On Hierarchical Joint Source–Channel Coding

Yossef Steinberg and Neri Merhav

Department of Electrical Engineering
Technion - Israel Institute of Technology
Technion City, Haifa 32000, ISRAEL
[ysteinbe,merhav]@ee.technion.ac.il

November 20, 2003

Abstract

We extend the setting of two–stage lossy source coding with successive refinement structures into a joint source–channel coding setting. In particular, we consider a problem where two descriptions of a memoryless source are to be transmitted across two independent memoryless channels and where the output of the channel corresponding to the first (coarse) description is also available to the decoder of the second (refinement) decoder. Side information, correlated to the source, may also be available to the decoders. Our first result is a separation theorem asserting that in the limit of long blocks, no optimality is lost by first applying lossy successive–refinement source coding, regardless of the channels, and then applying good channel codes to each one of the resulting bitstreams, regardess of the source and the side information. It is also shown that (even noiseless) feedback from the output of the first channel to the input of the second encoder cannot improve performance, but may sometimes facilitate the implementation of optimum codes significantly: In certain situations, even single–letter codes (of unit block length) may achieve optimum performance. Necessary and sufficient conditions are furnished for the optimality of single–letter codes with and without feedback.

Index terms — Hierarchical coding, joint source–channel coding, side information, successive refinement, systematic coding, Wyner-Ziv problem.

1 Introduction

The problem of lossy source coding in two or more stages of successive refinement has received quite considerable attention throughout the last few decades (see, e.g., [2], [5], [6], [8], [9], [10], [11],

[13], [15], [16], [17] and references therein). The successive refinement problem is a special case of a more general problem, called the multiple description problem, and it is about characterizing the region of pairs of rates and corresponding pairs of distortion levels, which are achievable by hierarchical codes that work simultaneously at two points in the rate–distortion plane, where the higher rate description of the source consists of the lower rate description plus an additional, refinement description. The concepts and most of the principal results extend to any finite number of stages of codes.

An interesting question that has not received attention thus far, to the best of our knowledge, evolves around the natural combination of successive refinement source coding and channel coding: Suppose that after transmitting a certain description of the source across a given channel, using a certain joint source–channel code (or, in particular, by separate source and channel coding), additional channel resources become available (time slots, bandwidth, etc.), and we wish then to transmit an additional (refinement) message, on the top of the one that has already been transmitted, in order to improve on the quality of the reproduction. Several questions then naturally arise: What would be the best stragtegy for joint source–channel coding in the two stages? What is the best performance attainable? Does the separation principle apply? What happens in the presence of feedback from the output of the first channel (of the coarse description) to the refinement encoder? What happens if there is side information (SI), correlated to the source, available at the two receivers?

In this paper, we make an attempt to address these questions. In the most basic setting of the problem, we consider a communication system with the following structure (see Fig. 1, ignoring, for the moment, the decoder inputs V_1 and V_2): a block drawn from a memoryless source is mapped, by two joint source–channel encoders, directly into two channel input vectors, the first corresponding to a coarse description, and the second – to an additional, refinement description. These two descriptions are transmitted separately via two independent memoryless channels, in compliance with certain limitations on average transmission costs (generalized power constraints), Γ_1 and Γ_2, and are received by two receivers that provide estimates of the source vector, with distortion levels D_1 and D_2, where it is assumed that receiver no. 2, that is connected to the second channel, has access also to the output of the first channel.

Our first result is a single–letter characterization of the region achievable distortion-cost quadruples $(D_1, \Gamma_1, D_2, \Gamma_2)$. The main feature of this characterization is that it admits a separation principle, which tells that no optimality is lost if one first applies optimum lossy successive–refinement source coding, independently of the channels, and then apply good channel coding for each one of the compressed bitstreams, independently of the source. Earlier, we raised the question of the potential impact of feedback from the output of the first channel to the input of the second, refinement transmitter. One might speculate that in the presence of such feedback, the second transmitter could be significantly improved, for example, by implementing, at the second stage encoder, a copy of the first decoder, subtracting the estimated source vector from the original source vector (whenever substraction is well–defined), and then encoding this estimation error vector of the first–stage decoder (in the spirit of the idea of differential/predicitive coding techniques, frequently encountered in applications of audio and video coding techniques). It is then perhaps somewhat surprising that neither this idea of using the feedback, nor any other idea one may have, can improve asymptotic performance: The same achievable region continues to apply even in the presence of feedback. In this context, it may not only be surprising that feedback does not help, but also that the separation principle continues to apply, because *both* the source encoder and the channel encoder of the refinement stage may ignore the feedback information altogether, without loss of asymptotic optimality, in the limit of long block coding. Also, from the mathematical point of view, feedback breaks the Markov structure of the communication system, which in the classical case, gives rise to the data processing theorem, the standard tool for proving the converse to the joint source–channel coding theorem.

