On Causal Source Codes with Side Information*

Tsachy Weissman Neri Merhav

August 23, 2004

Abstract

We study the effect of the introduction of side information into the causal source coding setting of Neuhoff and Gilbert. We find that the spirit of their result, namely the sufficiency of time-sharing scalar quantizers (followed by appropriate lossless coding) for attaining optimum performance within the family of causal source codes, extends to many scenarios involving availability of side information (at both encoder and decoder, or only on one side). For example, in the case where side information is available at both encoder and decoder, we find that time-sharing side-information-dependent scalar quantizers (at most two for each side-information symbol) attains optimum performance. This remains true even when the reproduction sequence is allowed non-causal dependence on the side information and even for the case where the source and the side information, rather than consisting of i.i.d. pairs, form, respectively, the output of a memoryless channel and its stationary ergodic input.

Key words and phrases: Causal source codes, Entropy coding, Lossy source coding, Scalar quantization, Side information.

1 Introduction

A (lossless or lossy) source code consists of two components: The encoder, which generates a bit stream upon observation of the source sequence, and the decoder, which reproduces (a possibly approximate version of) the source sequence based on its observation of that bit stream. A source code is *causal* if the k-th reproduction symbol depends on the source sequence only through its first k components.

The most striking fact about causal source codes was established by Neuhoff and Gilbert in [12]. Their main result is that for source coding of a memoryless source "if the future is not allowed to be looked into, the past is useless". More precisely put, the conclusion in [12] was that time sharing at most two scalar quantizers, followed by lossless entropy coding, attains optimum performance for memoryless sources among all causal source codes. At a first glance it may seem natural that, for a memoryless source, there is nothing to gain from the past sequence for reconstruction of the present symbol. The strikingness of the result, however, accentuates when contrasted with Shannon theory which renders other sequence components quite relevant for the coding of each symbol, even for i.i.d. sources.

The theory of causal source coding has been expanded since [12]. Causal source codes for sources with memory were considered to a limited extent in [2]. Recently, Linder and Zamir have extended the Neuhoff and Gilbert result

*Presented in part at the 40th Annual Allerton Conference on Communication, Control, and Computing, Monticello, IL, 2002. Part of this work was performed while the authors were visiting Hewlett-Packard Laboratories. T. Weissman is with the Information Systems Laboratory, Department of Electrical Engineering, Stanford University, Stanford, CA 94305. E-mail: tsachy@stanford.edu. N. Merhav is with the Department of Electrical Engineering, Technion- Israel Institute of Technology, Haifa 32000, Israel. E-mail: merhav@ee.technion.ac.il.

to the case of a general stationary source, as well as the individual sequence setting, in the high resolution (low distortion) limit [9]. The closely related setting of zero- and limited-delay source coding has also received attention lately: From error exponents [11, 13], to zero- and limited-delay coding [8, 15, 6, 7] and joint-source-channel-coding [10] in the individual-sequence setting. The reader is referred to [11] for a more comprehensive account of related literature.

In this work we study some of the effect of the introduction of side information into the Neuhoff-Gilbert setting. In particular, we seek to characterize optimum performance achievable by causal source codes in situations involving side information. Our finding is that the spirit of the result of [12], namely the sufficiency of time-sharing scalar quantizers for attaining optimum performance, extends, in senses that will be made precise, to many of the scenarios involving side information. We find, for example, for the case where side information is available at both encoder and decoder that time sharing side-information-dependent scalar quantizers attains optimum performance even when the reproduced sequence is allowed non-causal dependence on the side information. Furthermore, this remains true even when the source and the side information, rather than consisting of i.i.d. pairs, form, respectively, the output of a memoryless channel and its stationary ergodic input.

For the case where side information is available at the encoder only we find that when the reproduction sequence is restricted to causal dependence on both the source and the side information sequence, the side information is useless. This, evidently, is like the case of non-causal source coding, where side information at the encoder alone is useless. As is pointed out, however, this is no longer true when the reproduction sequence is allowed non-causal dependence on the side information sequence.

As we argue in Section 5, causal source coding with side information at the decoder alone is most motivated, from an operational viewpoint, when the reproduction is not allowed to depend on the side information. Using Slepian-Wolf coding it is seen that, in this case, there is no penalty for the absence of side information at the encoder.

Though the techniques underlying our proofs are similar in spirit to those of [12], various twists are necessary to accommodate the different settings considered and we have not found one "meta-theorem" from which the results can be derived as corollaries.

The remainder of this work is organized as follows. Section 2 presents the concrete problem formulation along with its "operational" motivation. Section 3 is dedicated to the setting of causal source coding when side information is available at both encoder and decoder. In Section 4 we look at causal source coding when side information is available at the encoder only and show that the side information is useless when the reproduction is constrained to causal dependence on the side information as well as on the source sequence. Section 5 addresses the case where side information is available solely to the *decoder*. We conclude in Section 6 with a summary of our findings and a mention of a few directions for related future work.

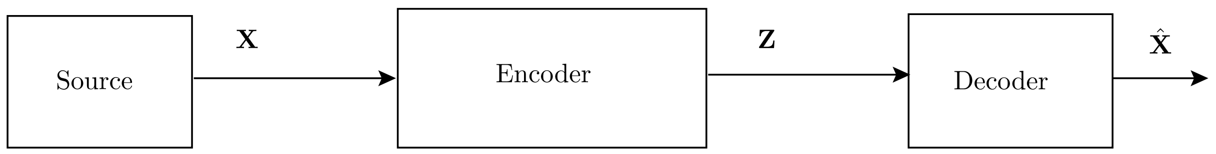

Figure 1: Source coding system.

2 Notation, Preliminaries, and Problem Formulation

2-A Notation and Conventions

Random variables will be denoted by upper case letters while specific values they may take will be denoted by the corresponding lower case letters. Double-sided infinite sequences will be denoted by boldface letters. Thus, the random variable V may assume the value v and the stochastic process $\mathbf{V} = \cdots, V_{-1}, V_0, V_1, \cdots$ may have a specific realization $\mathbf{v} = \cdots, v_{-1}, v_0, v_1, \cdots$. For $j \geq i$ let V_i^j denote the vector $(V_i, \ldots, V_j)$ and V^j denote the one-sided sequence $(\ldots, V_{j-1}, V_j)$. Throughout $\mathcal{X}, \mathcal{Y}, \hat{\mathcal{X}}$ will denote, respectively, the source-, side-information-, and reconstruction-alphabet. $\mathcal{Y}$ will be assumed finite[1] while $\mathcal{X}$ and $\hat{\mathcal{X}}$ can be assumed quite arbitrary measurable spaces (equipped with their σ-fields) . Mappings will be assumed measurable (with respect to the relevant σ-fields that will be clear from the context). For any set $\mathcal{A}$ (equipped with a σ-field) we let $\mathcal{M}(\mathcal{A})$ denote the set of all probability measures on $\mathcal{A}$ (and its σ-field). The inf of an empty set will be taken as infinity. H will denote the entropy (*not* differential entropy) functional. Thus, for example, $H(\hat{X}_1^k)$ will denote the entropy of $\hat{X}_1^k$ (being infinite if $\hat{X}_1^k$ is not discrete valued) and $H(\hat{X}_1^k|\mathbf{Y})$ the conditional entropy of $\hat{X}_1^k$ given $\mathbf{Y}$. A generic notation for a probability measure will be μ, with the argument revealing the random variable(s) to which it relates. Finally, $|\cdot|$ will denote cardinality.

2-B The Original Neuhoff-Gilbert Setting

Consider a discrete-time double-sided stochastic process $\mathbf{X} = \{X_k\}_{k=-\infty}^{\infty}$ with symbols in the alphabet $\mathcal{X}$. A source code (cf. Figure 1) operates as follows: The encoder accesses the source sequence $\mathbf{X}$ and creates a binary representation $\mathbf{Z} = \{Z_k\}_{k\geq 1}$. The decoder accepts $\mathbf{Z}$ and creates a reproduction $\hat{\mathbf{X}} = \{\hat{X}_k\}_{k\geq 1}$ of $\{X_k\}_{k\geq 1}$ with symbols in the reproduction space $\hat{\mathcal{X}}$. The system induced by the cascade of the encoder and the decoder is referred to as the *reproduction coder*. More precisely, the reproduction coder is characterized by a family of measurable mappings $\{f_k\}_{k=1}^{\infty}$ such that the reproduction $\hat{X}_k$ of the kth source output X_k is given by $\hat{X}_k = f_k(\mathbf{X})$.

When a source code with an induced reproduction code $\{f_k\}$ is applied to the source $\mathbf{X}$, the *average distortion* is defined by

$$d(\{f_k\}) = \limsup_{k\to\infty} E\left[d_k(X_1^k, \hat{X}_1^k)\right], \tag{1}$$

where $d_k(x_1^k, \hat{x}_1^k) = \frac{1}{k}\sum_{i=1}^k d(x_i, \hat{x}_i)$ is the block distortion induced by the single-letter distortion measure $d : \mathcal{X}\times\hat{\mathcal{X}} \to [0,\infty)$. The *average rate* r is defined by

$$r = \limsup_{k\to\infty} \frac{1}{k} E\left[L_k(\mathbf{X})\right], \tag{2}$$

[1] While the proofs of the converse results in the sequel can readily be verified to carry over to the case of a quite general $\mathcal{Y}$-alphabet, the direct proofs for a general alphabet become more tedious and, in fact, for the results to hold certain conditions must be imposed.

