such that $X_i = f_i(U, S^{i+1})$. We need to prove that

$$\frac{k}{k-1}C_{1,k} - \frac{1}{k-1}\log|\mathcal{Y}| \le C_1 \le C_{1,k}. \tag{81}$$

We first establish the achievability of $\frac{k}{k-1}C_{1,k} - \frac{1}{k-1}\log|\mathcal{Y}|$. Define

$$\tilde{C}_{1,k} = \frac{1}{k}\max I(U;Y^k) \tag{82}$$

where the max is over all $P(u)$ and $f_i, 1 \le i \le k$, such that $X_i = f_i(U, S^{i+1})$. Note that the difference between $\tilde{C}_{1,k}$ and $C_{1,k}$ is that in the latter U is allowed to depend on S_1. The achievability of $\tilde{C}_{1,k}$ follows from the same argument used by Shannon in [31] to establish achievability of C_0, applied to k-letter super-symbols. Thus, for every k, $\tilde{C}_{1,k} \le C_1$. On the other hand, considering the joint distribution that achieves the maximum in (80)

$$\begin{aligned} kC_{1,k} &= I(U;Y^k) \\ &\le I\left(U;Y_2^k\right) + \log|\mathcal{Y}| \\ &\le (k-1)\tilde{C}_{1,k-1} + \log|\mathcal{Y}| \end{aligned}$$

where the last inequality follows since (U, S_1) and S_2^k are independent. Now, letting $\tilde{U} = (U, S_1)$, it follows that

$(\tilde{U}, S_2^k, X_2^k, Y_2^k)$ is in the feasible set of the maximization associated with $\tilde{C}_{1,k-1}$. We thus have

$$\frac{k}{k-1}C_{1,k} - \frac{1}{k-1}\log|\mathcal{Y}| \le \tilde{C}_{1,k-1} \le C_1,$$

which proves the left inequality in (81). Turning to the right inequality in (81), we fix a scheme for block length n and define $U_i = (W, Y^{i-1}, S^{i-1})$. Then

$$\begin{aligned}
kI(W;Y^n) &= kH(Y^n) - kH(Y^n|W) \\
&= kH(Y^n) - \sum_{j=0}^{k-1}\sum_{i=-j+1}^{n-j} H(Y_{i+j}|Y^{i+j-1}, W) \\
&\overset{(a)}{=} kH(Y^n) - \sum_{j=0}^{k-1}\sum_{i=-j+1}^{n} H(Y_{i+j}|Y^{i+j-1}, W) \\
&\le kH(Y^n) - \sum_{j=0}^{k-1}\sum_{i=1}^{n} H(Y_{i+j}|Y^{i+j-1}, W) \\
&= kH(Y^n) - \sum_{i=1}^{n}\sum_{j=0}^{k-1} H(Y_{i+j}|W, Y^{i+j-1}) \\
&\overset{(b)}{\le} \sum_{j=0}^{k-1} \sum_{\substack{-k+1\le i\le n+k-1:\\ i\equiv j \bmod k}} H\left(Y_i^{i+k-1}\right) \\
&\quad - \sum_{i=1}^{n} H\left(Y_i^{i+k-1}|W, Y^{i-1}\right) \\
&= \sum_{i=1}^{n} H\left(Y_i^{i+k-1}\right) - \sum_{i=1}^{n} H\left(Y_i^{i+k-1}|W, Y^{i-1}\right) \\
&= \sum_{i=1}^{n} I\left(Y_i^{i+k-1}; W, Y^{i-1}\right) \\
&\le \sum_{i=1}^{n} I\left(Y_i^{i+k-1}; W, Y^{i-1}, S^{i-1}\right) \\
&= \sum_{i=1}^{n} I\left(Y_i^{i+k-1}; U_i\right) \\
&\overset{(c)}{\le} \sum_{i=1}^{n} kC_{1,k} \\
&= nkC_{1,k}
\end{aligned}$$

where

(a) follows by defining Y_m arbitrarily (but deterministically) for m outside the range $\{1, \ldots, n\}$, e.g., by taking

$$H(Y_m \,|\, Y^{m-1}, W) = 0$$

for such m;

(b) follows since, for any $0 \le j \le k-1$,

$$\begin{aligned}
H(Y^n) &= \sum_{\substack{-k+1\le i\le n+k-1:\\ i\equiv j \bmod k}} H\left(Y_i^{i+k-1} \,|\, Y^{i-1}\right) \\
&\le \sum_{\substack{-k+1\le i\le n+k-1:\\ i\equiv j \bmod k}} H\left(Y_i^{i+k-1}\right);
\end{aligned}$$

(c) follows by noting that, for $0 \le j \le k-1, X_{i+j} = X_{i+j}(U_i, S_i^{i+j+1})$, and (U_i, S_i) is independent of S_{i+1}^n (thus, (U_i, S_i) is independent of S_{i+1}^{i+k}). However, U_i and S_i can be dependent, since X_{i-1} can depend on S_i. Consequently, $U_i, X_i^{i+k-1}, Y_i^{i+k-1}, S_i^{i+k}$ satisfy the dependence structure of U, X^k, Y^k, S^{k+1} over which the maximization in (80) is taken. It then only remains to note that the constraint on the cardinality of $\mathcal{U}$ does not affect the value of the maximum. This follows by applying the argument that $|\mathcal{U}| \le \min\{|\mathcal{X}|, |\mathcal{Y}|\} + |\mathcal{S}| - 1$ does not affect the expressions of [31], [12], [15] to k super-symbols.

