Article

Source Coding with a Causal Helper

Shraga I. Bross

Faculty of Engineering, Bar-Ilan University, Ramat-Gan 52900, Israel; brosss@biu.ac.il

Abstract: A multi-terminal network, in which an encoder is assisted by a side-information-aided helper, describes a memoryless identically distributed source to a receiver, is considered. The encoder provides a non-causal one-shot description of the source to both the helper and the receiver. The helper, which has access to causal side-information, describes the source to the receiver sequentially by sending a sequence of causal descriptions depending on the message conveyed by the encoder and the side-information subsequence it has observed so far. The receiver reconstructs the source causally by producing on each time unit an estimate of the current source symbol based on what it has received so far. Given a reconstruction fidelity measure and a maximal allowed distortion, we derive the rates-distortion region for this setting and express it in terms of an auxiliary random variable. When the source and side-information are drawn from an independent identically distributed Gaussian law and the fidelity measure is the squared-error distortion we show that for the evaluation of the rates-distortion region it suffices to choose the auxiliary random variable to be jointly Gaussian with the source and side-information pair.

Keywords: source coding; causal helper; causal side-information

1. Introduction

In the classical source coding with decoder side information problem, the source and side information are generated by independent drawings (X_k, Y_k) of the pair $(X, Y) \sim P_{XY}$. The encoder forms a description of the source sequence $X^n = (X_1, \ldots, X_n)$ using a map $f^{(n)} \colon \mathcal{X}^n \to \{1, \ldots, \lfloor 2^{nR} \rfloor\}$, while the decoder forms its reconstruction $\hat{X}^n$ depending on both the side-information sequence Y^n and the index $T \in \{1, \ldots, \lfloor 2^{nR} \rfloor\}$ conveyed by the encoder. In their seminal work [1], Wyner and Ziv derived the rate distortion function for this setting, given a fidelity measure, when $\hat{X}^n$ can depend on Y^n in an arbitrary manner. Yet, Wyner–Ziv source coding with non-causal decoder side-information involves binning the implementation of which is complex.

A successive refinement for the Wyner–Ziv problem with side-information (Y, Z) is a variant of the Wyner–Ziv model, in which the encoder provides a two-layer description of the source sequence X^n to a pair of decoders. Decoder 1, which obtains just the course description layer, has available as non-causal side-information the memoryless vector Z^n, while Decoder 2, which obtains both description layers, has available as non-causal side-information the memoryless vector Y^n. It is further assumed that the reconstruction formed by Decoder 2 should be of smaller distortion compared to that formed by Decoder 1. Such a model has been considered in [2], wherein a complete characterization of the rates-distortion region has been obtained for the case where Z is stochastically degraded to Y with respect to X—i.e., $X \leftrightarrow Y \leftrightarrow Z$ forms a Markov chain.

The work [3] studies an extension of the model of [2] where a conference link of given capacity allows the unidirectional cooperation between Decoder 1 and Decoder 2—i.e., Decoder 1 functions also as a helper. The results in [3] are partially tight in the sense that the characterization of the encoder rates is conclusive, the remaining gap being in the characterization of the helper's rate.

Thus, with non-causal side-information at both decoders, the successive refinement for the Wyner–Ziv problem with a helper is yet unsolved.

Motivated by practical delay-constrained sequential source coding with decoder side information Weissman and El Gamal considered in [4], a scheme with causal side information at the decoder, where the sequence of reconstructions $\hat{X}^n = (\hat{X}_1, \ldots, \hat{X}_n)$ is formed sequentially in a causal manner according to $\hat{X}_k = \hat{X}_k(T, Y^k)$, and derived the corresponding rate distortion function. Similar to [1], the rate distortion function in [4] is expressed in terms of an auxiliary random variable, thus leaving the optimal choice of which an open issue depending on the specifics of the model. For the Gaussian setting where (X, Y) are Gaussian, the authors in [4] compute an upper bound on the rate distortion function by choosing the auxiliary random variable to be jointly Gaussian with X, while leaving the question of whether this choice is optimal yet an open problem.

With the vision that modern network design will support the use of cooperation links in favor of reduction of encoding/decoding complexity and network deployment constraints, this work considers an extension of the model [4], involving a causal helper and causal side information at the decoder, which is described as follows. The components of a trivariate independent identically distributed (IID) finite-alphabet source $\{X_k, Y_k, Z_k\}_{k=1}^{\infty}$ are observed by three terminals. The source component $\{X_k\}_{k=1}^{\infty}$ is observed by the encoder, while the source component $\{Y_k\}_{k=1}^{\infty}$ is observed by the helper. Both the encoder and the helper describe the length-n source sequence X^n according to a given fidelity to the decoder in two steps. First, the encoder when given X^n sends a rate-R description of it to both the helper and the decoder. Then, the helper sends to the decoder, per each source symbol Y_k, a causal description depending on the message it had received from the encoder and the source subsequence Y^k it had observed so far, with the aggregate rate not exceeding R_{h}. The decoder, which observes the source component $\{Z_k\}_{k=1}^{\infty}$ in a causal manner, per channel use k, uses the descriptions it had received so far and the source subsequence Z^k to form its reconstruction $\hat{X}_k$ for the source symbol X_k. Given a fidelity measure and a maximal allowed distortion, the goal is to determine the set of all rate pairs (R, R_{h}) that satisfy the distortion constraint.

Causal decoder side information has been considered as well in context with successive refinement. With the aim of reducing encoder/decoder complexity, a two layer description model with successive refinement has been considered in [5], under the setting that the side information is available causally at each of the decoders. A single-letter characterization of the rates-distortion region is obtained in [5], irrespective of the relative ordering of the side information quality at the decoders. Furthermore, the direct part in [5] demonstrates that similarly to [4] with causal side-information at the decoders the optimal code avoids binning, hence its implementation is practically appealing. The extension of the model [5] with a causal helper has recently been studied in [6].

2. Problem Formulation

Formally, our problem can be stated as follows. A discrete memoryless source (DMS) $(\mathcal{X}, P_X)$ is an infinite sequence $\{X_i\}_{i=1}^{\infty}$ of independent copies of a random variable X taking values in the finite set $\mathcal{X}$ with generic law P_X. Similarly, a triple source $(\mathcal{X}\mathcal{Y}\mathcal{Z}, P_{XYZ})$ is an infinite sequence of independent copies of the triplet of random variables (X, Y, Z), taking values in the finite sets $\mathcal{X}, \mathcal{Y}$ and $\mathcal{Z}$, respectively, with generic joint law P_{XYZ}. Since our goal is to encode the source X, let $\hat{\mathcal{X}}$ denote any finite reconstruction alphabet and let $d : \mathcal{X} \times \hat{\mathcal{X}} \to [0, \infty)$ be a single-letter distortion measure. The vector distortion measure is defined as

$$d(\boldsymbol{x}, \hat{\boldsymbol{x}}) = \frac{1}{n} \sum_{i=1}^{n} d(x_i, \hat{x}_i), \qquad \forall\, \boldsymbol{x} \in \mathcal{X}^n, \, \hat{\boldsymbol{x}} \in \hat{\mathcal{X}}^n.$$

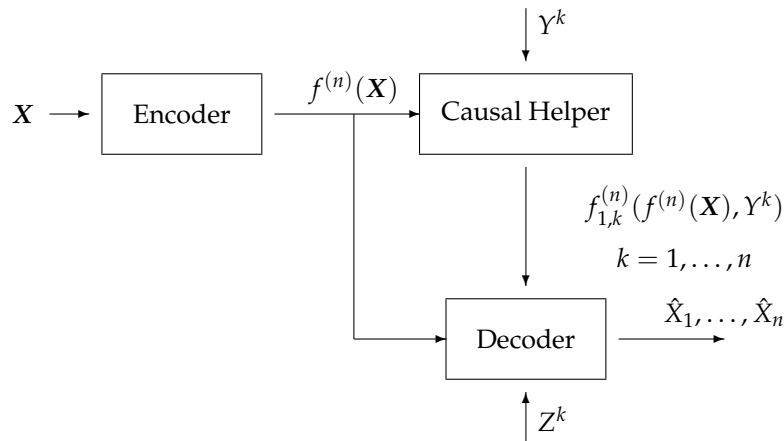A system diagram appears in Figure 1.

Figure 1. The causal listening-helper model.