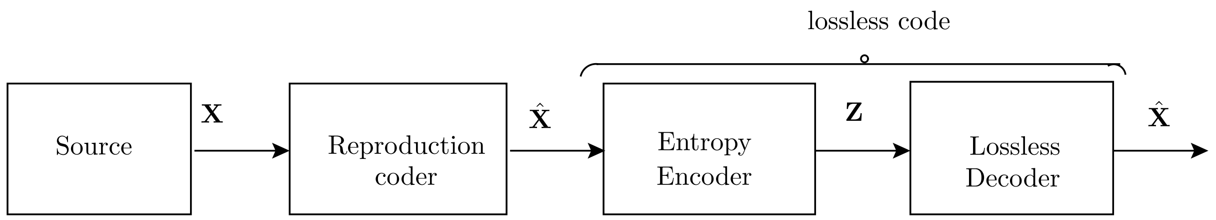

Figure 2: Equivalent representation of a source coding system

where $L_k(\mathbf{x})$ denotes the cumulative number of bits received by the decoder at the time it produces $\hat{x}_k$, when the source sequence is $\mathbf{x}$. In other words, if $z_1, z_2, \ldots$ is the sequence of bits produced by the encoder in response to $\mathbf{x}$, and if $\hat{x}_k$ is produced by the decoder after receiving z_l but before z_{l+1}, then $L_k(\mathbf{x}) = l$.

That the rate of a source code with a given reproduction coder can be improved (or not worsened) by concatenating the reproduction process with an appropriate lossless code follows from first information theoretic principles (cf. [12] for a formal proof). This implies no loss of optimality in confining attention to systems that first generate the reproduction process, and then losslessly encode it, as in Figure 2. This separation between the generation of the reconstruction and its lossless coding is more than merely conceptual. It is the basis for practical lossy source coding techniques [3].

Definition 1 *A reproduction coder is* causal *if for any $k \geq 1$,*

$$f_k(\mathbf{x}) = f_k(\tilde{\mathbf{x}}) \qquad \textit{whenever } x_{-\infty}^k = \tilde{x}_{-\infty}^k. \tag{3}$$

$\mathcal{F}_c$ will denote the class of all causal reproduction coders. A source code is causal if the reproduction coder it induces is causal.

A point to note is that in the above definition it is not required that causal source codes operate without introducing delay. This is because no restrictions are made on how the binary representation[2] $Z_1, Z_2, \ldots$ is generated by the encoder or on the way in which this sequence is used by the decoder. Thus, the class of delayless source codes is a (small) subset of the class of causal source codes.

Optimum performance theoretically attainable (OPTA) by causal source codes for the source $\mathbf{X}$ is given by $r_c(D)$, defined as the infimum of the average rates of all causal codes with average distortion $\leq D$. It was argued in [12] that $r_c(D)$ is determined solely by considering properties of reproduction coders (a consequence of the fact that for a general system of the type in Figure 1 there is one of the type in Figure 2 doing at least as well). Specifically, it was shown that $r_c(D)$ is given by

$$r_c(D) = \inf_{\{f_k\} \in \mathcal{F}_c : d(\{f_k\}) \leq D} H(\hat{\mathbf{X}}), \tag{4}$$

where

$$H(\hat{\mathbf{X}}) = \limsup_{k \to \infty} \frac{1}{k} H(\hat{X}_1^k) \tag{5}$$

denotes the lim sup entropy rate of the reproduction process.

[2]Note that each Z_i is allowed to depend on the whole of $\mathbf{X}$.

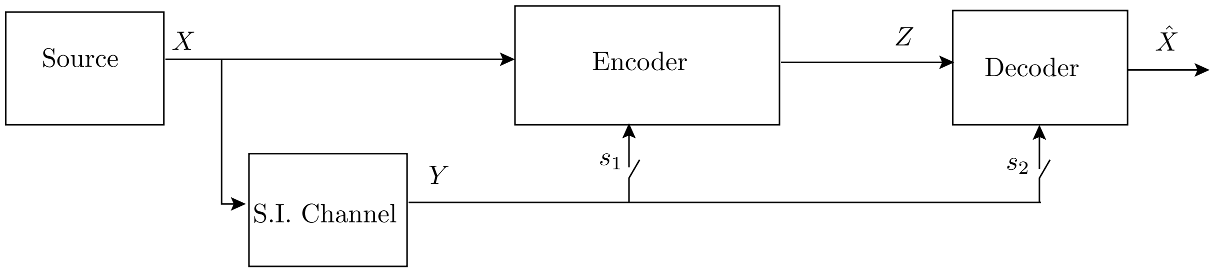

Figure 3: Source coding system with side information.

Define now

$$Q(P_X, D) = \inf_{f:\mathcal{X}\to\hat{\mathcal{X}}:Ed(X,f(X))\leq D} H(f(X)), \tag{6}$$

where $P_X \in \mathcal{M}(\mathcal{X})$ and X on the right side is distributed according to P_X. $Q(P_X, D)$ is nothing but the R-D function of the Entropy Constrained Scalar Quantizer (ECSQ) of the source P_X (cf. [4, 5] and references therein). The main result of [12] was the following.

Theorem 1 ([12]) *For $\mathbf{X} = \{X_k\}$, X_k i.i.d.$\sim P_X$,*

$$r_c(D) = \overline{Q}(P_X, D),$$

where $\overline{Q}(P_X, \cdot)$ denotes the lower convex hull of $Q(P_X, \cdot)$.

In words, time sharing (at most two) scalar quantizers, followed by lossless entropy coding, is optimal.

2-C Availability of Side Information

In this work we revisit the causal source coding setting of [12] for the case where side information is available (Figure 3). Thus, here and throughout we assume, in addition to the source $\mathbf{X}$ of the previous section, a side information sequence $\mathbf{Y}$, with components in the finite alphabet $\mathcal{Y}$, jointly distributed with $\mathbf{X}$. For this case, in general, the reproduction sequence will depend on the side information as well as on the source sequence, i.e., the reproduction coder is now characterized by a family $\{f_k\}$ such that $\hat{X}_k = f_k(\mathbf{X}, \mathbf{Y})$. We make a distinction between an induced reproduction coder depending causally on both the source and the side information and one depending causally on the source yet non-causally on the side information.

Definition 2 *A reproduction coder with side information is semi-causal if for any $k \geq 1$ and $\mathbf{y}$,*

$$f_k(\mathbf{x}, \mathbf{y}) = f_k(\tilde{\mathbf{x}}, \mathbf{y}) \quad \text{whenever } x_{-\infty}^k = \tilde{x}_{-\infty}^k. \tag{7}$$

A reproduction coder with side information is causal if for any $k \geq 1$,

$$f_k(\mathbf{x}, \mathbf{y}) = f_k(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}) \quad \text{whenever } x_{-\infty}^k = \tilde{x}_{-\infty}^k, y_{-\infty}^k = \tilde{y}_{-\infty}^k. \tag{8}$$

$\mathcal{F}_{si}, \mathcal{F}_{csi}$ will denote, respectively, the class of reproduction coders with side information that are semi-causal and causal.

When a switch s_1 or s_2 is closed (cf. Figure 3), we assume that the side information sequence is available to the corresponding system component in its entirety before the system begins operating. Note that it makes no sense to restrict the availability of the side information sequence to causality or finite delay of any type, since no such limitations are placed on the availability of the source sequence to the encoder (as in the original Neuhoff-Gilbert setting).

$R_{x,s_1,s_2}(D)$ will represent the analogue of $r_c(D)$, from the previous subsection, corresponding to the case where side information is available according to a switch configuration $s_1, s_2 \in \{0, 1\}$ (with 0 and 1 denoting, respectively, an open and a closed switch). More specifically, $R_{x,s_1,s_2}(D)$ will denote the infimum of the average rate attainable by a causal source coding system as in Figure 3 with a switch configuration corresponding to $s_1 s_2$, subject to the average distortion constraint D. Here $x \in \{0, csi, si\}$ according to whether the induced causal reproduction coder is not allowed to depend on side information ($x = 0$), is allowed causal dependence on the side information ($x = csi$), or non-causal dependence on the side information ($x = si$).

For the cases we address (in Sections 3 through 5 below) it will be seen that a "separation" holds, analogous to that leading from the systems of Figure 1 to systems as in Figure 2. It can be argued that this separation is what gave the result of [12] and, by analogy, what gives our results their operational significance. Viewed alternatively, our results can be seen as characterizing optimum performance for "two-stage" source coding systems that operate by first producing the reconstruction sequence, and then losslessly encoding it. Practical considerations may then dictate causal dependence of the reconstruction on the source sequence (as in [12]), as well as, possibly, causal dependence on the side information, or outright independence of it (if, say, the first stage is performed at a location where there is no access to the side information or if complexity considerations dictate either causal dependence or independence).