We have thus obtained

$$H(W \,|\, Y^n) \ge H(W) - nC_{1,k} = n(R - C_{1,k})$$

which establishes the converse part of the proof by Fano's inequality. □

VII. Conclusion

We characterized the rate distortion function for the source coding with decoder side information setting when the ith reconstruction symbol is allowed to depend only on the first $i + \ell$ side information symbols, for some finite look-ahead ℓ, in addition to the index from the encoder. For the case of causal side information, we found that the penalty of causality is the omission of the subtracted mutual information term in the Wyner–Ziv rate distortion function. For $\ell > 0$, we derived a computable expression for the rate distortion function. When specialized to the near-lossless case, our results were seen to characterize the best achievable rate for the Slepian–Wolf source coding problem with limited look-ahead side information. We found that for any ℓ, side information is useless when the joint pmf of the source and side information satisfy the positivity condition $P(x, y) > 0$ for all (x, y). On the other hand, we have seen that $H(X \,|\, Y)$ is achievable provided side information look-ahead ℓ_n is allowed to grow faster than logarithmic in the block length n. Finally, we applied our approach to derive a computable expression for channel capacity when state information is available to the sender with limited look-ahead.

Following are some open questions related to the work presented in this paper.

- For X, Y Gaussian and $\ell = 0$ (i.e., causal side information), is the rate distortion function given by the convexified version of $R_{\mathrm{ub}}(D)$, the bound attained when considering jointly Gaussian W, X, Y? Note that $R_{\mathrm{ub}}(D)$ is, in general, not convex (recall Lemma 1), so we already know that jointly Gaussian W, X, Y do not always attain the rate distortion function. Theorem 5 is perhaps another hint for the suboptimality of the Gaussian distribution in this context.
- Recall from Section V that $\lim_{\ell\to\infty} R_\ell(D) = R_{\mathrm{WZ}}(D)$ for $D > 0$, under Hamming loss. The argument used is readily seen to imply that this is true under any distortion measure, provided $D > D_{\min}$. It would be interesting to study the rate of this convergence. The techniques for computing the redundancy of rate distortion codes [27], [23], [43] should be extendable to Wyner–Ziv codes, leading to an upper bound $R_\ell(D) - R_{\mathrm{WZ}}(D) = O(\frac{1}{\ell}\log\ell)$, since

the redundancy of a Wyner–Ziv block code of length ℓ upper-bounds the difference $R_\ell(D) - R_{\mathrm{WZ}}(D)$. Proving a lower bound of matching order may be more challenging.

- Consider the lossless case. Assuming for simplicity the positivity condition, we have seen that $R^{\ell}_{\mathrm{LSCSI}} = H(X)$ for all ℓ, while, for every $R \in (H(X\,|\,Y), H(X)), R^{\{\ell_n\}}_{\mathrm{LSCSI}} \le R$ provided $\ell_n = C(R) \cdot \log n$ for a sufficiently large R-dependent constant $C(R)$. It would be interesting to determine whether or not a logarithmic growth rate is also necessary to achieve rates in $(H(X\,|\,Y), H(X))$ and, if so, to characterize the function $\alpha(R)$ for which $R^{\{\ell_n\}}_{\mathrm{LSCSI}} = R$ when $\ell_n = \alpha(R) \cdot \log n$.
- The fourth remark following the proof of Theorem 1 together with the observations in [25] establish that feedback/feedforward does not improve on the fundamental limits for the Gel'fand–Pinsker channel, the Shannon channel [31], the Wyner–Ziv source, and source coding with causal decoder side information (as considered in Section II). It would be interesting to determine whether the presence of feedback/feedforward can improve on the fundamental limits in the respective settings of Section VI and Section IV for positive ℓ. A negative answer would allow us to use finite-look-ahead adaptations of the schemes in [25] (in particular, for the Gaussian case, those based on the Schalkwijk–Kailath scheme [29], [28]) to deduce respective lower bounds on the capacity C_ℓ of Section VI and upper bounds on the rate distortion function $R_\ell(D)$ of Section IV.

Appendix I
Convexity of $R(D)$

To prove convexity of $R(D)$ in (1) we use the same approach as that used in [4, Lemma 14.9.1] for the the Wyner–Ziv function.

Let D_1 and D_2 be two distortion values and let W_1, f_1 and W_2, f_2 be the corresponding achievers of the minima in $R(D_1)$ and $R(D_2)$. Let Q be independent of X, Y, W_1, W_2, taking on the value 1 with probability λ and the value 2 with probability $1 - \lambda$. Define $W = (Q, W_Q)$ and let $\hat{X} = f(W, Y) = f_Q(W_Q, Y)$. Then

$$\begin{aligned} &E[\rho(X, \hat{X})] \\ &\quad = \lambda E[\rho(X, f_1(W_1, Y_1))] + (1-\lambda) E[\rho(X, f_2(W_2, Y_2))] \\ &\quad = \lambda D_1 + (1-\lambda) D_2. \end{aligned}$$

Thus,

$$\begin{aligned} &R(\lambda D_1 + (1-\lambda) D_2) \\ &\quad \le I(X; W) \\ &\quad = H(X) - H(X\,|\,W) \\ &\quad = H(X) - H(X|W_Q, Q) \\ &\quad = H(X) - \lambda H(X\,|\,W_1) - (1-\lambda) H(X|W_2) \\ &\quad = \lambda I(X; W_1) + (1-\lambda) I(X; W_2) \\ &\quad = \lambda R(D_1) + (1-\lambda) R(D_2). \end{aligned}$$