Definition 1. *An* $\left(n, M^{(n)}, \prod_{k=1}^{n} L_k^{(n)}, D\right)$ *code for the source X with causal side-information (SI) (Y, Z) and causal helper, consists of:*

1. *An encoder mapping*

$$f^{(n)} \colon \mathcal{X}^n \to \{1, \ldots, M^{(n)}\}. \tag{1}$$

2. *A unidirectional conference between the helper and the decoder consisting of a sequence of causal descriptions*

$$f_{1,k}^{(n)} \colon \{1, \ldots, M^{(n)}\} \times \mathcal{Y}^k \to \{1, \ldots, L_k^{(n)}\}, \qquad k = 1, \ldots, n \tag{2}$$

    and

3. *A sequence $g_1^{(n)}, \ldots, g_n^{(n)}$ of decoder reconstructions*

$$g_k^{(n)} \colon \{1, \ldots, M^{(n)}\} \times \{1, \ldots, L_1^{(n)}\} \times \ldots \times \{1, \ldots, L_k^{(n)}\} \times \mathcal{Z}^k \to \hat{\mathcal{X}}, \qquad k = 1, \ldots, n \tag{3}$$

    such that

$$\mathbb{E}\, d\left[X^n, \left(g_1^{(n)}\left(f^{(n)}(X^n), f_{1,1}^{(n)}(f^{(n)}(X^n), Y_1), Z_1\right), \ldots, \right.\right.$$
$$g_k^{(n)}\left(f^{(n)}(X^n), \{f_{1,j}^{(n)}(f^{(n)}(X^n), Y^j)\}_{j=1}^{k}, Z^k\right), \ldots,$$
$$\left.\left. g_n^{(n)}\left(f^{(n)}(X^n), \{f_{1,j}^{(n)}(f^{(n)}(X^n), Y^j)\}_{j=1}^{n}, Z^n\right)\right)\right] \le D. \tag{4}$$

The rate tuple (R, R_{h}) of the code is

$$R = \frac{1}{n} \log M^{(n)}$$
$$R_{\mathrm{h}} = \frac{1}{n} \sum_{k=1}^{n} \log L_k^{(n)}. \tag{5}$$

Given a non-negative distortion D, the tuple (R, R_{h}) is said to be D-achievable for X with causal SI (Y, Z) if, for any $\delta > 0$, $\epsilon > 0$, and sufficiently large n, there exists an $(n, \exp[n(R + \delta)], \exp[n(R_{\mathrm{h}} + \delta)], D + \epsilon)$ code for the source X with causal SI (Y, Z) and causal helper. The collection of all D-achievable rate tuples is the achievable source-coding region and is denoted by $\mathcal{R}(D)$.

In this work, we provide a single-letter characterization for $\mathcal{R}(D)$. In contrast to [5], a consequence of Definition 1 is that $\mathcal{R}(D)$ may depend on the joint law of the triple P_{XYZ}, not only through the

marginal laws P_{XY} and P_{XZ}. This is due to the fact that the decoder acquires, in addition to its private side information Z, some additional side information on Y via the conference link. As a result, the expectation in (4), which takes into account the mapping $g_k^{(n)}(\cdot)$, is taken over the joint law P_{XYZ} and not just over the marginal law P_{XZ}, as is the case in [5].

Finally, although not of the finite alphabet, of particular interest to us is the Gaussian source. This is a memoryless source, where (X, Y, Z) are centered jointly Gaussians with each pair (X_i, Y_i) drawn such that P_{XY} satisfies

$$Y_i = \sqrt{\rho}X_i + W_i \tag{6}$$

where $\rho > 0$ is a fixed constant, $X_i \sim \mathcal{N}(0,1)$ and $W_i \sim \mathcal{N}(0,1)$ are mutually independent. Moreover, Z_i is drawn according to

$$Z_i = aX_i + bY_i + T_i \tag{7}$$

where a and b are real numbers and $T_i \sim \mathcal{N}(0,1)$ is independent of (X_i, Y_i). Furthermore, in this case, the fidelity measure will be

$$\mathbb{E}\,d[X^n, \hat{X}^n] \triangleq \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\big[(X_i - \hat{X}_i)^2\big], \tag{8}$$

in which case we may restrict the reproduction functions $g_k^{(n)}$ to be the MMSE estimates of X_k, $k = 1, \ldots, n$ given $f^{(n)}(X^n)$, the sequence of causal descriptions $\left\{f_{1,j}^{(n)}(f^{(n)}(X^n), Y^j)\right\}_{j=1}^{k}$, and the side-information Z^k. That is,

$$\hat{X}_k = \mathbb{E}\left[X_k \,\middle|\, f^{(n)}(X^n), \left\{f_{1,j}^{(n)}(f^{(n)}(X^n), Y^j)\right\}_{j=1}^{k}, Z^k\right].$$

In the Gaussian network setting, our focus will be on determining the optimal choice of the auxiliary random variable by means of which the rates-distortion region is defined. Specifically, we will show that choosing it to be jointly Gaussian with X is optimal.

3. Main Results

Given a maximal allowed distortion D, define $\mathcal{R}^*(D)$ to be the set of all rate pairs (R, R_h) for which there exist random variables (U, V), taking values in finite alphabets $\mathcal{U}, \mathcal{V}$, respectively, such that

1. $U \leftrightarrow X \leftrightarrow (Y, Z)$ forms a Markov chain.
2. Conditioned on U, $X \leftrightarrow Y \leftrightarrow V$ forms a Markov chain.
3. There exist deterministic maps

$$\begin{aligned}\tilde{f}_1 &: \quad \mathcal{U} \times \mathcal{Y} \to \mathcal{V} \\ g &: \quad \mathcal{U} \times \mathcal{V} \times \mathcal{Z} \to \hat{\mathcal{X}}\end{aligned} \tag{9}$$

such that, with $V \triangleq \tilde{f}_1(U, Y)$,

$$\mathbb{E}d(X, g(U, V, Z)) \leq D. \tag{10}$$

4. The alphabets $\mathcal{U}, \mathcal{V}$ satisfy

$$\begin{aligned}|\mathcal{U}| &\leq |\mathcal{X}| + 2 \\ |\mathcal{V}| &\leq |\mathcal{X}|(|\mathcal{X}| + 2) + 1.\end{aligned} \tag{11}$$

5.   The rates R and R_h satisfy

$$R \geq I(X;U) \tag{12a}$$
$$R_h \geq H(V|U) = I(Y;V|U). \tag{12b}$$

Our first result is a single-letter characterization of the rates-distortion region.

Theorem 1. $\mathcal{R}(D) = \mathcal{R}^*(D)$.

Proof. See Section 4.1.   □

Remark 1. *The converse holds as well for the setting where the causal helper and the reconstructor benefit from causal disclosure—i.e., are cognizant of the past realizations of the source sequence, hence they are allowed to depend also on X^{k-1}, when forming $f_{1,k}^{(n)}$ and $\hat{X}_k$, respectively. That is,*

$$
\begin{aligned}
f_{1,k}^{(n)} &= f_{1,k}^{(n)}(f^{(n)}(X^n), Y^k, X^{k-1}), \\
\hat{X}_k &= g_k^{(n)}\left(f^{(n)}(X^n), \left\{f_{1,j}^{(n)}(f^{(n)}(X^n), Y^i, X^{i-1})\right\}_{j=1}^k, Z^k, X^{k-1}\right).
\end{aligned}
\tag{13}
$$

3.1. The Gaussian Setting with $Z = \emptyset$.

To simplify the presentation, we consider first the Gaussian setting with $Z = \emptyset$. In this case, the region $\mathcal{R}(D)$ is defined as the set of rate pairs (R, R_h) for which there exist random variables (U, V), taking real values, such that

1. $U \leftrightarrow X \leftrightarrow Y$ forms a Markov chain.
2.   Conditioned on U, $X \leftrightarrow Y \leftrightarrow V$ forms a Markov chain.
3.   There exist deterministic maps

$$
\begin{aligned}
\tilde{f}_1 &: \quad \mathcal{U} \times \mathcal{Y} \to \mathcal{V} \\
g &: \quad \mathcal{U} \times \mathcal{V} \to \hat{\mathcal{X}}
\end{aligned}
\tag{14}
$$

such that, with $V \triangleq \tilde{f}_1(U, Y)$,

$$D \geq \mathbb{E}\left[(X - g(U, V))^2\right] \triangleq \sigma_{X|UV}^2. \tag{15}$$

4.   The rates R and R_h satisfy (12a) and (12b), respectively.

Our second result characterizes the optimal choice of P_{XU} in the Gaussian setting.

Theorem 2. *For the evaluation of $\mathcal{R}(D)$ when (X, Y) are Gaussian, it suffices to assume that (X, U) are jointly Gaussian.*

Proof. For the treatment to follow, let us define the Gaussian channel $Y = \sqrt{\varrho}X + W$ where X has an arbitrary law with $\mathbb{E}[X^2] \leq 1$, $W \sim \mathcal{N}(0,1)$ is independent of X, and $\varrho > 0$ is the channel signal-to-noise ratio. We shall denote by $\mathbb{G}_{Y|X}^{(\varrho)}$ the conditional law of Y given X for this additive Gaussian model. Furthermore, the notation $(U_n, X_n, Y_n) \sim P_{U_n X_n} \mathbb{G}_{Y_n|X_n}^{(\varrho)}$ would imply that $(U_n, X_n) \sim P_{U_n X_n}$ and $Y_n = \sqrt{\varrho}X_n + W_n$ with $W_n \sim \mathcal{N}(0,1)$ independent of (U_n, X_n). Using this notation, the law P_{UXYV} defining $\mathcal{R}(D)$ (see also (45) ahead) may equivalently be expressed as

$$P_{XUYV} = P_{XU}\mathbb{G}_{Y|X}^{(\varrho)}P_{V|YU}. \tag{16}$$

Henceforth, we denote a law which factors as in (16) by $X \leftrightarrow Y \leftrightarrow V|_U$—i.e., conditioned on U, $X \leftrightarrow Y \leftrightarrow V$ forms a Markov chain and, furthermore, $\mathbb{P}_{Y|X} = \mathbb{G}_{Y|X}^{(\varrho)}$ independently of the rest.

Define the region (12a) and (12b) subject to constraint (15) by $\mathcal{O}_K$, where the subscript K denotes the covariance constraint (15). The region $\mathcal{O}_K$ is a closed convex set.