3 The Case $s_1 = s_2 = 1$

Assume throughout this section the case where both switches in Figure 3 are closed. Our goal is to characterize optimum performance when the reconstructed sequence is constrained to causal dependence on the source sequence, for the few possibilities of the way it is allowed to depend on the side information sequence. Arguing analogously as in Subsection 2-B, there is no loss of optimality in confining attention to systems that start by generating the reproduction process, followed by lossless coding (this time using the side information at both encoder and decoder), as in Figure 4. The switch on the wire connecting the side information to the reproduction coder can be in three different modes according to the value of x in $R_{x,1,1}$: completely open, closed but releasing the side information causally (i.e., y_t is released only after $\hat{x}_{t-1}$ has been produced), and closed with the whole side information sequence released upfront.

3-A Information Theoretic Representation of the OPTA Functions

The OPTA functions $R_{x,1,1}(D)$, $x \in \{0, csi, si\}$, can be given an information theoretic representation, analogous to (4). Indeed, if $\hat{\mathbf{X}}$ is the reconstruction process associated with any source code for this case then, by the converse to

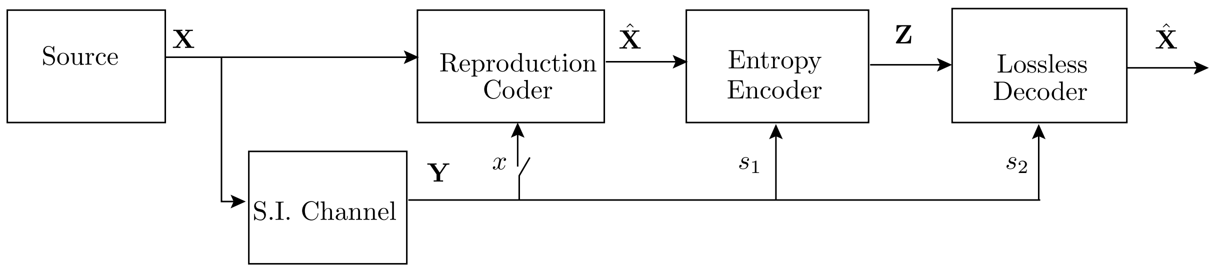

Figure 4: Alternative source coding system for the case $s_1 = s_2 = 1$.

lossless source coding with side information, $H(\hat{\mathbf{X}}|\mathbf{Y}) = \limsup_{k\to\infty} \frac{1}{k} H(\hat{X}_1^k|\mathbf{Y})$ lower bounds its average rate. On the other hand, if the average rate of a source code with reconstruction process $\hat{\mathbf{X}}$ is not close to $H(\hat{\mathbf{X}}|\mathbf{Y})$, it can be replaced by one with the same induced reproduction coder, but with average rate arbitrarily close to $H(\hat{\mathbf{X}}|\mathbf{Y})$ (by concatenation of a lossless entropy coder accessing the side information). This implies that

$$R_{x,1,1}(D) = \inf H(\hat{\mathbf{X}}|\mathbf{Y}), \tag{9}$$

where the infimum is over $\{\{f_k\} \in \mathcal{F}_c : d(\{f_k\}) \leq D\}$ if $x = 0$, over $\{\{f_k\} \in \mathcal{F}_{csi} : d(\{f_k\}) \leq D\}$ if $x = csi$, and over $\{\{f_k\} \in \mathcal{F}_{si} : d(\{f_k\}) \leq D\}$ if $x = si$.

3-B Statement of Results

For $P_{X,Y} \in \mathcal{M}(\mathcal{X} \times \mathcal{Y})$ and $D \geq 0$ define

$$Q_{si}(P_{X,Y}, D) = \inf_{\rho:\mathcal{Y}\to\mathbb{R}^+ : \int \rho(y) dP_Y(y) \leq D} \int_{\mathcal{Y}} \overline{Q}\left(P_{X|Y=y}, \rho(y)\right) dP_Y(y), \tag{10}$$

where P_Y and $P_{X|Y=y}$ on the right side denote, respectively, the Y-marginal and the conditional distribution of X given $Y = y$, both induced by $P_{X,Y}$ (and Q on the right side was defined in 6). To see the operative significance of $Q_{si}(P_{X,Y}, D)$ fix $\rho : \mathcal{Y} \to \mathbb{R}^+$ and consider, for each $y \in \mathcal{Y}$, a time-sharing of at most two scalar quantizers for the source $P_{X|Y=y}$ attaining distortion level $\rho(y)$. The minimum attainable entropy rate for the quantizer output process when the source is $P_{X|Y=y}$ is then $\overline{Q}\left(P_{X|Y=y}, \rho(y)\right)$. Suppose now that $(\mathbf{X}, \mathbf{Y})$ are formed by i.i.d. drawings from $P_{X,Y}$. Since the proportion of appearances of a symbol $y \in \mathcal{Y}$ is going to be $\approx P_Y(y)$, by time-sharing such side-information-dependent scalar quantizers the minimum attainable conditional entropy rate (conditioned on the side information) of the quantizers output process will be given by $\int_{\mathcal{Y}} \overline{Q}\left(P_{X|Y=y}, \rho(y)\right) dP_Y(y)$, while the resulting distortion will be given by $\int \rho(y) dP_Y(y)$. Thus, $Q_{si}(P_{X,Y}, D)$ denotes the minimum conditional entropy rate attainable by such side-information dependent time-sharing of scalar quantizers subject to a distortion constraint D.

Theorem 2 *Let $(\mathbf{X}, \mathbf{Y})$ be, respectively, the output of a memoryless channel and its stationary ergodic input. Then*

$$R_{si,1,1}(D) = R_{csi,1,1}(D) = Q_{si}(P_{X_1,Y_1}, D). \tag{11}$$

In words (made precise in the formal proof of the direct part in the next subsection), time-sharing side-information dependent scalar quantizers (at most two per each side information symbol) attains optimum performance when the reproduction sequence is allowed causal, or even non-causal, dependence on the side information.

The fact that optimal performance is attained by memoryless quantizers in this generality, where the pair process (X_i, Y_i) may be far from memoryless and where the reproduction process may depend non-causally on the side information, may seem surprising at first glance. To gain some intuition note that, conditioned on the side information sequence, $\mathbf{X}$ is an arbitrarily varying source, the side information sequence being the state sequence. Thus, since in the case considered here all system components access the side information sequence, for each realization of that sequence the problem reduces to that of causal source coding in the original Neuhoff-Gilbert sense for the arbitrarily varying source (AVS) with a known state sequence. The proof of the main result of [12] turns out to be extendable to the AVS (due to the key feature which remains, namely, the independence of the components), asserting that optimum performance is attained by state-dependent memoryless quantizers, time-sharing of at most two for each state.

Turning to the remaining case where the reproduction coder does not access the side information, define, for $P_{X,Y} \in \mathcal{M}(\mathcal{X} \times \mathcal{Y})$,

$$Q_{01}(P_{X,Y}, D) = \min_{f:Ed(X,f(X))\leq D} H(f(X)|Y), \tag{12}$$

where the expectation and conditional entropy functionals on the right side of (12) assume $(X, Y) \sim P_{X,Y}$.

Theorem 3 *For $(\mathbf{X}, \mathbf{Y})$ such that (X_i, Y_i) are i.i.d. $\sim P_{X,Y}$,*

$$R_{0,1,1}(D) = \overline{Q}_{01}(P_{X,Y}, D),$$

where $\overline{Q}_{01}(P_{X,Y}, \cdot)$ denotes the lower convex hull of $Q_{01}(P_{X,Y}, \cdot)$.

In words (which are made precise in the formal proof of the direct part in Appendix A), optimum performance in this case is attained by time sharing at most two scalar quantizers, as in the Neuhoff-Gilbert setting. As the proof of the converse shows, the result remains intact even if the reproduction coder is allowed to depend on past (but not present) side information symbols, i.e., allowing such dependence will not result in improved performance.

The following subsection is dedicated to the proof of Theorem 2. The proof of Theorem 3, which has similar components, is deferred to the Appendix.