Appendix II
Proof of Lemma 2

Condition (45) implies that for each y and any x with $P(x\,|\,y) > 0$, if $P(w\,|\,x) > 0$ then $P(w\,|\,x') = 0$ for any other x' with $P(x'\,|\,y) > 0$. Thus, Y and W uniquely determine X and hence $H(X\,|\,W, Y) = 0$. For the reverse implication, take $(x, x') \in G_X, w$ satisfying $P(w\,|\,x) > 0$, and y satisfying $P(y\,|\,x) > 0$ and $P(y\,|\,x') > 0$, then

$$P(y, w) \ge P(w, x, y) = P(w\,|\,x) P(x, y) > 0 \tag{A1}$$

and

$$P(x|y, w) \ge P(x, y, w) = P(w\,|\,x) P(x) P(y\,|\,x) > 0. \tag{A2}$$

Now $H(X\,|\,WY) = 0$ implies by (A1) that $H(X\,|\,W = w, Y = y) = 0$, Thus, by (A2), $P(x|y, w) = 1$, which implies that $P(x'|y, w) = 0$. But

$$P(x'|y, w) \ge P(x', y, w) = P(w\,|\,x') P(x') P(y\,|\,x')$$

thus $P(x'\,|\,y, w) = 0$ implies, by the positivity of $P(x')$ and $P(y\,|\,x')$, that $P(w\,|\,x') = 0$. Thus we have shown that for a pair $(x, x') \in G_X$ if $w \in N_W(x)$, then $w \notin N_W(x')$, or equivalently, $N_W(x) \cap N_W(x') = \emptyset$.

Appendix III
Proof of Theorem 4

We have established already that $R_\ell(D)$ is given by (51), where $R_{k,\ell}(D)$ is given in (46). Denoting (until the end of the proof) the right-hand side of (58) by $\tilde{R}_\ell(D)$, we need to show that $R_\ell(D) = \tilde{R}_\ell(D)$. Note first that an equivalent form of $R_{k,\ell}(D)$ is

$$R_{k,\ell}(D) = \frac{1}{k} \min I(X^k; W^k) \tag{A3}$$

where the minimum is over all functions $f_i : \mathcal{W}^k \times \mathcal{Y}^{i+\ell} \to \hat{\mathcal{X}}$, $1 \le i \le k - \ell, |\mathcal{W}| \le |\mathcal{X}| + 1$, and $P(w^k|x^k)$ such that

$$\frac{1}{k-\ell} \sum_{i=1}^{k-\ell} E[\rho(X_i, f_i(W^k, Y^{i+\ell}))] \le D. \tag{A4}$$

The equivalence follows by identifying W^k in (A3) as W of (46). We first prove that $R_\ell(D) \ge \tilde{R}_\ell(D)$. Assume that W^k, X^k, Y^k jointly achieve the minimum in (A3). We form the stationary triple $\tilde{\boldsymbol{W}}, \tilde{\boldsymbol{X}}, \tilde{\boldsymbol{Y}}$ by concatenating i.i.d. k-triplets distributed as W^k, X^k, Y^k, followed by a random time-shift, S, uniformly distributed on $\{1, \ldots, k\}$. Letting $\tilde{W}_i = (W_i, S)$, we note that $\tilde{\boldsymbol{W}}, \tilde{\boldsymbol{X}}, \tilde{\boldsymbol{Y}}$ are jointly stationary and satisfy the Markov relation $\tilde{\boldsymbol{W}} \to \tilde{\boldsymbol{X}} \to \tilde{\boldsymbol{Y}}$ (since when conditioning only on $\tilde{\boldsymbol{X}}$, or on both $\tilde{\boldsymbol{X}}$ and $\tilde{\boldsymbol{W}}$, the components of $\tilde{\boldsymbol{Y}}$ are independent, the distribution of Y_i depends only on X_i). Now, for any n,

$$\begin{aligned} I(\tilde{X}^n; \tilde{W}^n) &\le I(\tilde{X}^n; \tilde{W}^n\,|\,S) + \log k \\ &\le I(\tilde{X}^{\lceil n/k \rceil \cdot k}; \tilde{W}^{\lceil n/k \rceil \cdot k} | S) + \log k \\ &\le (\lceil n/k \rceil + 1) I(X^k; W^k) + \log k \end{aligned}$$

implying

$$\bar{I}(\tilde{\boldsymbol{X}}; \tilde{\boldsymbol{W}}) = \lim_n \frac{1}{n} I(\tilde{X}^n; \tilde{W}^n) \le \frac{1}{k} I(X^k; W^k) = R_{k,\ell}(D). \tag{A5}$$

On the other hand, letting $\tilde{X}_0^{\rm opt}(\tilde{\boldsymbol{W}},\tilde{Y}_{-\infty}^{\ell})$ denote the optimal estimate of $\tilde{X}_0$ based on $\tilde{\boldsymbol{W}},\tilde{Y}_{-\infty}^{\ell}$, and taking $\{f_i\}_{1\le i\le k-\ell}$ to be the achievers in (A4) while defining $\{f_i\}_{k-\ell+1\le i\le k}$ arbitrarily

$$\begin{aligned}
&E\left[\rho\left(\tilde{X}_0,\tilde{X}_0^{\rm opt}\left(\tilde{\boldsymbol{W}},\tilde{Y}_{-\infty}^{\ell}\right)\right)\right]\\
&\quad\overset{\rm (a)}{\le} E[\rho(X_S,f_S(W^k,Y^{S+\ell}))]\\
&\quad=\frac{1}{k}\sum_{i=1}^{k}E[\rho(X_i,f_i(W^k,Y^{i+\ell}))]\\
&\quad\le\frac{1}{k}\left[\ell\cdot\rho_{\max}+\sum_{i=1}^{k-\ell}E[\rho(X_i,f_i(W^k,Y^{i+\ell}))]\right]\\
&\quad\overset{\rm (b)}{\le}\frac{\ell\cdot\rho_{\max}}{k}+D,
\end{aligned}\tag{A6}$$