In line with [7,8], we define a λ-parametrized family of functions which are related to the sum rate associated with $\mathcal{R}(D)$.

Fix some $\lambda > 1$, and consider the minimization of the λ-sum rate defined as

$$\min_{(R,R_\mathrm{h}) \in \mathcal{O}_K} R + \lambda R_\mathrm{h}. \tag{17}$$

Observe that

$$\min_{(R,R_\mathrm{h}) \in \mathcal{O}_K} R + \lambda R_\mathrm{h} \overset{(a)}{\geq} \inf_{\substack{P_{XUYV} : X \leftrightarrow Y \leftrightarrow V|_U \\ \sigma^2_{X|UV} \leq D}} I(X;U) + \lambda I(Y;V|U), \tag{18}$$

where (a) follows using the lower bounds (12a) and (12b). Since the marginal law of X in our model (6) is Gaussian, the differential entropy $h(X)$ is fixed, thus for the minimization of (18) over a law of the form (16), we define the following functional of P_{XU} (i.e., of the conditional law $P_{X|U}$)

$$\mathsf{s}_\lambda(X, \varrho|U) \triangleq -h(X|U) + \inf_{\substack{V : X \leftrightarrow Y \leftrightarrow V|_U \\ \sigma^2_{X|UV} \leq D}} \lambda I(Y;V|U). \tag{19}$$

Thus, the minimum in (17) may be expressed as

$$\mathsf{V}_\lambda(\varrho) \triangleq \inf_{P_{XU} : E[X^2] \leq 1} \mathsf{s}_\lambda(X, \varrho|U). \tag{20}$$

Henceforth, the set of laws P_{UXYV} defined by (16) which satisfy $\sigma^2_{X|UV} \leq D$ will be denoted by $\mathscr{Q}$ and will be attributed as the feasible set.

As shown below, with a proper choice of λ, the functional (19) exhibits a "pair grouping" property with respect to the input X in the sense that the value of s_λ does not increase under this operation. Having established that, we follow the same steps as in the proof of ([9] Theorem 9) to establish that the objective (20) is attained when $P_{X|U}$ is Gaussian. More specifically,

- Lemma 1 shows that the value of s_λ "improves" under the pair grouping operation.
- With the proper time-sharing of two distributions attaining the infimum in (20) and satisfying the extremal property defined in (23) ahead, Lemma 2 proves that the pair grouping operation exhibits Gaussianity in the sense of Bernstein's characterization.

Lemma 1. *Let P_{XU} be a law on $\mathcal{X} \times \mathcal{U}$, let $P_{UXY} = P_{XU}\mathbb{G}_{Y|X}^{(\varrho)}$, and let (U_1, X_1, Y_1) and (U_2, X_2, Y_2) be two independent copies of (U, X, Y). Define*

$$X_+ \triangleq \frac{X_1 + X_2}{\sqrt{2}} \,, \quad X_- \triangleq \frac{X_1 - X_2}{\sqrt{2}} \tag{21}$$

and similarly

$$Y_+ \triangleq \frac{Y_1 + Y_2}{\sqrt{2}} \,, \quad Y_- \triangleq \frac{Y_1 - Y_2}{\sqrt{2}}.$$

with $\mathbf{U} = (U_1, U_2)$, there exists some $\lambda^* > 1$ such that for any $\lambda \geq \lambda^*$, the following inequalities hold

$$2\mathsf{s}_\lambda(X, \varrho|U) \;\geq\; \mathsf{s}_\lambda(X_+, \varrho|X_-, \mathbf{U}) + \mathsf{s}_\lambda(X_-, \varrho|Y_+, \mathbf{U}) \tag{22a}$$

$$2s_\lambda(X, \varrho|U) \quad \geq \quad s_\lambda(X_+, \varrho|Y_-, \mathbf{U}) + s_\lambda(X_-, \varrho|X_+, \mathbf{U}). \tag{22b}$$

Proof. See Section 4.2. □

Assume that the infimum in (20) is attained, and let $\mathcal{P}$ denote the subset of laws P_{XU} achieving the minimum. Suppose further that there exists a law $P_{XU} \in \mathcal{P}$ such that, for $(Y, X, U) \sim P_{XU}\mathbb{G}_{Y|X}^{(\varrho)}$, for any other $P_{X'U'} \in \mathcal{P}$ where $(Y', X', U') \sim P_{X'U'}\mathbb{G}_{Y'|X'}^{(\varrho)}$

$$h(Y|U) - h(X|U) \leq h(Y'|U') - h(X'|U'). \tag{23}$$

Denote the value of the LHS of (23) by $g^*(\varrho)$.

Lemma 2. *Fix $\epsilon > 0$. Let P_{XU} be an admissible law such that*

$$s_\lambda(X, \varrho|U) \quad \leq \quad \mathsf{V}_\lambda(\varrho) + \epsilon \tag{24a}$$
$$h(Y|U) - h(X|U) \quad \leq \quad g^*(\varrho) + \epsilon. \tag{24b}$$

There exists a law $(Y', X', U') \sim P_{X'U'}\mathbb{G}_{Y'|X'}^{(\varrho)}$ satisfying

$$s_\lambda(X', \varrho|U') \leq \mathsf{V}_\lambda(\varrho) + 2\epsilon \tag{25}$$

and

$$h(Y'|U') - h(X'|U') + \frac{1}{2}I(X_1 + X_2; X_1 - X_2|Y_1, Y_2, U_1, U_2) \leq g^*(\varrho) + \epsilon, \tag{26}$$

where (U_1, X_1, Y_1) and (U_2, X_2, Y_2) denote independent copies of (U, X, Y).

Proof. See Section 4.3. □

Inequality (26) combined with assumption (23) and Lemma 6 in ([9] Section VI.B) establish the following.

Lemma 3. *There exists a sequence $\{X_n, U_n\}$, such that for each $n \geq 1$, (X_n, U_n) is feasible,*

$$\lim_{n \to \infty} s_\lambda(X_n, \varrho|U_n) = \mathsf{V}_\lambda(\varrho) \tag{27}$$

and there exists a feasible law $P_{X_*U_*}$ on $\mathbb{R} \times \mathcal{U}$ such that

$$(X_n, U_n) \xrightarrow{\mathcal{D}} (X_*, U_*) \sim P_{X_*U_*}$$

and with $\mathbf{U}_n \triangleq (U_{1,n}, U_{2,n})$, $\mathbf{Y}_n \triangleq (Y_{1,n}, Y_{2,n})$, where the tuple $(U_{1,n}, X_{1,n}, Y_{1,n})$ and the tuple $(U_{2,n}, X_{2,n}, Y_{2,n})$ are two independent copies of $(U_n, X_n, Y_n) \sim P_{U_n X_n}\mathbb{G}_{Y_n|X_n}^{(\varrho)}$, we have

$$\liminf_{n \to \infty} I(X_{1,n} + X_{2,n}; X_{1,n} - X_{2,n}|\mathbf{Y}_n, \mathbf{U}_n = \mathbf{u}) = 0 \quad \text{for } P_{U_*} \times P_{U_*} - \text{a.e } \mathbf{u}. \tag{28}$$

Finally, we apply to identity (28) the extended form of Bernstein's theorem (see [10,11]), which is proved in ([9] Appendix A), to conclude that $X_{1,*}|_{\{U_{1,*}=u_1\}}$ (i.e., the conditional random variable: $X_{1,*}$ conditioned on $U_{1,*} = u_1$) and $X_{2,*}|_{\{U_{2,*}=u_2\}}$ are independent Gaussian random variables of equal variance.

Since the marginal of X is Gaussian, with no loss in generality, we may choose (X, U) to be jointly Gaussian, which establishes our claim in Theorem 2. □

3.2. The Gaussian Setting with Decoder Side-Information Z

In the Gaussian setting where the decoder side-information Z in non-void, using our previous definition $Y = \sqrt{\rho}X + W$, let

$$Z = aX + bY + T \tag{29}$$

for a pair of real numbers (a, b) where $T \sim \mathcal{N}(0, 1)$ is independent of (X, Y) and U. For arbitrary X with $\mathbb{E}[X^2] \leq 1$, let us define by $\tilde{\mathbb{G}}_{YZ|X}^{(\varrho,a,b)}$ the conditional law obtained by forming the pair (Y, Z), when given X, via the additive independent Gaussian pair (W, T) as described above.

The rates-distortion region $\mathcal{R}(D)$ is defined in Theorem 1 with

$$\mathbb{E}d(X, g(U, V, Z)) = \mathbb{E}\left[(X - g(U, V, Z))^2\right] \triangleq \sigma_{X|UVZ}^2 \leq D \tag{30}$$

and it is evaluated over a law of the form

$$P_{XUYZV} = P_{XU}\tilde{\mathbb{G}}_{YZ|X}^{(\varrho,a,b)}P_{V|YU}. \tag{31}$$

Consequently, for the minimization of (20), we consider the functional

$$\mathsf{s}_\lambda(X, \varrho|U) \triangleq -h(X|U) + \inf_{\substack{V: X \leftrightarrow Y \leftrightarrow V|_U \\ \sigma_{X|UVZ}^2 \leq D}} \lambda I(Y; V|U). \tag{32}$$

under a law of the form (31). Next, in Lemma 1, for P_{XU}, a law on $\mathcal{X} \times \mathcal{U}$ and $P_{UXYZ} = P_{XU}\tilde{\mathbb{G}}_{YZ|X}^{(\varrho,a,b)}$ we let (U_1, X_1, Y_1, Z_1) and (U_2, X_2, Y_2, Z_2) be two independent copies of (U, X, Y, Z).