3-C Proof of Theorem 2

Proof of Converse:

Lemma 1 *For any $P_{X,Y} \in \mathcal{M}(\mathcal{X} \times \mathcal{Y})$, $Q_{si}(P_{X,Y}, \cdot)$ is convex.*

Proof of Lemma 1: Fix an arbitrary $\varepsilon > 0$, $\lambda \in [0, 1]$ and $D_1, D_2 \geq 0$. It will suffice to show that

$$Q_{si}(P_{X,Y}, \lambda D_1 + (1 - \lambda)D_2) \leq \lambda Q_{si}(P_{X,Y}, D_1) + (1 - \lambda)Q_{si}(P_{X,Y}, D_2) + \varepsilon. \tag{13}$$

Let $\rho_i^*:\mathcal{Y}\to\mathbb{R}^+$ denote an ε-achiever of $Q_{si}(P_{X,Y},D_i)$, $i=1,2$, i.e.,

$$\int_{\mathcal{Y}}\overline{Q}\left(P_{X|Y=y},\rho_i^*(y)\right)dP_Y(y)\le Q_{si}(P_{X,Y},D_i)+\varepsilon, \tag{14}$$

and

$$\int\rho_i^*(y)dP_Y(y)\le D_i. \tag{15}$$

Define now $\rho:\mathcal{Y}\to\mathbb{R}^+$ via $\rho(y)=\lambda\rho_1^*(y)+(1-\lambda)\rho_2^*(y)$ so that, by (15),

$$\int\rho(y)dP_Y(y)\le\lambda D_1+(1-\lambda)D_2. \tag{16}$$

Consequently,

$$\begin{aligned}
Q_{si}(P_{X,Y},\lambda D_1+(1-\lambda)D_2) &\le \int_{\mathcal{Y}}\overline{Q}\left(P_{X|Y=y},\rho(y)\right)dP_Y(y)\\
&= \int_{\mathcal{Y}}\overline{Q}\left(P_{X|Y=y},\lambda\rho_1^*(y)+(1-\lambda)\rho_2^*(y)\right)dP_Y(y)\\
&\le \lambda\int_{\mathcal{Y}}\overline{Q}\left(P_{X|Y=y},\rho_1^*(y)\right)dP_Y(y)+(1-\lambda)\int_{\mathcal{Y}}\overline{Q}\left(P_{X|Y=y},\rho_2^*(y)\right)dP_Y(y)\\
&\le \lambda Q_{si}(P_{X,Y},D_1)+(1-\lambda)Q_{si}(P_{X,Y},D_2)+\varepsilon,
\end{aligned}$$

where the first inequality follows from (16) and the definition of Q_{si}, the second inequality follows from the convexity of $\overline{Q}$, and the last one is due to (14). $\square$

Proof of Converse Part of Theorem 2: Assume $(\mathbf{X},\mathbf{Y})$ as in the statement of the theorem. Our goal is to show that $R_{si,1,1}(D)\ge Q_{si}(P_{X_1,Y_1},D)$. To this end, fix a reproduction coder $\{f_k\}_{k=1}^{\infty}\in\mathcal{F}_{si}$ for which

$$d(\{f_k\})\le D. \tag{17}$$

It will suffice to show that

$$H(\hat{\mathbf{X}}|\mathbf{Y})\ge Q_{si}(P_{X_1,Y_1},D). \tag{18}$$

For $k\ge 1$,

$$\begin{aligned}
H(\hat{X}_1^k|\mathbf{Y}) &= \sum_{i=1}^{k}H(\hat{X}_i|\hat{X}_1^{i-1},\mathbf{Y})\\
&\ge \sum_{i=1}^{k}H(\hat{X}_i|X^{i-1},\mathbf{Y})\\
&= \sum_{i=1}^{k}H(f_i(X^i,\mathbf{Y})|X^{i-1},\mathbf{Y}),
\end{aligned} \tag{19}$$

where the inequality follows since for the causal source code with non-causal side information $(\hat{X}_1^{i-1},\mathbf{Y})$ is uniquely determined by $(X^{i-1},\mathbf{Y})$. Now,

$$\begin{aligned}
&H(f_i(X^i,\mathbf{Y})|X^{i-1},\mathbf{Y})\\
=\ &\int H\left(f_i((x^{i-1}\ X_i),\mathbf{y})|X^{i-1}=x^{i-1},\mathbf{Y}=\mathbf{y}\right)d\mu(x^{i-1},\mathbf{y})
\end{aligned}$$

$$
\begin{aligned}
&= \int H\left(f_i((x^{i-1}\ X_i),\mathbf{y})\middle|\, Y_i=y_i\right) d\mu(x^{i-1},\mathbf{y}) && (20)\\
&= \int \left[\int H\left(f_i((x^{i-1}\ X_i),\mathbf{y})\middle|\, Y_i=y_i\right) d\mu(x^{i-1},y^{i-1},y_{i+1}^{\infty}|y_i)\right] d\mu(y_i)\\
&\geq \int \left[\int \overline{Q}\left(P_{X_i|Y_i=y_i}, E[d(X_i,f_i((x^{i-1}\ X_i),\mathbf{y})|Y_i=y_i]\right) d\mu(x^{i-1},y^{i-1},y_{i+1}^{\infty}|y_i)\right] d\mu(y_i) && (21)\\
&\geq \int \overline{Q}\left(P_{X_i|Y_i=y_i}, \int E[d(X_i,f_i((x^{i-1}\ X_i),\mathbf{y})|Y_i=y_i] d\mu(x^{i-1},y^{i-1},y_{i+1}^{\infty}|y_i)\right) d\mu(y_i) && (22)\\
&= \int \overline{Q}\left(P_{X_i|Y_i=y_i}, E[d(X_i,f_i(X^i,\mathbf{Y}))|Y_i=y_i]\right) d\mu(y_i)\\
&\geq Q_{si}\left(P_{X_i,Y_i}, \int E[d(X_i,f_i(X^i,\mathbf{Y}))|Y_i=y_i] d\mu(y_i)\right)\\
&= Q_{si}\left(P_{X_i,Y_i}, Ed(X_i,f_i(X^i,\mathbf{Y}))\right) && (23)\\
&= Q_{si}\left(P_{X_1,Y_1}, Ed(X_i,\hat{X}_i)\right) && (24)
\end{aligned}
$$

where (20) follows since, by hypothesis (that $\mathbf{X}$ is the output of a memoryless channel whose input is $\mathbf{Y}$), X_i is conditionally independent of $(X^{i-1},Y^{i-1},Y_{i+1}^{\infty})$ given Y_i. Inequality (21) follows from the definition of Q and (22) follows from convexity. Inequality (23) follows directly from the definition of Q_{si} and the equality in (24) is due to stationarity. Consequently,

$$
\begin{aligned}
H(\hat{\mathbf{X}}|\mathbf{Y}) &= \limsup_{k\to\infty} \frac{1}{k} H(\hat{X}_1^k|\mathbf{Y})\\
&\geq \limsup_{k\to\infty} \frac{1}{k}\sum_{i=1}^{k} H(f_i(X^i,\mathbf{Y})|X^{i-1},\mathbf{Y}) && (25)\\
&\geq \limsup_{k\to\infty} \frac{1}{k}\sum_{i=1}^{k} Q_{si}\left(P_{X_1,Y_1}, Ed(X_i,f_i(X^i,\mathbf{Y}))\right) && (26)\\
&\geq \limsup_{k\to\infty} Q_{si}\left(P_{X_1,Y_1}, \frac{1}{k}\sum_{i=1}^{k} Ed(X_i,\hat{X}_i)\right) && (27)\\
&\geq Q_{si}\left(P_{X_1,Y_1}, \limsup_{k\to\infty}\frac{1}{k}\sum_{i=1}^{k} Ed(X_i,\hat{X}_i)\right) && (28)\\
&= Q_{si}\left(P_{X_1,Y_1}, d(\{f_k\})\right)\\
&\geq Q_{si}\left(P_{X_1,Y_1}, D\right), && (29)
\end{aligned}
$$

where (25) follows from (19), (26) follows from (24), (27) is due to the convexity of $Q_{si}(P_{X_1,Y_1},\cdot)$ (recall Lemma 1), (28) is due to the monotonicity of $Q_{si}(P_{X_1,Y_1},\cdot)$, and (29) follows from (17). This establishes (18) and completes the proof. □

The proof of the direct part is straightforward, formally establishing the fact that time-sharing side-information-dependent scalar quantizers (at most two per each side information symbol) attains $Q_{si}(P_{X_1,Y_1},D)$, when $\mathbf{X}$ is the output of a memoryless channel whose input is $\mathbf{Y}$.