where (a) follows by considering a suboptimal estimator, in lieu of $\tilde{X}_0^{\rm opt}(\tilde{\boldsymbol{W}},\tilde{Y}_{-\infty}^{\ell})$, which employs the f_i corresponding to the value of the shift S (which belongs to $\tilde{\boldsymbol{W}}$) on the corresponding components of $\tilde{\boldsymbol{W}},\tilde{\boldsymbol{Y}}$, and (b) follows from (A4). The combination of (A5) and (A6) implies that

$$\tilde{R}_\ell\left(D+\frac{\ell\rho_{\max}}{k}\right)\le R_{k,\ell}(D).\tag{A7}$$

By the arbitrariness of k and the continuity of $\tilde{R}_\ell$ (which follows from its easily verified convexity), this gives

$$\tilde{R}_\ell(D)\le\lim_{k\to\infty}R_{k,\ell}(D)=R_\ell(D).\tag{A8}$$

For the reverse direction, fix $\varepsilon>0$ and assume that $\boldsymbol{W},\boldsymbol{X},\boldsymbol{Y}\varepsilon$ achieve the infimum in (58). Let m be sufficiently large such that

$$E\left[\rho\left(X_0,\hat{X}_0^{\rm opt}\left(W_{-m}^{m},Y_{-m}^{\ell}\right)\right)\right]\le D+\varepsilon,\tag{A9}$$

where $\hat{X}_0^{\rm opt}(W_{-m}^{m},Y_{-m}^{\ell}))$ denotes the optimal estimate of X_0 based on W_{-m}^{m},Y_{-m}^{ℓ} (that such an m exists follows from the fact that

$$E[\rho(X_0,\hat{X}_0^{\rm opt}(\boldsymbol{W},Y_{-\infty}^{\ell}))]=\lim_{m}E[\rho(X_0,\hat{X}_0^{\rm opt}(W_{-m}^{m},Y_{-m}^{\ell}))]$$

which is a consequence of martingale convergence [3, Theorem 5.21], cf. also [37, Lemma 4]). Now let k be sufficiently large so that

$$\frac{1}{k}I(X^k;W^k)\le\bar{I}(\boldsymbol{X};\boldsymbol{W})+\varepsilon\tag{A10}$$

and

$$\frac{(2m-\ell)\rho_{\max}}{k-\ell}\le\varepsilon.\tag{A11}$$

Letting $g(w_{-m}^{m},y_{-m}^{\ell})=\hat{X}_0^{\rm opt}(w_{-m}^{m},y_{-m}^{\ell})$, set

$$f_i(w^k,y^{i+\ell})=g\left(w_{i-m}^{i+m},y_{i-m}^{i+\ell}\right),\qquad\text{for } m+1\le i\le k-\ell\tag{A12}$$

and f_i to be arbitrary, otherwise. Then

$$\begin{aligned}
&\frac{1}{k-\ell}\sum_{i=1}^{k-\ell}E[\rho(X_i,f_i(W^k,Y^{i+\ell}))]\\
&\quad\le\frac{1}{k-\ell}\left[(2m-\ell)\rho_{\max}+\sum_{i=m+1}^{k-m}E[\rho(X_i,f_i(W^k,Y^{i+\ell}))]\right]\\
&\quad=\frac{(2m-\ell)\rho_{\max}}{k-\ell}+\frac{1}{k-\ell}\sum_{i=m+1}^{k-m}E\left[\rho\left(X_i,g\left(W_{i-m}^{i+m},Y_{i-m}^{i+\ell}\right)\right)\right]\\
&\quad\overset{\rm (a)}{=}\frac{(2m-\ell)\rho_{\max}}{k-\ell}+\frac{k-2m}{k-\ell}E\left[\rho\left(X_0,g\left(W_{-m}^{m},Y_{-m}^{\ell}\right)\right)\right]\\
&\quad=\frac{(2m-\ell)\rho_{\max}}{k-\ell}\\
&\qquad+\frac{k-2m}{k-\ell}E\left[\rho\left(X_0,\hat{X}_0^{\rm opt}\left(W_{-m}^{m},Y_{-m}^{\ell}\right)\right)\right]\\
&\quad\le\varepsilon+(D+\varepsilon)
\end{aligned}\tag{A13}$$

where (a) is due to the joint stationarity of $\boldsymbol{W},\boldsymbol{X},\boldsymbol{Y}$ and the last inequality follows from (A9) and (A11). The combination of (A10) and (A13) implies that $R_{k,\ell}(D+2\varepsilon)\le\bar{I}(\boldsymbol{X};\boldsymbol{W})+\varepsilon$, which, when combined with the fact that $\bar{\boldsymbol{W}},\bar{\boldsymbol{X}}$ ε achieve the infimum in (58), gives

$$R_{k,\ell}(D+2\varepsilon)\le\tilde{R}_\ell(D)+2\varepsilon.\tag{A14}$$

Finally, since (A14) holds for *all* sufficiently large k

$$R_\ell(D+2\varepsilon)\le\tilde{R}_\ell(D)+2\varepsilon\tag{A15}$$

which, by the arbitrariness of $\varepsilon>0$ and the continuity of $R_\ell(\cdot)$ (which follows from its convexity), implies $R_\ell(D)\le\tilde{R}_\ell(D)$.