Upon defining (X_-, X_+) and (Y_-, Y_+) as before, we also define

$$Z_+ \triangleq \frac{Z_1 + Z_2}{\sqrt{2}} , \quad Z_- \triangleq \frac{Z_1 - Z_2}{\sqrt{2}}.$$

It can be verified that

$$Z_+ = aX_+ + bY_+ + \frac{1}{\sqrt{2}}(T_1 + T_2) \triangleq aX_+ + bY_+ + T_+$$

$$Z_- = aX_- + bY_- + \frac{1}{\sqrt{2}}(T_1 - T_2) \triangleq aX_- + bY_- + T_- \tag{33}$$

where T_- and T_+ are independent. Furthermore, the pair (T_-, T_+) is independent of (U, X_-, X_+, Y_-, Y_+) and is equal in distribution to the pair (T_1, T_2). Thus, the simultaneous unitary transformation $(Y_1, Y_2) \mapsto (Y_-, Y_+)$ and $(Z_1, Z_2) \mapsto (Z_-, Z_+)$ preserves the Gaussian nature of the channel and factors according to

$$P_{UX_-X_+Y_-Y_+Z_-Z_+V} = P_{X_-X_+U}\tilde{\mathbb{G}}_{Y_-Z_-|X_-}^{(\varrho,a,b)}\tilde{\mathbb{G}}_{Y_+Z_+|X_+}^{(\varrho,a,b)}P_{V|UY_-Y_+}. \tag{34}$$

Observe that with the choice of $U \triangleq \mathbf{U} = (U_1, U_2)$ and $V = (V_1, V_2)$ where $(U_1, X_1, Y_1, Z_1, V_1)$ and $(U_2, X_2, Y_2, Z_2, V_2)$ are two independent copies of $(\tilde{U}, X, Y, Z, \tilde{V})$ such that $X \leftrightarrow Y \leftrightarrow \tilde{V}|_{\tilde{U}}$ and $\sigma_{X|\tilde{U}\tilde{V}Z}^2 \leq D$ –i.e., $P_{\tilde{U}XYZ\tilde{V}}$ is feasible for the minimization of (20), we have

$$\begin{aligned}
\sigma_{X_+|UY_-Z_+V}^2 &\leq \sigma_{X_+|UVZ_+}^2 \leq D \\
\sigma_{X_-|UY_+Z_-V}^2 &\leq \sigma_{X_-|UVZ_-}^2 \leq D.
\end{aligned} \tag{35}$$

Thus, Lemma 1 holds for this setting with decoder side-information Z as well.

4. Proofs

4.1. Proof of Theorem 1

Converse: Assume that the pair (R, R_{h}) is D-achievable. For $j = 1, \ldots, n$ with the convention $Z^0 \triangleq \varnothing$, define the RV's

$$
\begin{align}
T &\triangleq f^{(n)}(X^n) \tag{36a}\\
V_j &\triangleq f^{(n)}_{1,j}(f^{(n)}(X^n), Y^j), \quad j = 1, \ldots, n \tag{36b}\\
U_j &\triangleq (T, X^{j-1}, Y^{j-1}, Z^{j-1}), \quad j = 1, \ldots, n. \tag{36c}
\end{align}
$$

The rate R is lower bounded as follows

$$
\begin{align}
nR &\geq \log M^{(n)} \geq H(T) \notag\\
&= I(X^n; T) = \sum_{k=1}^{n} I(X_k; T | X^{k-1}) \notag\\
&\overset{(a)}{=} \sum_{k=1}^{n} I(X_k; T X^{k-1}) \notag\\
&\overset{(b)}{=} \sum_{k=1}^{n} I(X_k; T X^{k-1} Y^{k-1} Z^{k-1}) \notag\\
&= \sum_{k=1}^{n} I(X_k; U_k). \tag{37}
\end{align}
$$

Here, (a) follows since X^n is memoryless, and (b) follows since $X_k \leftrightarrow (T, X^{k-1}) \leftrightarrow (Y^{k-1}, Z^{k-1})$ forms a Markov chain.

We may now lower bound R_{h} as follows

$$
\begin{align}
nR_{\mathrm{h}} &\geq \sum_{k=1}^{n} \log L_k^{(n)} \geq H(V_1, V_2, \ldots, V_n) \notag\\
&\geq H(V_1, V_2, \ldots, V_n | T) \notag\\
&= \sum_{k=1}^{n} H(V_k | T V^{k-1}) \notag\\
&\geq \sum_{k=1}^{n} H(V_k | T V^{k-1} X^{k-1} Y^{k-1} Z^{k-1}) \notag\\
&\overset{(c)}{=} \sum_{k=1}^{n} H(V_k | T X^{k-1} Y^{k-1} Z^{k-1}) \notag\\
&\overset{(c)}{=} \sum_{k=1}^{n} I(Y_k; V_k | T X^{k-1} Y^{k-1} Z^{k-1}) \notag\\
&= \sum_{k=1}^{n} I(Y_k; V_k | U_k), \tag{38}
\end{align}
$$

where (c) follows since V_j is a deterministic function of (T, Y^j).

The sum-rate can be lower bounded as follows

$$
\begin{align}
n(R + R_{\mathrm{h}}) &\geq \log M^{(n)} + \sum_{k=1}^{n} \log L_k^{(n)} \notag\\
&\geq H(T) + H(V_1, \ldots, V_n) \notag\\
&\geq H(T) + H(V_1, \ldots, V_n | T) \notag
\end{align}
$$

$$
\begin{aligned}
&= \; I(T; X^n) + H(V_1, \ldots, V_n | T) \\
&= \; \sum_{k=1}^{n} I(X_k; T | X^{k-1}) + H(V_1, \ldots, V_n | T) \\
&\overset{(d)}{\geq} \; \sum_{k=1}^{n} \left[I(X_k; U_k) + I(Y_k; V_k | U_k) \right],
\end{aligned}
\tag{39}
$$

where (d) follows by equality (37) and inequality (38).

Draw J uniformly from $\{1, \ldots, n\}$ independently of $\{(X_k, Y_k, Z_k, V_k, U_k)\}_{k=1}^{n}$, and define the RV's $U = (U_J, J)$, $V = V_J$, $Z = Z_J$, $Y = Y_J$, and $X = X_J$. Using J, we may express (37) as follows

$$
\begin{aligned}
R &\geq \; \frac{1}{n} \sum_{k=1}^{n} I(X_k; U_k) = I(X_J; U_J | J) \\
&= \; I(X_J; U_J, J) - I(X_J; J) \\
&= \; I(X_J; U_J, J) = I(X; U),
\end{aligned}
\tag{40}
$$

and we may express (38) as follows

$$
\begin{aligned}
R_{\mathrm{h}} &\geq \; \frac{1}{n} \sum_{k=1}^{n} I(Y_k; V_k | U_k) \\
&= \; I(Y_J; V_J | U_J, J) = I(Y; V | U) = H(V | U).
\end{aligned}
\tag{41}
$$

With regard to the expected distortion, we may write

$$
\begin{aligned}
D &\geq \; \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[d(X_i, \hat{X}_i)] \\
&= \; \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[d(X_i, g_i^{(n)}(T, V_1, \ldots, V_i, Z^i))] \\
&\overset{(e)}{\geq} \; \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[d(X_i, g_i^*(T, X^{i-1}, Y^{i-1}, Z^{i-1}, V_i, Z_i))] \\
&= \; \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[d(X_i, g_i^*(U_i, V_i, Z_i))] \\
&= \; \mathbb{E}[d(X_J, g(U_J, J, V_J, Z_J))] \\
&= \; \mathbb{E}[d(X, g(U, V, Z))].
\end{aligned}
\tag{42}
$$

Step (e) is justified as follows: Since $V_1, \ldots, V_{i-1}$ are deterministic functions of (T, Y^{i-1})

$$
(X^{i-1}, Z^{i-1}, Z_i) \leftrightarrow (T, Y^{i-1}) \leftrightarrow (V_1, \ldots, V_{i-1})
$$

is a Markov chain and, given (U_k, V_k, Z_k), the tuple $(T, V_1, \ldots, V_k, Z^k)$ is independent of $(V_1, \ldots, V_{k-1})$. As a consequence of that, Lemma 1 in ([12] Section II.B) guarantees the existence of a reconstruction $\hat{X}_k^*(U_k, V_k, Z_k)$ which dominates $\hat{X}_k$ in the sense that

$$
\mathbb{E}\left[d(X_k, \hat{X}_k^*(U_k, V_k, Z_k))\right] \leq \mathbb{E}\left[d(X_k, \hat{X}_k(T, V_1, \ldots, V_k, Z_k, Z^{k-1}))\right].
\tag{43}
$$

This observation interpreted as the "data processing inequality" for estimation has already been made in ([12] Lemma 1).