Proof of Direct: Fix an arbitrary $\varepsilon>0$ and let $\{\rho(y)\}_{y\in\mathcal{Y}}$ be an ε-achiever of $Q_{si}(P_{X_1,Y_1},D)$, so that both

$$\sum_{y\in\mathcal{Y}} \rho(y)\Pr(Y_1=y) \leq D \tag{30}$$

and

$$\sum_{y\in\mathcal{Y}} \overline{Q}\left(P_{X_1|Y_1=y}, \rho(y)\right) \Pr(Y_1 = y) \le Q_{si}(P_{X_1,Y_1}, D) + \varepsilon \tag{31}$$

hold. For each $y \in \mathcal{Y}$, by the definition of $\overline{Q}\left(P_{X_1|Y_1=y}, \rho(y)\right)$, there exist $\lambda_y \in [0,1]$ and $D_y^{(0)}, D_y^{(1)} \ge 0$ such that

$$\overline{Q}\left(P_{X_1|Y_1=y}, \rho(y)\right) = \lambda_y Q\left(P_{X_1|Y_1=y}, D_y^{(0)}\right) + (1-\lambda_y) Q\left(P_{X_1|Y_1=y}, D_y^{(1)}\right) \tag{32}$$

and

$$\rho(y) = \lambda_y D_y^{(0)} + (1-\lambda_y) D_y^{(1)}. \tag{33}$$

Let further $f_y^{(0)}, f_y^{(1)}$ denote respective ε-achievers of $Q\left(P_{X_1|Y_1=y}, D_y^{(0)}\right)$ and $Q\left(P_{X_1|Y_1=y}, D_y^{(1)}\right)$ so that

$$E[d(X_1, f_y^{(0)}(X_1))|Y_1 = y] \le D_y^{(0)}, \qquad E[d(X_1, f_y^{(1)}(X_1))|Y_1 = y] \le D_y^{(1)} \tag{34}$$

and

$$H(f_y^{(0)}(X_1)|Y_1 = y) \le Q\left(P_{X_1|Y_1=y}, D_y^{(0)}\right) + \varepsilon, \qquad H(f_y^{(1)}(X_1)|Y_1 = y) \le Q\left(P_{X_1|Y_1=y}, D_y^{(1)}\right) + \varepsilon. \tag{35}$$

For each $y \in \mathcal{Y}$ let now $\{b_i^{(y)}\}$ be a deterministic binary sequence such that

$$\frac{1}{k}\sum_{i=1}^{k} b_i^{(y)} \stackrel{k\to\infty}{\longrightarrow} 1 - \lambda_y. \tag{36}$$

For $k \ge 1$ let further:

$$N(Y^k) = |\{1 \le i \le k : Y_i = Y_k\}|. \tag{37}$$

Finally, let the reproduction coder be defined by

$$\hat{X}_k = f_{Y_k}^{\left(b^{(Y_k)}_{N(Y^k)}\right)}(X_k) \tag{38}$$

(so that, in particular, it is a member of $\mathcal{F}_{csi}$). Now, for $\mathbf{y} \in \mathcal{Y}^\infty$,

$$\begin{aligned}
& E[d_k(X_1^k, \hat{X}_1^k)|\mathbf{Y} = \mathbf{y}] \\
= \; & \frac{1}{k}\sum_{i=1}^{k} E[d(X_i, \hat{X}_i)|\mathbf{Y} = \mathbf{y}] \\
= \; & \frac{1}{k}\sum_{i=1}^{k} E\left[d\left(X_i, f_{y_i}^{\left(b^{(y_i)}_{N(y^i)}\right)}(X_i)\right) \middle| \mathbf{Y} = \mathbf{y}\right] \\
= \; & \frac{1}{k}\sum_{i=1}^{k} E\left[d\left(X_i, f_{y_i}^{\left(b^{(y_i)}_{N(y^i)}\right)}(X_i)\right) \middle| Y_i = y_i\right] & (39) \\
= \; & \sum_{y\in\mathcal{Y}} E\left[d\left(X_1, f_y^{(0)}(X_1)\right) \middle| Y_1 = y\right] \frac{|\{1 \le i \le k : y_i = y, b^{(y)}_{N(y^i)} = 0\}|}{k} \\
& + \sum_{y\in\mathcal{Y}} E\left[d\left(X_1, f_y^{(1)}(X_1)\right) \middle| Y_1 = y\right] \frac{|\{1 \le i \le k : y_i = y, b^{(y)}_{N(y^i)} = 1\}|}{k}, & (40)
\end{aligned}$$

where (39) follows by the conditional independence of X_i and $(Y^{i-1}, Y_{i+1}^{\infty})$ given Y_i, and (40) follows by stationarity. The ergodicity of $\mathbf{Y}$, combined with (36), imply that for $P_{\mathbf{Y}}$-almost every $\mathbf{y}$ and all $y \in \mathcal{Y}$

$$\frac{|\{1 \le i \le k : y_i = y, b^{(y)}_{N(y^i)} = 0\}|}{k} \stackrel{k\to\infty}{\longrightarrow} \lambda_y \Pr(Y_1 = y), \qquad \frac{|\{1 \le i \le k : y_i = y, b^{(y)}_{N(y^i)} = 1\}|}{k} \stackrel{k\to\infty}{\longrightarrow} (1-\lambda_y) \Pr(Y_1 = y). \tag{41}$$

Combining (41) with (40) gives, for $P_{\mathbf{Y}}$-almost every $\mathbf{y}$,

$$\begin{aligned}
&\lim_{k\to\infty} \left[E[d_k(X_1^k, \hat{X}_1^k)|\mathbf{Y}=\mathbf{y}]\right] \\
= &\sum_{y\in\mathcal{Y}} \Pr(Y_1 = y) \left(E\left[d\left(X_1, f_y^{(0)}(X_1)\right)\Big| Y_1 = y\right]\lambda_y + E\left[d\left(X_1, f_y^{(1)}(X_1)\right)\Big| Y_1 = y\right](1-\lambda_y)\right). \qquad (42)
\end{aligned}$$

Consequently,

$$\begin{aligned}
&\limsup_{k\to\infty} E d_k(X_1^k, \hat{X}_1^k) \\
= &\limsup_{k\to\infty} E\left[E[d_k(X_1^k, \hat{X}_1^k)|\mathbf{Y}]\right] \\
\le & E \lim_{k\to\infty} \left[E[d_k(X_1^k, \hat{X}_1^k)|\mathbf{Y}]\right] && (43)\\
= &\sum_{y\in\mathcal{Y}} \Pr(Y_1 = y) \left(E\left[d\left(X_1, f_y^{(0)}(X_1)\right)\Big| Y_1 = y\right]\lambda_y + E\left[d\left(X_1, f_y^{(1)}(X_1)\right)\Big| Y_1 = y\right](1-\lambda_y)\right) && (44)\\
\le &\sum_{y\in\mathcal{Y}} \Pr(Y_1 = y) \left(D_y^{(0)}\lambda_y + D_y^{(1)}(1-\lambda_y)\right) && (45)\\
= &\sum_{y\in\mathcal{Y}} \Pr(Y_1 = y)\rho(y) && (46)\\
\le & D, && (47)
\end{aligned}$$

where (43) follows by Fatou's lemma, (44) follows from (42), (45) from (34), (46) from (33), and (47) from (30).

The conditional entropy rate can now be bounded analogously. For each $\mathbf{y} \in \mathcal{Y}^\infty$,

$$\begin{aligned}
\frac{1}{k} H(\hat{X}_1^k|\mathbf{Y}=\mathbf{y}) &= \frac{1}{k} H\left(\left(f_{y_1}^{\left(b^{(y_1)}_{N(y^1)}\right)}(X_1), \ldots, f_{y_k}^{\left(b^{(y_k)}_{N(y^k)}\right)}(X_k)\right)\Bigg| \mathbf{Y}=\mathbf{y}\right) \\
&= \frac{1}{k}\sum_{i=1}^k H\left(f_{y_i}^{\left(b^{(y_i)}_{N(y^i)}\right)}(X_i)\Bigg| \mathbf{Y}=\mathbf{y}\right) && (48)\\
&= \frac{1}{k}\sum_{i=1}^k H\left(f_{y_i}^{\left(b^{(y_i)}_{N(y^i)}\right)}(X_i)\Bigg| Y_i=y_i\right) && (49)\\
&= \sum_{y\in\mathcal{Y}} H\left(f_y^{(0)}(X_1)\Big| Y_1 = y\right)\frac{|\{1 \le i \le k : y_i = y, b^{(y)}_{N(y^i)} = 0\}|}{k} \\
&\quad + \sum_{y\in\mathcal{Y}} H\left(f_y^{(1)}(X_1)\Big| Y_1 = y\right)\frac{|\{1 \le i \le k : y_i = y, b^{(y)}_{N(y^i)} = 1\}|}{k}, && (50)
\end{aligned}$$

where (48) is due to the memorylessness of the process $\mathbf{X}$ when conditioned on $\mathbf{Y}$, (49) follows by the conditional independence of X_i and $(Y^{i-1}, Y_{i+1}^{\infty})$ given Y_i, and (50) follows by stationarity. By (41) it now follows that for

$P_{\mathbf{Y}}$-almost every $\mathbf{y}$

$$\lim_{k\to\infty}\frac{1}{k}H(\hat{X}_1^k|\mathbf{Y}=\mathbf{y})=\sum_{y\in\mathcal{Y}}\Pr(Y_1=y)\left[H\left(f_y^{(0)}(X_1)\middle|Y_1=y\right)\lambda_y+H\left(f_y^{(1)}(X_1)\middle|Y_1=y\right)(1-\lambda_y)\right]. \tag{51}$$