Appendix IV
Proof of Theorem 5

Trivially, $\underline{R}_\ell^G(D)\le\underline{R}_0^G(D)\le\underline{R}_{\rm ub}(D)$, so it only remains to prove that $\underline{R}_\ell^G(D)\ge\underline{R}_{\rm ub}(D)$ for every ℓ. Toward this end define

$$R_\infty^G(D)=\inf\left\{\bar{I}(\boldsymbol{X};\boldsymbol{W}):E\left(X_0-\hat{X}_0^{\rm opt}(\boldsymbol{W},\boldsymbol{Y})\right)^2\le D\right\}\tag{A16}$$

where the infimum is over jointly stationary and Gaussian $\boldsymbol{W},\boldsymbol{X},\boldsymbol{Y}$ with the Markov relation $\boldsymbol{W}\to\boldsymbol{X}\to\boldsymbol{Y}$. Clearly, $R_\ell^G(D)\ge R_\infty^G(D)$ for every ℓ, so it suffices to show that $R_\infty^G(D)\ge\underline{R}_{\rm ub}(D)$. Note that the most general scenario of zero mean $\boldsymbol{W},\boldsymbol{X},\boldsymbol{Y}$ that are jointly Gaussian, jointly stationary, and satisfy the Markov relation $\boldsymbol{W}\to\boldsymbol{X}\to\boldsymbol{Y}$ is when $\boldsymbol{W}$ is the output of a linear filter F whose input is $\boldsymbol{X}$, "corrupted" by additive Gaussian (not necessarily white) noise $\boldsymbol{Z}$ that is independent of $\boldsymbol{X},\boldsymbol{Y}$. However, neither the mutual information rate nor the MSE are affected when considering $\tilde{\boldsymbol{W}}$ instead of $\boldsymbol{W}$, where $\tilde{\boldsymbol{W}}$ is the output of the whitening filter for the power spectral density (psd) of $\boldsymbol{Z}$ (since the transformation from $\boldsymbol{W}$ to $\tilde{\boldsymbol{W}}$ is invertible). This implies that the infimum in (A16) is not affected by restricting ourselves to $\boldsymbol{W}$ that is the

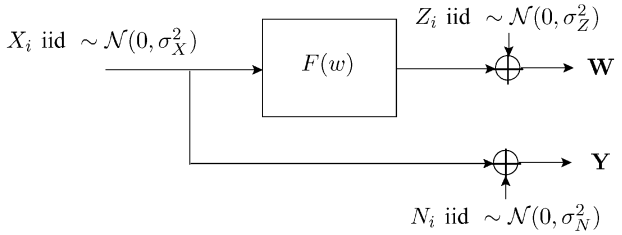

Fig. 9. The infimum in the definition of $R_\infty^G(D)$ is not affected by restricting to $\boldsymbol{W},\boldsymbol{X},\boldsymbol{Y}$ that are generated as in the figure.

output of a linear filter F "corrupted" by white Gaussian noise $\boldsymbol{Z}$ independent of $\boldsymbol{X},\boldsymbol{Y}$. This scenario is depicted in Fig. 9. As is well know (cf., e.g., [16]), the minimum MSE estimate of $\boldsymbol{X}$ given $\boldsymbol{W},\boldsymbol{Y}$ is, in general, obtained by passing $\boldsymbol{W}$ through the filter $A,\boldsymbol{Y}$ through the filter B, and adding the outputs (as depicted in Fig. 10), where A and B are given in the spectral domain by

$$\begin{aligned} A &= \frac{S_{\boldsymbol{XW}}S_{\boldsymbol{Y}} - S_{\boldsymbol{XY}}S_{\boldsymbol{YW}}}{S_{\boldsymbol{W}}S_{\boldsymbol{Y}} - |S_{\boldsymbol{WY}}|^2} \\ B &= \frac{S_{\boldsymbol{XY}}S_{\boldsymbol{W}} - S_{\boldsymbol{XW}}S_{\boldsymbol{WY}}}{S_{\boldsymbol{W}}S_{\boldsymbol{Y}} - |S_{\boldsymbol{WY}}|^2}. \end{aligned} \tag{A17}$$

Substituting the simple forms of the psd and cross-psd in our case gives

$$\begin{aligned} A &= \frac{\sigma_X^2\sigma_N^2\overline{F}}{\sigma_X^2\sigma_N^2|F|^2 + \sigma_Z^2\sigma_Y^2} \\ B &= \frac{\sigma_X^2\sigma_Z^2}{\sigma_X^2\sigma_N^2|F|^2 + \sigma_Z^2\sigma_Y^2}. \end{aligned} \tag{A18}$$

Letting $\boldsymbol{R},\boldsymbol{S}$ denote the respective outputs of the filters A,B, the associated minimum mean-square error (MMSE) is readily computed to be

$$\begin{aligned} &E[(X_0 - \hat{X}_0)^2] \\ &\quad = E[(X_0 - R_0 - S_0)^2] \\ &\quad = \sigma_X^2 - \frac{1}{\pi}\int_0^{2\pi} A\sigma_X^2 F dw \\ &\qquad - \frac{1}{\pi}\int_0^{2\pi} B\sigma_X^2 dw + \frac{1}{2\pi}\int_0^{2\pi}\left[\sigma_X^2|F|^2 + \sigma_Z^2\right]|A|^2 dw \\ &\qquad + \frac{1}{2\pi}\int_0^{2\pi}\left(\sigma_X^2 + \sigma_N^2\right)|B|^2 dw \\ &\qquad + \frac{1}{\pi}\int_0^{2\pi} A\overline{B}\sigma_X^2 F dw \qquad \text{(A19)} \\ &\quad = \int_0^{2\pi} G(|F|^2) dw. \qquad \text{(A20)} \end{aligned}$$

The particular form of G is not important here. The key point is that the MMSE can be expressed as an integral of a function that depends on w only through the $|F(w)|^2$.