Furthermore,

$$
V_J \; = \; f_{1,J}^{(n)}(f^{(n)}(X^n), Y^J) = f_{1,J}^{(n)}(f^{(n)}(X^n), Y^{J-1}, Y_J)
$$

$$= \tilde{f}_{1,J}^{(n)}(T, X^{J-1}, Y^{J-1}, Z^{J-1}, Y_J) = \tilde{f}_{1,J}^{(n)}(U_J, J, Y_j). \tag{44}$$

By (1) and the memoryless property of the sequence $(X_k, Y_k, Z_k), k = 1, \ldots, n$ one can verify the Markov relation $U_k \leftrightarrow X_k \leftrightarrow (Y_k, Z_k)$ which implies the Markov relation $U \leftrightarrow X \leftrightarrow (Y, Z)$. Similarly, the definitions of U_k and V_k and the memoryless property of $(X_k, Y_k), k = 1, \ldots, n$ imply that, conditioned on U_k, $X_k \leftrightarrow Y_k \leftrightarrow V_k$ forms a Markov chain.

Thus, conditioned on U, $X \leftrightarrow Y \leftrightarrow V$ forms a Markov chain, hence

$$P_{XUYV} = P_{XU}\mathbb{P}_{Y|X}P_{V|YU}, \tag{45}$$

where $\mathbb{P}_{Y|X}$ denotes the conditional law induced by the marginal law P_{XY}. The combination of (40)–(42) and (44) together with the latter Markov relations establish the converse.

We shall now obtain an alternative characterization for the lower bound (38). For a law

$$P_{UXZYV} = P_U P_{X|U} P_{ZY|X} P_{V|YU},$$

and its induced conditional law $P_{XZYV|U}$, let

$$Q(P_{XZYV|U}, \tilde{D}) \triangleq \inf_{g^*: \mathcal{U} \times \mathcal{V} \times \mathcal{Z} \to \hat{\mathcal{X}} \, : \, \mathbb{E}[d(X, g^*(U,V,Z))|U=u] \leq \tilde{D}} H(V|U = u),$$

and let $\bar{Q}(P_{XZYV|U}, \cdot)$ denote the lower convex envelope of $Q(P_{XZYV|U}, \cdot)$.

Define

$$Q_s(P_{UXZYV}, D) \quad \triangleq \quad \inf_{\rho(u): \, \int \rho(u)dP_U(u) \leq D} \int_{\mathcal{U}} \bar{Q}(P_{XZYV|U}, \rho(u))dP_U(u),$$

then by ([13] Section III.C, Lemma 1) $Q_s(P_{UXZYV}, \cdot)$ is convex.

Note that

$$\begin{aligned}
H(V_k|T, X^{k-1}, Y^{k-1}, Z^{k-1}) &= H(f_{1,k}^{(n)}(T, Y^{k-1}, Y_k)|U_k) \\
&= \int H(f_{1,k}^{(n)}(t, y^{k-1}, Y_k)|U_k = u_k) \, d\mu(u_k). \tag{46}
\end{aligned}$$

The integrand on the RHS of (46) is the entropy of the scalar quantizer of $V_k \triangleq f_{1,k}^{(n)}(T, Y^{k-1}, Y_k)$ conditioned on $U_k = u_k$ where U_k is defined in (36c). Now, conditioned on $U_k = u_k$, Lemma 1 in ([12] Section II.B) ensures that

$$\begin{aligned}
\mathbb{E}\big[d(X_k, \hat{X}_k(t, y^{k-1}, f_{1,k}^{(n)}(t, y^{k-1}, Y_k), Z^k))|U_k = u_k\big] \\
\geq \mathbb{E}\big[d(X_k, g_k^*(U_k, V_k, Z_k))|U_k = u_k\big].
\end{aligned}$$

Consequently, we may lower bound the RHS of (46) as follows

$$\begin{aligned}
\int H(f_{1,k}^{(n)}(t, y^{k-1}, Y_k)|U_k = u_k) \, d\mu(u_k) &\overset{(a)}{\geq} \int \bar{Q}(P_{XZYV|U}, \mathbb{E}[d(X_k, \hat{X}_k)|U_k = u_k]) \, d\mu(u_k) \\
&\geq \int \left[\int \bar{Q}(P_{XZYV|U}, \mathbb{E}[d(X_k, g(u, V, z))|U_k = u]) \, d\mu(z|u)\right] d\mu(u) \\
&\overset{(b)}{\geq} \int \bar{Q}(P_{XZYV|U}, \mathbb{E}[d(X_k, g(u, V, Z))|U_k = u]) \, d\mu(u) \\
&\geq Q_s(P_{UXZYV}, D), \tag{47}
\end{aligned}$$

where in (a) we used the definition of Q and in (b) its convexity. The lower bound (47) may be interpreted as follows. Fix $\rho: \mathcal{U} \to \mathbb{R}^+$ and consider, for each $u \in \mathcal{U}$, time shraing of at most two scalar

quantizers for the "source" $P_{V|U=u}$ attaining a distortion level $\rho(u)$. The optimal helper time shares side-information-dependent scalar quantizers of V_k (at most two per each side-information symbol U_k), while the reconstruction at the decoder is a function of (U, V, Z).

Direct: To establish the achievability of $\mathcal{R}^*(D)$, consider the codebook construction as follows. The codebook $\mathcal{A} = \{u_1, \ldots, u_M\}$, $u_k \in \mathcal{U}^n$ is obtained by drawing the n-length sequences u_k independently of $T^\delta_{P_U}$. (For the definition of $T^\delta_{P_U}$—the set of δ-strongly typical n-sequences corresponding to a marginal law P_U and a few properties of these sequences see [14–16]).

Given the source sequence x, $f^{(n)}(x)$ is defined as follows.

1. If $x \in T^\delta_{P_X}$ the encoder searches for the first sequence $U_k = u$ in $\mathcal{A}$ such that (s.t.) $(x, u) \in T^{2\delta}_{P_{XU}}$ and sets $f^{(n)}(x) = k$.
2. If $x \notin T^\delta_{P_X}$, or $\nexists U_k \in \mathcal{A}$ s.t. $(x, u) \in T^{2\delta}_{P_{XU}}$, an encoding error is declared.

Given $f^{(n)}(x) = k$, the helper forms the sequence of descriptions $V_i = \tilde{f}_1(U_{k,i}, Z_i)$, $V_i \in [1, \ldots, L_i]$ that is sent causally to the decoder.

Decoding: Given $f^{(n)}(x) = k$ as well as the sub-sequence $V_1, \ldots, V_i$, the decoder forms the reconstruction sequence $\hat{X}_i = g(U_{k,i}, V_i, Z_i)$, $i \in [1, \ldots, n]$.

Given that (U, X) are jointly typical since (X, Y, Z) is memoryless, the Markov lemma guarantees that, for large n, with high probability (U, X, Y, Z) are also jointly typical. Since $X \leftrightarrow (U, Y) \leftrightarrow V$ forms a Markov chain, by the Markov lemma, with high probability, (X, V) as well as (X, V, Z) are jointly typical. Thus, with high probability, $(X, \hat{X})$ are jointly typical, hence the distortion constraint (10) is fulfilled for large n. That the sequence $V_1, \ldots, V_n$ can be described at a conditional entropy rate satisfying (12b)–(47) can be established along similar lines as the proof of the direct part of Theorem 2 in ([13] Section III.C). Finally, standard error probability analysis verifies that, with high probability, (U, X) are jointly typical as long as (12a) holds.

4.2. Proof of Lemma 1

If $Y_i = \sqrt{\rho} X_i + W_i$, $i = 1, 2$, then

$$Y_+ = \frac{\sqrt{\rho}}{\sqrt{2}}(X_1 + X_2) + \frac{1}{\sqrt{2}}(W_1 + W_2) \triangleq \sqrt{\rho} X_+ + W_+ \tag{48a}$$

$$Y_- = \frac{\sqrt{\rho}}{\sqrt{2}}(X_1 - X_2) + \frac{1}{\sqrt{2}}(W_1 - W_2) \triangleq \sqrt{\rho} X_- + W_- \tag{48b}$$

where W_- and W_+ are independent and the pair (W_-, W_+) is equal in distribution to the pair (W_1, W_2). Thus, the unitary transformation $(Y_1, Y_2) \mapsto (Y_-, Y_+)$ preserves the Gaussian nature of the channel and factors according to (see (53) ahead)

$$P_{UX_-X_+Y_-Y_+V} = P_{X_-X_+U}\mathbb{G}^{(\varrho)}_{Y_-|X_-}\mathbb{G}^{(\varrho)}_{Y_+|X_+}P_{V|UY_-Y_+}. \tag{49}$$

To show (22a), consider the sequence of identities

$$\lambda I(Y_+Y_-; V|U) - h(X_+X_-|U)$$
$$= \lambda I(Y_+; V|U) + \lambda I(Y_-; V|UY_+) - h(X_-|UY_+) - h(X_+|UX_-) - I(X_-; Y_+|U)$$
$$= \lambda I(Y_+; V|UX_-) + \lambda I(Y_-; V|UY_+) - h(X_+|UX_-) - h(X_-|UY_+)$$
$$+ (\lambda - 1)\big[I(Y_+; X_-|U) - I(Y_+; X_-|UV)\big] - I(Y_+; X_-|UV). \tag{50}$$

Moreover, to show (22b), consider the sequence of identities

$$\lambda I(Y_+Y_-; V|U) - h(X_+X_-|U)$$
$$= \lambda I(Y_-; V|U) + \lambda I(Y_+; V|UY_-) - h(X_+|UY_-) - h(X_-|UX_+) - I(X_+; Y_-|U)$$

$$
\begin{aligned}
&= \lambda I(Y_-; V|UX_+) + \lambda I(Y_+; V|UY_-) - h(X_+|UY_-) - h(X_-|UX_+) \\
&\quad + (\lambda - 1)\big[I(Y_-; X_+|U) - I(Y_-; X_+|UV)\big] - I(Y_-; X_+|UV).
\end{aligned}
\tag{51}
$$

Starting with (50), consider the difference

$$
I(Y_+; X_-|U) - I(Y_+; X_-|UV) = I(X_-; V|U) - I(X_-; V|UY_+)
\tag{52}
$$

under a law of the form

$$
P_{UX_-X_+Y_-Y_+V} = P_{X_-X_+U} P_{Y_-|X_-} P_{Y_+|X_+} P_{V|UY_-Y_+}
\tag{53}
$$

i.e., that $X_- \leftrightarrow Y_- \leftrightarrow V|_{(U,Y_+)}$ forms a Markov chain (see also ([9] Section VI.A, Lemma 4)).