Consequently,

$$\begin{aligned}
H(\hat{\mathbf{X}}|\mathbf{Y}) &= \limsup_{k\to\infty}\frac{1}{k}H(\hat{X}_1^k|\mathbf{Y}) \\
&= \limsup_{k\to\infty}\frac{1}{k}\int H(\hat{X}_1^k|\mathbf{Y}=\mathbf{y})d\mu(\mathbf{y}) \\
&\le \int\left[\limsup_{k\to\infty}\frac{1}{k}H(\hat{X}_1^k|\mathbf{Y}=\mathbf{y})\right]d\mu(\mathbf{y}) && (52)\\
&= \sum_{y\in\mathcal{Y}}\Pr(Y_1=y)\left[H\left(f_y^{(0)}(X_1)\middle|Y_1=y\right)\lambda_y+H\left(f_y^{(1)}(X_1)\middle|Y_1=y\right)(1-\lambda_y)\right] && (53)\\
&\le \varepsilon+\sum_{y\in\mathcal{Y}}\Pr(Y_1=y)\left[Q\left(P_{X_1|Y_1=y},D_y^{(0)}\right)\lambda_y+Q\left(P_{X_1|Y_1=y},D_y^{(1)}\right)(1-\lambda_y)\right] && (54)\\
&= \varepsilon+\sum_{y\in\mathcal{Y}}\Pr(Y_1=y)\overline{Q}\left(P_{X_1|Y_1=y},\rho(y)\right) && (55)\\
&= Q_{si}(P_{X_1,Y_1},D)+2\varepsilon, && (56)
\end{aligned}$$

where (52) follows from Fatou's lemma, (53) follows from (51), (54) follows from (35), (55) follows from (32), and (56) follows from (31). The arbitrariness of $\varepsilon>0$ completes the proof. □

4 The Case $s_1 = 1$, $s_2 = 0$

4-A Information Theoretic Representation of the OPTA Functions

Consider the case where switch s_1 in Figure 3 is closed and s_2 is open. For any source code of this type, if $\hat{\mathbf{X}}$ denotes the reproduction sequence, then by the converse to lossless source coding the average rate is lower bounded by $H(\hat{\mathbf{X}})$. This is because, as is well-known (cf., e.g., [14]), side information at the encoder alone does not enhance compression performance[3]. On the other hand, if this particular source code does not come close to attaining rate $H(\hat{\mathbf{X}})$, it can be replaced by one which does, essentially by concatenation of a lossless entropy coder. Thus we have

$$R_{x,1,0}(D)=\inf H(\hat{\mathbf{X}}), \tag{57}$$

where, similarly as in (9), the infimum is over $\{\{f_k\}\in\mathcal{F}_c : d(\{f_k\})\le D\}$ if $x=0$, over $\{\{f_k\}\in\mathcal{F}_{csi} : d(\{f_k\})\le D\}$ if $x=csi$, and over $\{\{f_k\}\in\mathcal{F}_{si} : d(\{f_k\})\le D\}$ if $x=si$. This implies also that for this setting there is no loss of optimality in confining attention to systems of the type depicted in Figure 5 (where, as in the previous section, the switch is in one of three modes corresponding to the value of x).

4-B Statement of Result

The fact that side information at the encoder alone does not enhance performance in the classical source coding setting does not imply that this should be the case for causal source coding. Indeed, $R_{si,1,0}(D)$ will, in general, be

[3] The side information sequence in this case can be thought of as (useless) extra randomness available to the encoder.

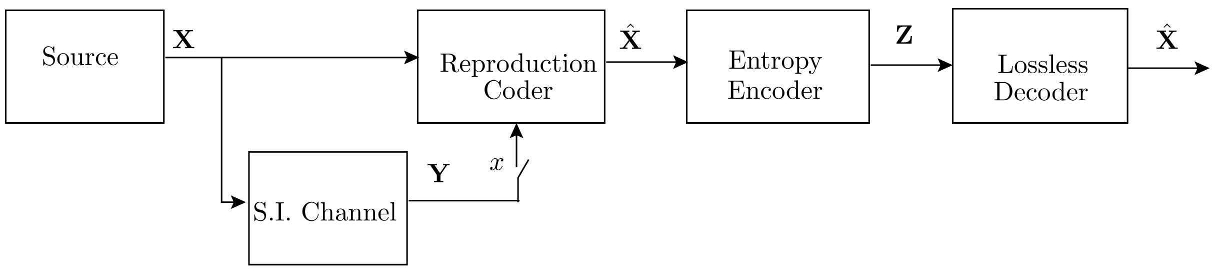

Figure 5: Alternative source coding system for the case $s_1 = 1, s_2 = 0$.

strictly less than $r_c(D)$ from the original setting of [12]. To see this note that, for example, in the extreme case where the side information sequence perfectly reveals the source sequence (say $\mathbf{Y} \equiv \mathbf{X}$), $R_{si,1,0}(D)$ is nothing but the classical rate distortion function of the source $\mathbf{X}$. As it turns out, however, when the dependence on the side information in the reproduction coder is constrained to causality, the side information can be discarded, with no degradation in performance. More formally:

Theorem 4 *For* $(\mathbf{X}, \mathbf{Y})$ *such that* (X_i, Y_i) *are i.i.d.* $\sim P_{X,Y}$, $R_{csi,1,0}(D) = \overline{Q}(P_X, D)$.

4-C Proof of Theorem 4

The direct part of this result, given Theorem 1 is, of course, immediate as, by definition, $R_{csi,1,0}(D) \leq r_c(D)$. Turning to the proof of the converse, define

$$Q_r(P_X, D) = \inf_{f: \mathcal{U} \times \mathcal{X} \to \hat{\mathcal{X}} : Ed(X, f(U,X)) \leq D} H(f(U, X)), \tag{58}$$

where the infimum is over all joint distributions of the pair (U, X) consistent with the P_X-marginal for X, where U can take values in any measurable space $\mathcal{U}$, as well as over all $f : \mathcal{U} \times \mathcal{X} \to \hat{\mathcal{X}}$ satisfying the indicated constraint.
Claim 1

$$\overline{Q}_r(P_X, D) = \overline{Q}(P_X, D),$$

$\overline{Q}_r(P_X, \cdot)$ *denoting the lower convex hull of* $Q_r(P_X, \cdot)$.

In words, when time-sharing is allowed, randomized scalar quantizers are not advantageous to non-randomized ones.
Proof of Claim 1: Trivially, by the definitions, $Q_r(P_X, D) \leq Q(P_X, D)$ and therefore $\overline{Q}_r(P_X, D) \leq \overline{Q}(P_X, D)$. It will therefore suffice to show that

$$Q_r(P_X, D) \geq \overline{Q}(P_X, D). \tag{59}$$

To this end, assume $Q_r(P_X, D) < \infty$ (otherwise there is nothing to prove), fix $\varepsilon > 0$, and let $(P_{X,U}, f)$ denote an ε-achiever of $Q_r(P_X, D)$, i.e., assuming $(X, U) \sim P_{X,U}$ and denoting $Z = f(U, X)$,

$$H(Z) \leq Q_r(P_X, D) + \varepsilon, \quad Ed(X, Z) \leq D. \tag{60}$$

Since Z is discrete-valued, it follows that $P_{Z|X=x}$ is a distribution of a discrete-valued r.v. for P_X-a.e. x. For $U' \sim U[0, 1]$ independent of X and all such x let $g_x : [0, 1] \to \hat{\mathcal{X}}$ be a mapping such that $g_x(U') \sim P_{Z|X=x}$ (such a mapping is of course given by the pseudo-inverse of the cumulative distribution function associated with $P_{Z|X=x}$). Define now $g : [0,1] \times \mathcal{X} \to \hat{\mathcal{X}}$ via $g(u', x) = g_x(u')$. Clearly $(X, g(U', X))$ and (X, Z) are identically distributed and therefore

$$H(g(U', X)) \leq Q_r(P_X, D) + \varepsilon, \quad Ed(X, g(U', X)) \leq D. \tag{61}$$

Thus, letting $\mathcal{A} \subseteq \hat{\mathcal{X}}$ denote the discrete set where $g(U', X)$ takes its values,

$$\begin{aligned}
Q_r(P_X, D) + \varepsilon &\geq H(g(U', X)) \\
&= -\sum_{a \in \mathcal{A}} \Pr(g(U', X) = a) \log \Pr(g(U', X) = a) \\
&= -\sum_{a \in \mathcal{A}} \left[\int_0^1 \Pr(g(u', X) = a) du' \right] \log \left[\int_0^1 \Pr(g(u', X) = a) du' \right] \\
&\geq -\sum_{a \in \mathcal{A}} \int_0^1 \Pr(g(u', X) = a) \log \Pr(g(u', X) = a) du' \\
&= \int_0^1 H(g(u', X)) du' \\
&\geq \int_0^1 \overline{Q}(P_X, Ed(X, g(u', X))) du' \\
&\geq \overline{Q}\left(P_X, \int_0^1 Ed(X, g(u', X)) du' \right) \\
&= \overline{Q}\left(P_X, Ed(X, g(U', X))\right) \\
&\geq \overline{Q}\left(P_X, D\right),
\end{aligned}$$

where the first and the last inequalities follow from (61), the second inequality from convexity of the function $x \log x$, the third inequality from the definition of the function $Q(P_X, \cdot)$ and the fourth from the convexity of $\overline{Q}(P_X, \cdot)$. The arbitrariness of $\varepsilon > 0$ implies (59). $\square$