The mutual information rate is given by

$$\begin{aligned} \bar{I}(\boldsymbol{X};\boldsymbol{W}) &= \frac{1}{2}\log\frac{\exp\left\{\frac{1}{2\pi}\int_0^{2\pi}\ln\left[\sigma_X^2|F(w)|^2 + \sigma_Z^2\right]dw\right\}}{\sigma_Z^2} \\ &= \frac{1}{4\pi}\int_0^{2\pi}\log\left[1 + \frac{\sigma_X^2|F(w)|^2}{\sigma_Z^2}\right]dw \\ &= \int_0^{2\pi} J(|F|^2) dw. \end{aligned} \tag{A21}$$

Here too, the mutual information can be expressed as an integral of a function that depends on w only through the value $|F(w)|^2$. Thus, we have shown that every point on the curve $R_\infty^G(\cdot)$ is of the form given by (A20) and (A21) or, equivalently, of the form

$$D = \int_0^\infty G(u)d\mu(u), \quad R = \int_0^\infty J(u)d\mu(u) \tag{A22}$$

for some probability measure μ. Specifically, to get from (A20) and (A21) to (A22), let μ be the distribution of $|F(U)|^2$, where $U \sim \mathrm{Unif}[0,2\pi]$. To conclude it suffices to show that, for every $\alpha \geq 0$, $(G(\alpha), J(\alpha))$ is a point on the curve $R_{\mathrm{ub}}(\cdot)$ of Section II (in other words, $R_{\mathrm{ub}}(G(\alpha)) = J(\alpha)$). To see this, note that the MMSE and mutual information corresponding to a point on the curve $R_{\mathrm{ub}}(\cdot)$ are obtained in the setting of Fig. 10 by a filter $F(w)$, which is constant $F(w) \equiv \sqrt{\alpha}$, but the value of this point, by (A20) and (A21), is given by $(G(\alpha), J(\alpha))$. To sum up, we have shown (by (A22)) that every point on the curve $R_\infty^G(\cdot)$ is a convex combination of points on the curve $R_{\mathrm{ub}}(\cdot)$, implying that $R_\infty^G(D) \geq \underline{R}_{\mathrm{ub}}(D)$.

Appendix V
Proof That $\tilde{R}^{k,\ell}_{\mathrm{LSCSI}} \leq R^{k,\ell}_{\mathrm{LSCSI}} + \frac{\ell}{k}R_{\mathrm{LSCSI}}$

Let $P^{k,\ell}_{\mathrm{LSCSI}}(w|x^k)$ and $P_{\mathrm{LSCSI}}(w\,|\,x)$ be the conditional probability distributions that achieve $R^{k,\ell}_{\mathrm{LSCSI}}$, i.e., the minimum in ((63)), and R_{LSCSI}, i.e., the minimum (in (35)), respectively. Now, under the conditional distribution

$$P(\hat{w},\tilde{w}^\ell|x^k) = P^{k,\ell}_{\mathrm{LSCSI}}(\hat{w}\,|\,x^k) \times \prod_{j=1}^{\ell} P_{\mathrm{LSCSI}}(\tilde{w}_j\,|\,x_{k-\ell+j}),$$

distribution

$$\begin{aligned} &I(X^k;\hat{W},\tilde{W}^\ell) \\ &\quad = H(\hat{W},\tilde{W}^\ell) - H(\hat{W},\tilde{W}^\ell\,|\,X^k) \\ &\quad \leq H(\hat{W}) + H(\tilde{W}^\ell) - H(\hat{W},\tilde{W}^\ell\,|\,X^k) \\ &\quad = H(\hat{W}) + H(\tilde{W}^\ell) - H(\hat{W}\,|\,X^k) - H\left(\tilde{W}^\ell\,|\,X^k_{k-\ell+1}\right) \\ &\quad = I(X^k;\hat{W}) + I\left(X^k_{k-\ell+1};\tilde{W}^\ell\right) \\ &\quad = kR^{k,\ell}_{\mathrm{LSCSI}} + \ell R_{\mathrm{LSCSI}}. \end{aligned} \tag{A23}$$

On the other hand, since under

$$P^{k,\ell}_{\mathrm{LSCSI}}(w\,|\,x^k)H(X_i\,|\,W,Y^{i+\ell}) = 0, \quad \text{for } 1 \leq i \leq k-\ell$$

and under $P_{\mathrm{LSCSI}}(w\,|\,x)$ $H(X|W,Y) = 0$

$$\begin{aligned} H(X_i\,|\,\hat{W},\tilde{W}^\ell,Y^{i+\ell}) \leq H(X_i\,|\,\hat{W},Y^{i+\ell}) = 0, \\ \text{for all } 1 \leq i \leq k-\ell \end{aligned} \tag{A24}$$

and

$$\begin{aligned} H\left(X^k_{k-\ell+1}|\hat{W},\tilde{W}^\ell,Y^k_{k-\ell+1}\right) &\leq H\left(X^k_{k-\ell+1}\,|\,\tilde{W}^\ell,Y^k_{k-\ell+1}\right) \\ &\leq \sum_{j=1}^{\ell} H(X_{k-\ell+j}|\tilde{W}_j,Y_{k-\ell+j}) = 0. \end{aligned} \tag{A25}$$