We distinguish between the two cases:

(1)　In case that $I(X_-; Y_+|UV) = 0$, the non-negativity of mutual information implies that the expression (52) is non-negative, hence the inequalities (63) ahead hold for any $\lambda > 1$.

(2)　In case that $I(X_-; Y_+|UV) > 0$, we prove first that for the set of laws which are feasible for the optimization problem (20), the expression (52) is strictly positive.

Observe that with the choice of $U \triangleq \mathbf{U} = (U_1, U_2)$ and $V = (V_1, V_2)$ where (U_1, X_1, Y_1, V_1) and (U_2, X_2, Y_2, V_2) are two independent copies of $(\tilde{U}, X, Y, \tilde{V})$ such that $X \leftrightarrow Y \leftrightarrow \tilde{V}|_{\tilde{U}}$ and $\sigma^2_{X|\tilde{U}\tilde{V}} \le D$ i.e., $P_{\tilde{U}XY\tilde{V}} \in \mathcal{Q}$, we have

$$
\begin{aligned}
\sigma^2_{X_+|UY_-V} &\le \sigma^2_{X_+|UV} = \frac{1}{2}\sigma^2_{X_1|U_1V_1} + \frac{1}{2}\sigma^2_{X_2|U_2V_2} \le D \\
\sigma^2_{X_-|UY_+V} &\le \sigma^2_{X_-|UV} = \frac{1}{2}\sigma^2_{X_1|U_1V_1} + \frac{1}{2}\sigma^2_{X_2|U_2V_2} \le D.
\end{aligned}
\tag{54}
$$

Thus, the unitary transformation $(Y_1, Y_2) \mapsto (Y_-, Y_+)$ picks a pair of independent copies of a law $P_{\tilde{U}XY\tilde{V}} \in \mathcal{Q}$ and "creates" a pair of laws, $X_- \leftrightarrow Y_- \leftrightarrow V|_{UY_+}$ and $X_+ \leftrightarrow Y_+ \leftrightarrow V|_{UY_-}$, which factor jointly according to (53) with

$$
\begin{aligned}
P_{Y_-|X_-} &= \mathbb{G}^{(\varrho)}_{Y_-|X_-} \\
P_{Y_+|X_+} &= \mathbb{G}^{(\varrho)}_{Y_+|X_+},
\end{aligned}
$$

hence are symmetric w.r.t. the inputs X_- and X_+. We shall define the latter set of laws by $\mathscr{P}^*$ and, as shown above, $\mathscr{P}^* \subseteq \mathcal{Q}$.

Remark 2. *Suppose that $(\tilde{U}, X, Y) \sim P_{X\tilde{U}}\mathbb{G}^{(\varrho)}_{Y|X}$ with $\tilde{U} \in \mathcal{U}$ then, for both laws $X_- \leftrightarrow Y_- \leftrightarrow V|_{UY_+}$ and $X_+ \leftrightarrow Y_+ \leftrightarrow V|_{UY_-}$, we have $(U, Y_+) \in \mathcal{U} \times \mathcal{U} \times \mathbb{R}$ and $(U, Y_-) \in \mathcal{U} \times \mathcal{U} \times \mathbb{R}$. Since the image of the map*

$$
P_{X\tilde{U}} \mapsto (\mathbb{E}[X^2], \sigma^2_{X|\tilde{U}V}, \mathsf{s}_\lambda(X, \varrho|\tilde{U}))
\tag{55}
$$

is a convex set, standard dimensionality reduction argument can be used to establish the existence of a law P_{XU} where P_U is supported on a finite set and achieves any point in the image of the map (55) (See ([9] Section IV, Remark 2)).

By rate-distortion theory, the constraint $\sigma^2_{X_-|U,V} \le D$ with both U and V non-void implies that

$$
I(X_-; U) < I(X_-; UV) \quad \Rightarrow \quad I(X_-; V|U) > 0.
\tag{56}
$$

Now, for any $P_{UX_-X_+Y_-Y_+V}$ as per (49), conditioned on (U, Y_+), the random variable V is dependent on Y_- i.e. $I(Y_-; V|UY_+) > 0$. As a consequence of that, since the mutual information $I(X_-; Y_-)$ is strictly positive and, conditioned on (U, Y_+), $X_- \leftrightarrow Y_- \leftrightarrow V$ forms a Markov chain

$$I(X_-; V|UY_+) > 0. \tag{57}$$

Since a law of the form (53) dictates

$$I(X_-; V|UY_-Y_+) = 0, \tag{58}$$

the combination of (56)–(58) yields that

$$I(X_-; V|U) \geq I(X_-; V|UY_+) > I(X_-; V|UY_-Y_+) = 0.$$

Thus, the conditional mutual information $I(X_-; V|U)$ is non-increasing under the conditioning on (Y_-, Y_+).

By the symmetry of the pair of laws $X_- \leftrightarrow Y_- \leftrightarrow V|_{UY_+}$ and $X_+ \leftrightarrow Y_+ \leftrightarrow V|_{UY_-}$ induced by $P_{UX_-X_+Y_-Y_+V}$, conditioned on (U, Y_-), the random variable V is dependent on Y_+, hence

$$I(Y_+; V|UY_-) > 0. \tag{59}$$

Now

$$\begin{aligned}
P_{UX_-X_+Y_-Y_+V} &= P_{X_-X_+}u P_{Y_-|X_-} P_{Y_+|X_+} P_{V|UY_-Y_+} \\
&= P_{X_-X_+}u P_{Y_+|X_+} P_{VY_-|UX_-Y_+},
\end{aligned} \tag{60}$$

hence

$$\begin{aligned}
P_{UX_-X_+Y_+V} &= \sum_{Y_-} P_{X_-X_+}uY_-Y_+V \\
&= \sum_{Y_-} P_{X_-X_+}u P_{Y_+|X_+} P_{VY_-|UX_-Y_+} \\
&= P_{X_-X_+}u P_{Y_+|X_+} P_{V|UX_-Y_+}.
\end{aligned} \tag{61}$$

However, with the latter factorization, if indeed $I(X_-; V|U) = I(X_-; V|UY_+)$, i.e.,

$$I(X_-; V|U) = I(X_-; V|UY_+) > I(X_-; V|UY_-Y_+) = 0$$

then, since the conditional mutual information $I(X_-; V|U)$ is non-increasing under the conditioning on (Y_-, Y_+), it follows that $P_{V|UX_-Y_+} = P_{V|UX_-}$, which is in contradiction with (59). Consequently, for $P_{UX_-X_+Y_-Y_+V} \in \mathscr{P}^*$, $I(X_-; V|U) - I(X_-; V|UY_+) = \Delta > 0$, thus establishing that the expression (52) is positive.

Consequently, there exists some $\lambda^* > 1$, such that for any $\lambda \geq \lambda^* \triangleq \frac{I(Y_+; X_-|UV)}{\Delta} + 1$

$$(\lambda - 1)\big[I(Y_+; X_-|U) - I(Y_+; X_-|UV)\big] - I(Y_+; X_-|UV) > 0. \tag{62}$$

The combination of (50) and (62) implies that, for any $\lambda \geq \lambda^*$,

$$\begin{aligned}
\lambda I(Y_+Y_-&; V|\mathbf{U}) - h(X_+X_-|\mathbf{U}) \\
&\overset{(a)}{\geq} -\big[h(X_+|\mathbf{U}X_-) + h(X_-|\mathbf{U}Y_+)\big] + \lambda\big[I(Y_+; V|\mathbf{U}X_-) + I(Y_-; V|\mathbf{U}Y_+)\big] \\
&\overset{(b)}{\geq} \mathsf{s}_\lambda(X_+, \varrho|X_-, \mathbf{U}) + \mathsf{s}_\lambda(X_-, \varrho|Y_+, \mathbf{U}).
\end{aligned} \tag{63}$$

Here, (a) follows by (50) and (62), and (b) is true, since the set of laws over which the infimum on the RHS of (63) is evaluated is not empty. Indeed, the choice of $\mathbf{U} = (U_1, U_2)$ and $V = (V_1, V_2)$ where (U_1, X_1, Y_1, V_1) and (U_2, X_2, Y_2, V_2) are two independent copies of $(\tilde{U}, X, Y, \tilde{V})$ such that $X \leftrightarrow Y \leftrightarrow \tilde{V}|_{\tilde{U}}$ and $\sigma^2_{X|\tilde{U}\tilde{V}} \leq D$ –i.e., $P_{\tilde{U}XY\tilde{V}} \in \mathcal{Q}$ (hence it is feasible for $\mathsf{s}_\lambda(X, \varrho|U)$ in (19)), satisfies (54), hence it belongs to the feasible set.