Proof of Converse Part of Theorem 4: We need to prove that $R_{csi,1,0}(D) \geq \overline{Q}(P_X, D)$. To this end, fix any reproduction coder $\{\hat{X}_k\}$ in $\mathcal{F}_{csi}$, given by $\hat{X}_k = f_k(X^k, Y^k)$, satisfying

$$d(\{f_k\}) \leq D. \tag{62}$$

It will suffice to show that

$$H(\hat{\mathbf{X}}) \geq \overline{Q}\left(P_X, D\right). \tag{63}$$

For $k \geq 1$,

$$\begin{aligned}
H(\hat{X}_k | \hat{X}_1^{k-1}) &\geq H(\hat{X}_k | X^{k-1}, Y^{k-1}) \\
&= H(f_k(X^k, Y^k) | X^{k-1}, Y^{k-1}) \\
&= \int H\left(f_k(x^{k-1}, X_k, y^{k-1}, Y_k) | X^{k-1} = x^{k-1}, Y^{k-1} = y^{k-1}\right) d\mu(x^{k-1}, y^{k-1}) \\
&= \int H\left(f_k(x^{k-1}, X_k, y^{k-1}, Y_k)\right) d\mu(x^{k-1}, y^{k-1}) \\
&\geq \int \overline{Q}_r\left(P_X, Ed(X_k, f_k(x^{k-1}, X_k, y^{k-1}, Y_k))\right) d\mu(x^{k-1}, y^{k-1}) \qquad (64)
\end{aligned}$$

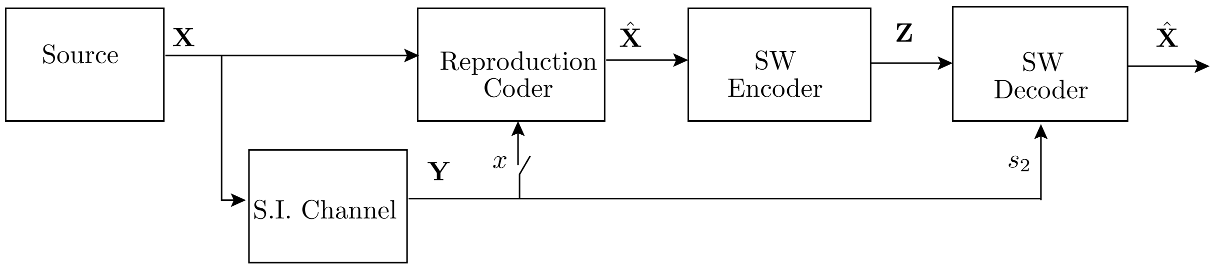

Figure 6: Reproduction coding followed by essentially lossless coding with the side information at the decoder only.

$$\geq \overline{Q}\left(P_X, \int Ed(X_k, f_k(x^{k-1}, X_k, y^{k-1}, Y_k)) d\mu(x^{k-1}, y^{k-1})\right) \quad (65)$$

$$= \overline{Q}\left(P_X, Ed(X_k, \hat{X}_k)\right), \quad (66)$$

where (64) follows from the definition of the function Q_r, (65) follows by convexity, and (66) from Claim 1. Consequently,

$$\begin{aligned}
H(\hat{\mathbf{X}}) &= \limsup_{n\to\infty} \frac{1}{n} H(\hat{X}_1^n) \\
&= \limsup_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} H(\hat{X}_k|\hat{X}_1^{k-1}) \\
&\geq \limsup_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} \overline{Q}\left(P_X, Ed(X_k, \hat{X}_k)\right) && (67) \\
&\geq \limsup_{n\to\infty} \overline{Q}\left(P_X, E\left[d_k(X_1^k, \hat{X}_1^k)\right]\right) && (68) \\
&= \overline{Q}\left(P_X, \liminf_{n\to\infty} E\left[d_k(X_1^k, \hat{X}_1^k)\right]\right) && (69) \\
&\geq \overline{Q}\left(P_X, \limsup_{n\to\infty} E\left[d_k(X_1^k, \hat{X}_1^k)\right]\right) && (70) \\
&= \overline{Q}\left(P_X, d(\{f_k\})\right) \\
&\geq \overline{Q}\left(P_X, D\right), && (71)
\end{aligned}$$

where (67) follows from (66), (68) by convexity, (69) and (70) by monotonicity and continuity of $\overline{Q}(P_X, \cdot)$, and (71) by monotonicity of $\overline{Q}(P_X, \cdot)$ and (62). This establishes (63). □

5 The Case $s_1 = 0, s_2 = 1$

Consider the problem of causal source coding when the *decoder* accesses the side information, yet the encoder does not. Recall first from Section 3 that when side information is available to both encoder and decoder, attention can be restricted to systems of the form depicted in Figure 4 (regardless of whether or not causality constraints are imposed on the reproduction). However, the performance of the system in Figure 4 can also be attained when the middle switch is open, as in Figure 6, using Slepian-Wolf coding [14, 1] in the lossless part in lieu of the conventional conditional entropy coding performed when the switch is closed. This also leads to:

Observation 1 $R_{0,0,1}(D) = R_{0,1,1}(D) = \overline{Q}_{01}(P_{X,Y}, D)$.

Proof: For the converse note that trivially $R_{0,0,1}(D) \geq R_{0,1,1}(D)$ and invoke Theorem 3. For the direct part, use the same time sharing of at most two scalar quantizers used in the proof of the direct part of Theorem 3 (in Appendix A below) to generate the side-information-independent reconstruction sequence. Then use Slepian-Wolf coding [14, 1] to essentially losslessly encode the reproduction sequence using the side information at the decoder only. □

Note that the reason why the system of Figure 6 is relevant in the setting corresponding to $R_{0,0,1}(D)$ (in which case the x-switch is open) is that in this case the reproduction coder and the lossless encoder jointly play the role of an encoder that has no access to the side information, thus forming a system complying with the allowed structure, as depicted in Figure 3 with an open s_1 and a closed s_2. In contrast, when the x-switch in Figure 6 is closed (causally or non-causally), the reproduction coder and lossless encoder cannot be thought of as jointly playing the role of the encoder in Figure 3 (with an open s_1) since their composition will, in general, depend on the side information. Thus, the "separation" between reproduction coding and lossless coding that has prevailed throughout all the scenarios considered thus far breaks down in the settings pertaining to $R_{csi,0,1}(D)$ and $R_{si,0,1}(D)$, the characterization of which remains open.

6 Summary and Open Questions

We revisited the causal source coding setting of Neuhoff and Gilbert for the case where side information is present. For the scenarios considered, the optimality of time sharing scalar quantizers (side-information dependent when possible) was established. More specifically, for the case of side information at both encoder and decoder, all possibilities for the dependence of the reconstruction on the side information were considered and the optimality of time-sharing scalar quantizers was established. For the case where side information is available at the encoder only and the reconstruction sequence is constrained to causal dependence on both the source and the side information, it was shown that, similarly as in non-causal source coding, the side information sequence can be disregarded without loss. Characterization of optimum performance when the reconstruction is allowed *non*-causal dependence on the side information, namely, the characterization of $R_{si,1,0}(D)$, remains open. For the case where side information is at the decoder only and the reproduction does not depend on the side information, a Slepian-Wolf coding argument was seen to imply that there is no penalty for the absence of side information at the encoder. Characterization of the remaining cases involving side information at the decoder remains open. These cases, however, seem less motivated from a practical viewpoint. It is not clear for these cases whether confining the induced reproduction to causal dependence on the source and (causal or non-causal) dependence on the side information has any operational interpretation.

Our findings are summarized in the following table:

	$s_1 = s_2 = 0$	$s_1 = 1, s_2 = 0$	$s_1 = 0, s_2 = 1$	$s_1 = s_2 = 1$
$x = 0$	$R_{0,0,0}(D) = \overline{Q}(P_X, D)$	$R_{0,1,0}(D) = \overline{Q}(P_X, D)$	$R_{0,0,1}(D) = \overline{Q}_{01}(P_{X,Y}, D)$	$R_{0,1,1}(D) = \overline{Q}_{01}(P_{X,Y}, D)$
$x = csi$	$R_{csi,0,0}(D) = \overline{Q}(P_X, D)$	$R_{csi,1,0}(D) = \overline{Q}(P_X, D)$	$R_{csi,0,1}(D) = ?$	$R_{csi,1,1}(D) = Q_{si}(P_{X,Y}, D)$
$x = si$	$R_{si,0,0}(D) = \overline{Q}(P_X, D)$	$R_{si,1,0}(D) = ?$	$R_{si,0,1}(D) = ?$	$R_{si,1,1}(D) = Q_{si}(P_{X,Y}, D)$

The first column is the result in [12] (the bottom two lines following trivially), the middle line in the second column is Theorem 4 (from which the first line in the second column follows trivially). The first line of the third column follows from Observation 1 in Section 5. The fourth column summarizes the results of Section 3: the first line is Theorem 3, while the second two are Theorem 2.