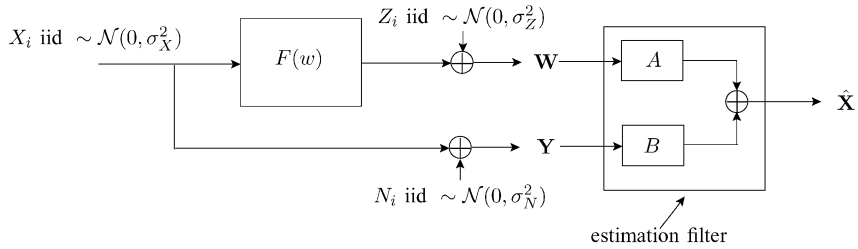

Fig. 10. Form of $\boldsymbol{W}, \boldsymbol{X}, \boldsymbol{Y}, \hat{\boldsymbol{X}}$ associated with $R_\infty^G(D)$.

Thus, for $\breve{W} = (\hat{W}, \tilde{W}^\ell)$, $P(\breve{w} \,|\, x^k)$ is in the feasible set associated with the minimization in (67), which by (A23) implies that $\tilde{R}^{k,\ell}_{\mathrm{LSCSI}} \le R^{k,\ell}_{\mathrm{LSCSI}} + \frac{\ell}{k} R_{\mathrm{LSCSI}}$.

Appendix VI
Proof of Lemma 3

We need to show that for any function f, if $H(f(X) \,|\, W, Y) > 0$ then $H(f(X) \,|\, W, Y') > 0$. Assume that $H(f(X) \,|\, W, Y) > 0$. This implies the existence of an $a \neq b, w$, or y such that

$$P(f(X) = a, W = w, Y = y) > 0$$

and

$$P(f(X) = b, W = w, Y = y) > 0. \tag{A26}$$

Since

$$\begin{aligned} &P(f(X) = a, W = w, Y = y) \\ &\quad = \sum_{x: f(x) = a} P(w \,|\, x) P(X = x, Y = y) \end{aligned}$$

and

$$\begin{aligned} &P(f(X) = a, W = w, Y' = y) \\ &\quad = \sum_{x: f(x) = a} P(w \,|\, x) P(X = x, Y' = y) \end{aligned}$$

the fact that X, Y and X, Y' induce the same bipartite graph implies that each summand in the first sum is positive if and only if its corresponding summand in the second sum is positive. When combined with (A26), this implies that $P(f(X) = a, W = w, Y' = y) > 0$ and, similarly, $P(f(X) = b, W = w, Y' = y) > 0$, which in turn implies that $H(f(X)|W, Y') > 0$.

Acknowledgment

The authors gratefully acknowledge Ertem Tuncel for insightful comments and suggestions, and for working out the explicit form of R_{LSCSI} for the sixth configuration in Fig. 6. They are also grateful to the referees for comments that have helped to improve the manuscript.

References

[1] N. Alon and A. Orlitsky, "Source coding and graph entropies," *IEEE Trans. Inf. Theory*, vol. 42, no. 5, pp. 1329–1339, Sep. 1996.

[2] T. Berger, *Rate-Distortion Theory: A Mathematical Basis for Data Compression*. Englewood Cliffs, N.J.: Prentice-Hall, 1971.

[3] L. Breiman, *Probability*. Philadelphia, PA: SIAM, 1992.

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

[5] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*. New York: Academic, 1981.

[6] A. Das and P. Narayan, "Capacities of time-varying multiple-access channels with side information," *IEEE Trans. Inf. Theory*, vol. 48, no. 1, pp. 4–25, Jan. 2002.

[7] A. Dembo and T. Weissman, "The minimax distortion redundancy in noisy source coding," *IEEE Trans. Inf. Theory*, vol. 49, no. 11, pp. 3020–3030, Nov. 2003.

[8] Y. Ephraim and N. Merhav, "Hidden Markov processes," *IEEE Trans. Inf. Theory*, vol. 48, no. 6, pp. 1518–1569, Jun. 2002.

[9] U. Erez, S. Shamai (Shitz), and R. Zamir, "Capacity and lattice strategies for canceling known interference," *IEEE Trans. Inf. Theory*, vol. 51, no. 11, pp. 3820–3833, Nov. 2005.

[10] T. Ericson, "A result on delay-less information transmission," in *Proc. Int. Symp. Information Theory*, Grignano, Italy, Jun. 1979.

[11] N. T. Gaarder and D. Slepian, "On optimal finite-state digital transmission systems," *IEEE Trans. Inf. Theory*, vol. IT-28, no. 2, pp. 167–186, Mar. 1982.

[12] S. I. Gel'fand and M. S. Pinsker, "Coding for channel with random parameters," *Probl. Contr. and Inf. Theory*, vol. 9, no. I, pp. 19–31, 1980.

[13] R. M. Gray, D. L. Neuhoff, and P. C. Shields, "A generalization of Ornstein's d-bar distance with applications to information theory," *Ann. Probab.*, vol. 3, no. 2, pp. 315–328, Apr. 1975.

[14] R. M. Gray, D. L. Neuhoff, and J. Omura, "Process definitions of distortion-rate functions and source coding theorems," *IEEE Trans. Inf. Theory*, vol. IT-21, no. 5, pp. 524–532, Sep. 1975.

[15] C. Heegard and A. El Gamal, "On the capacity of computer memory with defects," *IEEE Trans. Inf. Theory*, vol. IT-29, no. 5, pp. 731–739, Sep. 1983.

[16] T. Kailath, A. H. Sayed, and B. Hassibi, *Linear Estimation*. Upper Saddle River, NJ: Prentice-Hall, 2000.