On the other hand, since both mappings $(X_1, X_2) \mapsto (X_+, X_-)$ and $(Y_1, Y_2) \mapsto (Y_+, Y_-)$ are invertible, then with $\mathbf{U} = (U_1, U_2)$

$$
\begin{aligned}
\lambda I(Y_+ Y_-; V|\mathbf{U}) - h(X_+ X_-|\mathbf{U}) &= \lambda I(Y_1 Y_2; V|\mathbf{U}) - h(X_1 X_2|\mathbf{U}) \\
&= \lambda I(Y_1; V|\mathbf{U}) + \lambda I(Y_2; V|\mathbf{U}Y_1) - h(X_1|\mathbf{U}) - h(X_2|\mathbf{U}X_1) \\
&\overset{(a)}{=} \lambda I(Y_1; V|\mathbf{U}) + \lambda I(Y_2; V|\mathbf{U}Y_1) - h(X_1|\mathbf{U}) - h(X_2|\mathbf{U}) \\
&= \lambda I(Y_1; V|\mathbf{U}) + \lambda I(Y_2; V|\mathbf{U}) - h(X_1|\mathbf{U}) - h(X_2|\mathbf{U}) \\
&\quad - \lambda[I(Y_2; V|\mathbf{U}) - I(Y_2; V|\mathbf{U}Y_1)] \\
&= \lambda I(Y_1; V|\mathbf{U}) + \lambda I(Y_2; V|\mathbf{U}) - h(X_1|\mathbf{U}) - h(X_2|\mathbf{U}) \\
&\quad - \lambda[I(Y_2; V|\mathbf{U}) - h(Y_2|\mathbf{U}Y_1) + h(Y_2; |\mathbf{U}VY_1)] \\
&\overset{(b)}{=} \lambda I(Y_1; V|\mathbf{U}) + \lambda I(Y_2; V|\mathbf{U}) - h(X_1|\mathbf{U}) - h(X_2|\mathbf{U}) \\
&\quad - \lambda[I(Y_2; V|\mathbf{U}) - h(Y_2|\mathbf{U}) + h(Y_2; |\mathbf{U}VY_1)] \\
&= \lambda I(Y_1; V|\mathbf{U}) + \lambda I(Y_2; V|\mathbf{U}) - h(X_1|\mathbf{U}) - h(X_2|\mathbf{U}) \\
&\quad + \lambda I(Y_2; Y_1|\mathbf{U}V).
\end{aligned}
\tag{64}
$$

Here, (a) follows since, conditioned on $\mathbf{U}$, X_1 and X_2 are independent, and (b) follows since, conditioned on $\mathbf{U}$, Y_1 and Y_2 are independent.

When $\mathbf{U} = (U_1, U_2)$ and $V = (V_1, V_2)$ where (U_1, X_1, Y_1, V_1) and (U_2, X_2, Y_2, V_2) are two independent copies of $(\tilde{U}, X, Y, \tilde{V})$, such that $X \leftrightarrow Y \leftrightarrow \tilde{V}|_{\tilde{U}}$ and $\sigma^2_{X|\tilde{U}\tilde{V}} \leq D$, then $Y_2 \leftrightarrow (\mathbf{U}, V) \leftrightarrow Y_1$ forms a Markov chain. Thus, $I(Y_1; Y_2|\mathbf{U}V) = 0$ and the RHS of (64) becomes $\lambda I(Y_1; V_1|U_1) + \lambda I(Y_2; V_2|U_2) - h(X_1|U_1) - h(X_2|U_2)$ thus establishing that,

$$
\begin{aligned}
&\inf_{\substack{V: \, X \leftrightarrow Y \leftrightarrow V|_{\mathbf{U}} \\ \sigma^2_{X|UV} \leq D}} \left\{ -h(X_1, X_2|\mathbf{U}) + \lambda I(Y_1, Y_2; V|\mathbf{U}) \right\} \\
&\qquad\qquad \leq \sum_{i=1}^{2} \inf_{\substack{V: \, X_i \leftrightarrow Y_i \leftrightarrow V|_{U_i} \\ \sigma^2_{X_i|U_iV} \leq D}} \left\{ -h(X_i|U_i) + \lambda I(Y_i; V|U_i) \right\} = 2\,\mathsf{s}_\lambda(X, \varrho|U),
\end{aligned}
\tag{65}
$$

where the inequality follows, since the infimum on the LHS is taken over a larger set than that on the RHS. The combination of (63) and (65) establishes (22a) for $\lambda \geq \lambda^*$.

Now, returning to (51), consider the difference

$$
I(Y_-; X_+|U) - I(Y_-; X_+|UV) = I(X_+; V|U) - I(X_+; V|UY_-)
\tag{66}
$$

under the law (53) i.e., that $X_+ \leftrightarrow Y_+ \leftrightarrow V|_{(U,Y_-)}$ forms a Markov chain. By rate-distortion theory, the constraint $\sigma^2_{X_+|U,V} \leq D$ with both U and V non-void implies that

$$
I(X_+; U) < I(X_+; UV) \quad \Rightarrow \quad I(X_+; V|U) > 0.
\tag{67}
$$

Moreover, an argument similar to that leading to the conclusion that (52) is non-negative establishes that (66) is non-negative.

Consequently, there exist some $\lambda^* > 1$ such that for any $\lambda \geq \lambda^*$

$$(\lambda - 1)\left[I(Y_-; X_+|U) - I(Y_-; X_+|UV)\right] - I(Y_-; X_+|UV) > 0. \tag{68}$$

In addition, the combination of (51) and (68) implies that, for any $\lambda \geq \lambda^*$,

$$\lambda I(Y_+Y_-; V|\mathbf{U}) - h(X_+X_-|\mathbf{U}) \geq \mathsf{s}_\lambda(X_-, \varrho|X_+, \mathbf{U}) + \mathsf{s}_\lambda(X_+, \varrho|Y_-, \mathbf{U}). \tag{69}$$

The combination of (69) and (65) establishes (22b) for $\lambda \geq \lambda^*$.

4.3. Proof of Lemma 2

Consider the tuple $(X_+, X_-, Y_+, Y_-, \mathbf{U})$ as defined in Lemma 1 that is constructed from independent copies of $(U, X, Y) \sim P_{XU}\mathbb{G}_{Y|X}^{(\varrho)}$. With the transformation (21) X_+ and X_- preserve the variance of $X_i, i = 1, 2$ and W_+ and W_- preserve the variance of $W_i, i = 1, 2$. Furthermore, as shown in the proof of Lemma 1, the unitary transformation $(Y_1, Y_2) \mapsto (Y_-, Y_+)$ picks a pair of independent copies of a law $P_{UXYV} \in \mathscr{Q}$ and "creates" a pair of laws, $X_- \leftrightarrow Y_- \leftrightarrow V|_{UY_+}$ and $X_+ \leftrightarrow Y_+ \leftrightarrow V|_{UY_-}$, which factor jointly according to (53) with

$$\begin{aligned} P_{Y_-|X_-} &= \mathbb{G}_{Y_-|X_-}^{(\varrho)} \\ P_{Y_+|X_+} &= \mathbb{G}_{Y_+|X_+}^{(\varrho)}, \end{aligned}$$

hence are symmetric w.r.t., the inputs X_- and X_+ and both are in the feasible set. Similarly, the unitary transformation $(Y_1, Y_2) \mapsto (Y_-, Y_+)$ picks a pair of independent copies of a law $P_{UXYV} \in \mathscr{Q}$ and "creates" a pair of laws, $X_- \leftrightarrow Y_- \leftrightarrow V|_{UX_+}$ and $X_+ \leftrightarrow Y_+ \leftrightarrow V|_{UX_-}$, which are both in the feasible set (see the factorization (61)).