Appendix

A Proof of Theorem 3

Proof of Direct Part: By definition, there exist $\lambda \in [0,1]$ and $D_0, D_1 \geq 0$ such that $D = \lambda D_0 + (1-\lambda)D_1$ and

$$\overline{Q}_{01}(P_{X,Y}, D) = \lambda Q_{01}(P_{XY}, D_0) + (1-\lambda)Q_{01}(P_{XY}, D_1). \tag{A.1}$$

Fix an arbitrary $\varepsilon > 0$ and let $f^{(0)}, f^{(1)}$ denote ε-achievers of $Q_{01}(P_{XY}, D_0)$, $Q_{01}(P_{XY}, D_1)$, respectively, so that both

$$Ed(X_i, f^{(0)}(X_i)) \leq D_0, \quad Ed(X_i, f^{(1)}(X_i)) \leq D_1 \tag{A.2}$$

and

$$H(f^{(0)}(X_i)|Y_i) \leq Q_{01}(P_{XY}, D_0) + \varepsilon, \quad H(f^{(1)}(X_i)|Y_i) \leq Q_{01}(P_{XY}, D_1) + \varepsilon \tag{A.3}$$

hold. Let $\{b_i\}$ be a deterministic binary sequence satisfying

$$\frac{1}{n}\sum_{i=1}^{n} b_i \overset{n\to\infty}{\longrightarrow} 1-\lambda. \tag{A.4}$$

Letting the reproduction coder be given by $\hat{X}_i = f^{(b_i)}(X_i)$ (so that, in particular, it is a member of $\mathcal{F}_c$), (A.4) and (A.2) imply

$$Ed_k(X_1^k, \hat{X}_1^k) \overset{n\to\infty}{\longrightarrow} \lambda Ed(X_1, f^{(0)}(X_1)) + (1-\lambda)Ed(X_1, f^{(1)}(X_1)) \leq \lambda D_0 + (1-\lambda)D_1. \tag{A.5}$$

On the other hand

$$\begin{aligned}
\frac{1}{k}H(\hat{X}_1^k|\mathbf{Y}) &= \frac{1}{k}\sum_{i=1}^{k} H(\hat{X}_i|\hat{X}_1^{i-1}, \mathbf{Y}) \\
&= \frac{1}{k}\sum_{i=1}^{k} H(f^{(b_i)}(X_i)|Y_i) \\
&\overset{k\to\infty}{\longrightarrow} \lambda H(f^{(0)}(X_1)|Y_1) + (1-\lambda)H(f^{(1)}(X_1)|Y_1) && \text{(A.6)} \\
&\leq \lambda Q_{01}(P_{XY}, D_0) + (1-\lambda)Q_{01}(P_{XY}, D_1) + \varepsilon && \text{(A.7)} \\
&\leq \overline{Q}_{01}(P_{X,Y}, D) + \varepsilon, && \text{(A.8)}
\end{aligned}$$

where (A.6) follows by (A.4) and stationarity, (A.7) follows from (A.3), and (A.8) from (A.1). The proof is complete by the arbitrariness of $\varepsilon > 0$. □

Proof of Converse Part of Theorem 3: As indicated in Section 3, we shall prove a result somewhat stronger than the converse part of Theorem 3, allowing reproduction coders of the form $\hat{X}_k = f_k(X^k, Y^{k-1})$. More specifically, we shall show that if a reproduction coder of the form $\hat{X}_k = f_k(X^k, Y^{k-1})$ satisfies

$$d(\{f_k\}) \leq D \tag{A.9}$$

then

$$H(\hat{\mathbf{X}}|\mathbf{Y}) \geq \overline{Q}_{01}\left(P_{X,Y}, D\right). \tag{A.10}$$

For $k \geq 1$,

$$\begin{aligned}
H(\hat{X}_k|\hat{X}^{k-1}, \mathbf{Y}) &\geq H(\hat{X}_k|X^{k-1}, \mathbf{Y}) \\
&= H(f_k(X^k, Y^{k-1})|X^{k-1}, \mathbf{Y}) \\
&= \int H(f_k(x^{k-1}X_k, y^{k-1})|X^{k-1} = x^{k-1}, \mathbf{Y} = \mathbf{y})d\mu(x^{k-1}, \mathbf{y}) \\
&= \int H(f_k(x^{k-1}X_k, y^{k-1})|Y_k = y_k)d\mu(x^{k-1}, \mathbf{y}) \qquad \text{(A.11)} \\
&= \int \left[\int H(f_k(x^{k-1}X_k, y^{k-1})|Y_k = y_k)d\mu(y_k^\infty|x^{k-1}, y^{k-1})\right] d\mu(x^{k-1}, y^{k-1}) \\
&= \int \left[\int H(f_k(x^{k-1}X_k, y^{k-1})|Y_k = y_k)d\mu(y_k)\right] d\mu(x^{k-1}, y^{k-1}) \qquad \text{(A.12)} \\
&= \int \left[H(f_k(x^{k-1}X_k, y^{k-1})|Y_k)\right] d\mu(x^{k-1}, y^{k-1}) \\
&\geq \int \overline{Q}_{01}(P_{X,Y}, Ed(X_k, f_k(x^{k-1}X_k, y^{k-1})))d\mu(x^{k-1}, y^{k-1}) \qquad \text{(A.13)} \\
&\geq \overline{Q}_{01}\left(P_{X,Y}, \int Ed(X_k, f_k(x^{k-1}X_k, y^{k-1}))d\mu(x^{k-1}, y^{k-1})\right) \qquad \text{(A.14)} \\
&= \overline{Q}_{01}\left(P_{X,Y}, \int E[d(X_k, f_k(X^k, Y^{k-1}))|X^{k-1} = x^{k-1}, Y^{k-1} = y^{k-1}]d\mu(x^{k-1}, y^{k-1})\right) \\
&= \overline{Q}_{01}\left(P_{X,Y}, Ed(X_k, \hat{X}_k)\right), \qquad \text{(A.15)}
\end{aligned}$$

where (A.11) and (A.12) follow by the fact that the pairs (X_i, Y_i) are i.i.d., (A.13) follows from the definition of the function Q_{01}, and (A.14) follows by convexity. Inequality (A.10) is now established using (A.15) and the assumption (A.9) through a chain of inequalities analogous to that leading to (71). □

References

[1] T. M. Cover. A proof of the data compression theorem of Slepian and Wolf for ergodic sources. *IEEE Trans. Inform. Theory*, IT-21:226–228, March 1975.

[2] R. K. Gilbert and D. L. Neuhoff. Bounds to the performance of causal codes for Markov sources. *Proc. Allerton Conf. Comm. Contr. and Comput.*, pages 284–292, 1979.

[3] R. M. Gray and D. L. Neuhoff. Quantization. *IEEE Trans. Inform. Theory*, 44(6):2325–2383, October 1998.

[4] A. György and T. Linder. Optimal entropy-constrained scalar quantization of a uniform source. *IEEE Trans. Inform. Theory*, 46(7):2704–2711, November 2000.

[5] A. György and T. Linder. On the structure of optimal entropy-constrained scalar quantizers. *IEEE Trans. Inform. Theory*, 48(2):416–427, February 2002.

[6] A. György, T. Linder, and G. Lugosi. Efficient adaptive algorithms and minimax bounds for zero-delay lossy source coding. *IEEE Trans. on Signal Processing*, 52:2337–2347, August 2004.

[7] A. György, T. Linder, and G. Lugosi. A "follow the perturbed leader"-type algorithm for zero-delay quantization of individual sequences. *Proc. 2004 Data Compression Conference (DCC'04)*, March 2004. Snowbird, Utah, USA.

[8] T. Linder and G. Lugosi. A zero-delay sequential scheme for lossy coding of individual sequences. *IEEE Trans. Inform. Theory*, 47:2533–2538, September 2001.

[9] T. Linder and R. Zamir. Causal source coding of stationary sources and individual sequences with high resolution. *IEEE Trans. Inform. Theory.* Accepted for publication.

[10] S. Matloub and T. Weissman. On competitive zero-delay joint source-channel coding. *Proceedings of 38th Annual Conference on Information Sciences and Systems*, pages 555–559, March 2004.

[11] N. Merhav and I. Kontoyiannis. Source coding exponents for zero-delay coding with finite memory. *IEEE Trans. Inform. Theory*, IT-49:609–625, March 2003.

[12] D. L. Neuhoff and R. K. Gilbert. Causal source codes. *IEEE Trans. Inform. Theory*, IT-28(5):701–713, September 1982.

[13] E. Sabbag. Large deviations performance of zero–delay finite–memory lossy source codes and source–channel codes. Master's thesis, Technion-I.I.T., December 2003.

[14] D. Slepian and J. K. Wolf. Noiseless coding of correlated information sources. *IEEE Trans. Inform. Theory*, IT-19(4):471–480, July 1973.

[15] T. Weissman and N. Merhav. Finite-delay lossy coding and filtering of individual sequences corrupted by noise. *IEEE Trans. Inform. Theory*, 48(3):721–733, March 2002.