[17] J. Körner, "Coding of an information source having ambiguous alphabet and the entropy of graphs," in *Proc. 6th Pargue Conf. Information Theory, 1973*, Prague, Czech Republic, 2003, pp. 411–425.

[18] J. Körner, "Fredman-Komlós bounds and information theory," *SIAM J. Alg. Discr. Meth.*, vol. 7, pp. 560–570, 1986.

[19] P. Koulgi, E. Tuncel, S. L. Regunathan, and K. Rose, "On zero-error source coding with decoder side information," *IEEE Trans. Inf. Theory*, vol. 49, no. 1, pp. 99–111, Jan. 2003.

[20] T. Linder and G. Lugosi, "A zero-delay sequential scheme for lossy coding of individual sequences," *IEEE Trans. Inf. Theory*, vol. 47, no. 6, pp. 2533–2538, Sep. 2001.

[21] T. Linder and R. Zamir, "Causal source coding of stationary sources and individual sequences with high resolution," *IEEE Trans. Inf. Theory*, accepted for publication.

[22] S. Matloub and T. Weissman, "On competitive zero—delay joint source—channel coding," in *Proc. Conf. Information Science and Systems*, Princeton, NJ, 2004, pp. 555–559.

[23] N. Merhav, "A comment on 'A rate of convergence result for a universal D-semifaithful code'," *IEEE Trans. Inf. Theory*, vol. 41, no. 4, pp. 1200–1202, Jul. 1995.

[24] N. Merhav and I. Kontoyiannis, "Source coding exponents for zero-delay coding with finite memory," *IEEE Trans. Inf. Theory*, vol. 49, no. 3, pp. 609–625, Mar. 2003.

[25] N. Merhav and T. Weissman, "Coding for the feedback Gel'fand-Pinsker channel and the feedforward Wyner-Ziv source," in *Proc. IEEE Int. Symp. Information Theory*, Adelaide, Australia, Sep. 2005, pp. 1506–1510.
[26] D. L. Neuhoff and R. K. Gilbert, "Causal source codes," *IEEE Trans. Inf. Theory*, vol. IT-28, no. 5, pp. 701–713, Sep. 1982.
[27] R. J. Pilc, "The transmission distortion of a source as a function of the encoding block length," *Bell Syst. Tech. J.*, vol. 47, pp. 827–885, 1968.
[28] J. P. M. Schalkwijk, "A coding scheme for additive noise channels with feedback—II: Band-limited signals," *IEEE Trans. Inf. Theory*, vol. IT-12, no. 2, pp. 183–189, Apr. 1966.
[29] J. P. M. Schalkwijk and T. Kailath, "A coding scheme for additive noise channels with feedback—I: No bandwidth constraint," *IEEE Trans. Inf. Theory*, vol. IT-12, no. 2, pp. 183–189, Apr. 1966.
[30] C. E. Shannon, "The zero-error capacity of a noisy channel," *IRE Trans. Inf. Theory*, vol. IT-2, no. 3, pp. 8–19, Sep. 1956.
[31] ——, "Channels with side information at the transmitter," *IBM J. Res. Dev.*, vol. 2, pp. 289–293, 1958.
[32] D. Slepian and J. K. Wolf, "Noiseless coding of correlated information sources," *IEEE Trans. Inf. Theory*, vol. IT-19, no. 4, pp. 471–480, Jul. 1973.
[33] D. Teneketzis, "Optimal real-time encoding-decoding of markov sources in noisy environments," in *Proc. Math. Theory of Networks and Syst. (MTNS)*, Leuwen, Belgium, 2004.
[34] J. C. Walrand and P. Varaiya, "Optimal causal coding-decoding problems," *IEEE Trans. Inf. Theory*, vol. IT-29, no. 6, pp. 814–820, Nov. 1983.
[35] T. Weissman and N. Merhav, "On limited-delay lossy coding and filtering of individual sequences," *IEEE Trans. Inf. Theory*, vol. 48, no. 3, pp. 721–733, Mar. 2002.
[36] ——, "On causal source codes with side information," *IEEE Trans. Inf. Theory*, vol. 51, no. 11, pp. 4003–4013, Nov. 2005.
[37] T. Weissman, E. Ordentlich, G. Seroussi, S. Verdú, and M. Weinberger, "Universal discrete denoising: Known channel," *IEEE Trans. Inf. Theory*, vol. 51, no. 1, pp. 5–28, Jan. 2005.
[38] H. S. Witsenhausen, "On the structure of real-time source coders," *Bell Syst. Tech. J.*, vol. 58, no. 6, pp. 1437–1451, Jul./Aug. 1979.
[39] ——, "On sequences of pairs of dependent random variables," *SIAM J. Appl. Math.*, vol. 28, pp. 100–113, Jan. 1975.
[40] ——, "The zero-error side information problem and chromatic numbers," *IEEE Trans. Inf. Theory*, vol. IT-22, no. 5, pp. 592–593, Sep. 1976.
[41] A. D. Wyner, "The rate distortion function for source coding with side information at the decoder—II: General sources," *Inf. Contr.*, vol. 38, pp. 60–80, 1978.
[42] A. D. Wyner and J. Ziv, "The rate distortion function for source coding with side information at the decoder," *IEEE Trans. Inf. Theory*, vol. IT-22, no. 1, pp. 1–10, Jan. 1976.
[43] Z. Zhang, E. H. Yang, and V. K. Wei, "The redundancy of source coding with a fidelity criterion—Part I: Known statistics," *IEEE Trans. Inf. Theory*, vol. 43, no. 1, pp. 71–91, Jan. 1997.