Consequently, by (20), each of the quantities $\mathsf{s}_\lambda(X_+, \varrho|X_-, \mathbf{U})$, $\mathsf{s}_\lambda(X_-, \varrho|Y_+, \mathbf{U})$, $\mathsf{s}_\lambda(X_-, \varrho|X_+, \mathbf{U})$ and $\mathsf{s}_\lambda(X_+, \varrho|Y_-, \mathbf{U})$ is at least $\mathsf{V}_\lambda(\varrho)$. The assumption $P_{XU} \in \mathscr{P}$ leads by (22) of Lemma 1 to an opposite conclusion. Therefore, all four quantities are equal and coincide with the minimum, i.e.,

$$\begin{aligned} \mathsf{s}_\lambda(X_-, \varrho|Y_+, \mathbf{U}) &= \mathsf{s}_\lambda(X_+, \varrho|Y_-, \mathbf{U}) = \mathsf{V}_\lambda(\varrho) \tag{70a} \\ \mathsf{s}_\lambda(X_-, \varrho|X_+, \mathbf{U}) &= \mathsf{s}_\lambda(X_+, \varrho|X_-, \mathbf{U}) = \mathsf{V}_\lambda(\varrho). \tag{70b} \end{aligned}$$

Let $B \sim \mathrm{Ber}(1/2)$ be a Bernoulli random variable with values in the set $\{+, -\}$ independent of $(X_-, X_+, Y_-, Y_+, \mathbf{U})$. Let $\bar{B}$ be the complement of B in the set $\{+, -\}$ and define $(\tilde{X}, \tilde{\mathbf{U}})$ by $\tilde{X} \triangleq X_B$ and $\tilde{\mathbf{U}} \triangleq (B, Y_{\bar{B}}, \mathbf{U})$. Then

$$\begin{aligned} \mathsf{s}_\lambda(\tilde{X}, \varrho|\tilde{\mathbf{U}}) &= \frac{1}{2}\mathsf{s}_\lambda(X_-, \varrho|Y_+, \mathbf{U}) + \frac{1}{2}\mathsf{s}_\lambda(X_+, \varrho|Y_-, \mathbf{U}) \\ &\leq \mathsf{V}_\lambda(\varrho) + 2\epsilon, \tag{71} \end{aligned}$$

where the inequality follows by assumption (24a).

Furthermore, $\mathbb{E}[(\tilde{X})^2] = \mathbb{E}[X^2]$ and by (54)

$$\sigma_{\tilde{X}|\mathbf{U}V}^2 \leq \frac{1}{2}\sigma_{X_+|\mathbf{U}V}^2 + \frac{1}{2}\sigma_{X_-|\mathbf{U}V}^2 \leq D. \tag{72}$$

Next

$$\begin{aligned} h(\tilde{Y}|\tilde{\mathbf{U}}) &= h(Y_B|B, Y_{\bar{B}}, \mathbf{U}) = \frac{1}{2}h(Y_+|Y_-\mathbf{U}) + \frac{1}{2}h(Y_-|Y_+\mathbf{U}) \\ h(\tilde{X}|\tilde{\mathbf{U}}) &= h(X_B|B, Y_{\bar{B}}, \mathbf{U}) = \frac{1}{2}h(X_+|Y_-\mathbf{U}) + \frac{1}{2}h(X_-|Y_+\mathbf{U}). \end{aligned}$$

Thus

$$
\begin{aligned}
h(\tilde{Y}|\tilde{\mathbf{U}}) - h(\tilde{X}|\tilde{\mathbf{U}}) &= \frac{1}{2}\big[h(Y_+|Y_-\mathbf{U}) + h(Y_-|Y_+\mathbf{U})\big] - \frac{1}{2}\big[h(X_+|Y_-\mathbf{U}) + h(X_-|Y_+\mathbf{U})\big] \\
&= \frac{1}{2}\big[h(Y_+Y_-|\mathbf{U}) - h(Y_-|\mathbf{U}) + h(Y_-|Y_+\mathbf{U})\big] - \frac{1}{2}\big[h(X_+|Y_-\mathbf{U}) + h(X_-|Y_+\mathbf{U})\big] \\
&= \frac{1}{2}\big[h(Y_+Y_-|\mathbf{U}) - h(X_+X_-|\mathbf{U})\big] - \frac{1}{2}I(Y_-;Y_+|\mathbf{U}) \\
&\qquad + \frac{1}{2}\big[h(X_+|\mathbf{U}) + h(X_-|X_+\mathbf{U})\big] - \frac{1}{2}\big[h(X_+|Y_-\mathbf{U}) + h(X_-|Y_+\mathbf{U})\big] \\
&\overset{(a)}{=} \frac{1}{2}\big[h(Y_+Y_-|\mathbf{U}) - h(X_+X_-|\mathbf{U})\big] - \frac{1}{2}I(Y_-;Y_+|\mathbf{U}) + \frac{1}{2}I(X_+;Y_-|\mathbf{U}) \\
&\qquad + \frac{1}{2}\big[h(X_-|X_+Y_+\mathbf{U}) - h(X_-|Y_+\mathbf{U})\big] \\
&\overset{(b)}{=} \frac{1}{2}\big[h(Y_+Y_-|\mathbf{U}) - h(X_+X_-|\mathbf{U})\big] + \frac{1}{2}\big[h(Y_-|Y_+\mathbf{U}) - h(Y_-|X_+Y_+\mathbf{U})\big] \\
&\qquad - \frac{1}{2}I(X_-;X_+|Y_+\mathbf{U}) \\
&= \frac{1}{2}\big[h(Y_+Y_-|\mathbf{U}) - h(X_+X_-|\mathbf{U})\big] + \frac{1}{2}I(Y_-;X_+|Y_+\mathbf{U}) - \frac{1}{2}I(X_-;X_+|Y_+\mathbf{U}) \\
&\overset{(c)}{=} \frac{1}{2}\big[h(Y_+Y_-|\mathbf{U}) - h(X_+X_-|\mathbf{U})\big] + \frac{1}{2}I(Y_-;X_+|Y_+\mathbf{U}) - \frac{1}{2}I(X_-Y_-;X_+|Y_+\mathbf{U}) \\
&= \frac{1}{2}\big[h(Y_+Y_-|\mathbf{U}) - h(X_+X_-|\mathbf{U})\big] - \frac{1}{2}I(X_-;X_+|Y_-Y_+\mathbf{U}) \\
&= h(Y|U) - h(X|U) - \frac{1}{2}I(X_-;X_+|Y_-Y_+\mathbf{U}).
\end{aligned}
\tag{73}
$$

Here,

(*a*)　follows since $X_- \leftrightarrow (X_+\mathbf{U}) \leftrightarrow Y_+$ forms a Markov chain,

(*b*)　follows since $Y_- \leftrightarrow (X_+\mathbf{U}) \leftrightarrow Y_+$ forms a Markov chain, and

(*c*)　follows since $Y_- \leftrightarrow (X_-Y_+\mathbf{U}) \leftrightarrow X_+$ forms a Markov chain.

The combination of identity (73) with inequality (24b) proves (26).

Since $\tilde{X} \in \mathbb{R}$ while $\tilde{\mathbf{U}} \in \{+,-\} \times \mathbb{R} \times \mathcal{U} \times \mathcal{U}$ a dimensionality reduction argument as in Remark 2 establishes the existence of a law $P_{X'U'}$, where U' has finite support, such that (25) and (26) are fulfilled, hence (U', X') are in the feasible set.

Conflicts of Interest: The author declares no conflict of interest.

References

1. Wyner, A.D.; Ziv, J. The rate-distortion function for source coding with side information at the receiver. *IEEE Trans. Inform. Theory* **1976**, *IT-22*, 1–11. [CrossRef]
2. Steinberg, Y.; Merhav, N. On successive refinement for the Wyner-Ziv problem. *IEEE Trans. Inform. Theory* **2004**, *50*, 1636–1654. [CrossRef]
3. Bross, S.I.; Weissman, T. On successive refinement for the Wyner-Ziv problem with partially cooperating decoders. In Proceedings of the 2008 IEEE International Symposium on Information Theory (ISIT), Toronto, ON, Canada, 6–11 July 2008.
4. Weissman, T.; Gamal, A.E. Source coding with limited-look-ahead side information at the decoder. *IEEE Trans. Inform. Theory* **2006**, *52*, 5218–5239. [CrossRef]
5. Maor, A.; Merhav, N. On successive refinement with causal side information at the decoders. *IEEE Trans. Inform. Theory* **2008**, *54*, 332–343. [CrossRef]
6. Bross, S.I. Scalable source coding with causal side information and a causal helper. In Proceedings of the 2020 IEEE International Symposium on Information Theory (ISIT), Los Angeles, CA, USA, 21–26 June 2020.

7.　Geng, Y.; Gohari, A.; Nair, C.; Yu, Y. The capacity region of classes of product broadcast channels. *arXiv* **2011**, arXiv:1105.5438.

8.　Geng, Y.; Nair, C. The capacity region of the two-receiver Gaussian vector broadcast channel with private and common messages. *IEEE Trans. Inform. Theory* **2014**, *IT-60*, 2087–2104. [CrossRef]

9.　Courtade, T.A. A strong entropy power inequality. *IEEE Trans. Inform. Theory* **2018**, *IT-64*, 2173–2192. [CrossRef]

10.　Bernstein, S.N. On a property characteristic of the normal law. *Trudy Leningrad. Polytech. Inst.* **1941**, *3*, 21–22.

11.　Bryc, W. *The Normal Distribution: Characterizations with Applications*; Springer: New York, NY, USA, 2012; Volume 100.

12.　Choudhuri, C.; Kim, Y.-H.; Mitra, U. Causal state communication. *IEEE Trans. Inform. Theory* **2013**, *59*, 3709–3719. [CrossRef]

13.　Weissman, T.; Merhav, N. On causal source codes with side information. *IEEE Trans. Inform. Theory* **2005**, *51*, 4003–4013. [CrossRef]

14.　Cover, T.M.; Thomas, J.A. *Elements of Information Theory*; John Wiley and Sons: New York, NY, USA, 1991.

15.　Csiszár, I.; Körner, J. *Information Theory: Coding Theorems for Discrete Memoryless Systems*; Academic: New York, NY, USA, 1981.

16.　Gamal, A.E.; Kim, Y.H. *Network Information Theory*; Cambridge University Press: Cambridge, UK, 2012.