It turns out, however, that similarly to the well–known features of feedback in the classical case,[1] here too, feedback sometimes enables optimum performance using extremely simple systems, such as single–letter codes (i.e., codes of unit block length). On the other hand, in the absence of feedback, optimum performance may not be achievable, in general, using single–letter codes. Indeed, as we demonstrate in two examples, optimum performance using single–letter codes is sometimes achievable by applying the above described idea of transmitting, at the second stage, the estimation error signal of the first stage decoder. This observation motivates us to furnish general conditions under which single–letter codes are optimal with and without feedback, in the spirit of [7] (see also [4]). An interesting corollary of these conditions is that, under the quadratic distortion measure, feedback is necessary to achieve optimality with single-letter codes.

Another ingredient that we incorporate into our model of two–stage joint source–channel coding

[1]For example, the capacity of the erasure channel with feedback is well–known to be easily achievable by repeated repetition requests until the first successful reception.

system is the possible availability of side information (SI) streams, correlated to the source, at the two decoders (cf. V_1 and V_2 in Fig. 1). The main motivation is that it makes the present work a joint extension of two earlier works, where the models considered include SI at the decoder(s). The first is [15], where the pure source coding problem in two stages and in the presence of SI at the decoders (that is, successive refinement for the Wyner–Ziv problem [18]) was studied. In other words, [15] is a special case of the present work, where the channels are noiseless. The other work is [14], where there is only one stage of coding and only one (noisy) channel through which coded information is transmitted. One of the motivations of the model in [14] is, as its title suggests, systematic lossy joint source–channel coding, where the SI channel (corresponding to the conditional probability distribution of the SI given the source) is thought of as a channel that conveys uncoded transmission of the source, that is, the systematic part of the code. The model we present in this work then gives rise to three layers of information, the first layer being of uncoded information (the systematic part), the second is a relatively weak code, and the third one is a stronger code. Alternatively, another way to look at it, is as a system of systematic source–channel coding with a successive–refinement structure.

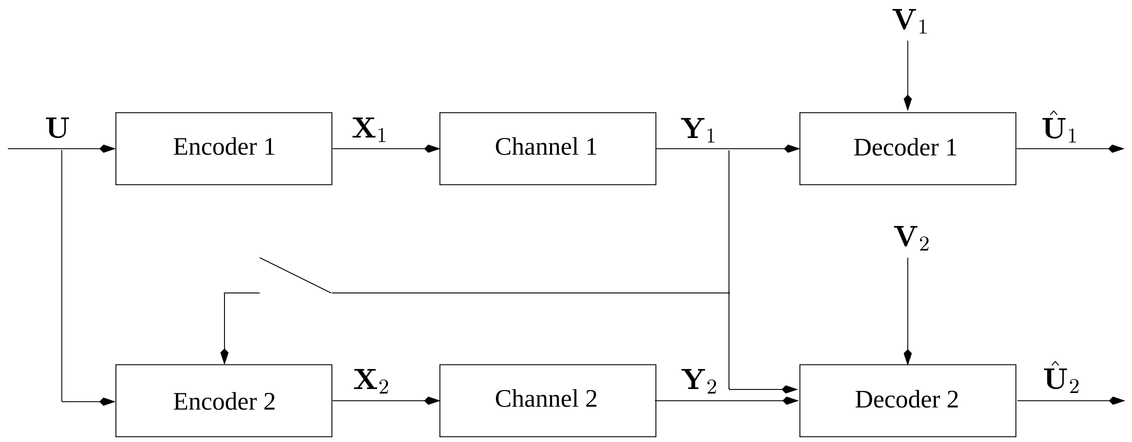

Figure 1: Hierarchical joint source-channel coding with side information at the receivers.

To summarize then, the model we consider serves as a common umbrella to the notions of Wyner–Ziv coding, successive refinement, and joint source–channel coding with the option of (block-wise) feedback. Beyond the motivations that have already been dicussed, it is hoped that this unified perspective will give rise to new insights.

The outline of the paper is as follows. In Section 2, we define the notation conventions as well as some terminology and give a formal definition of the problem. In Section 3, we give the main result on the characterization of the achievable region $(D_1, \Gamma_1, D_2, \Gamma_2)$ in the presence of SI, with/without feedback. Section 4 is a preparatory section (for Section 5), which defines notions of optimality of a point $(D_1, \Gamma_1, D_2, \Gamma_2)$ and relates it to optimality of a working point in successive refinement source coding for the Wyner–Ziv problem [15]. In Section 5, the results of Sections 3 and 4 are used in order to establish conditions on optimality of single–letter codes. Finally, in Section 6, the proofs of the main results are provided.

2 Notation, Preliminaries, and Problem Formulation

We begin by setting up notation conventions. Throughout this paper, scalar RVs will be denoted by capital letters, their sample values will be denoted by the respective lower case letters, and their alphabets will be denoted by the respective calligraphic letters. A similar convention will apply to random vectors and their sample values, which will be denoted by the same symbols superscripted by the dimension. When the dimension is clear from the context, boldface fonts will sometimes be used instead of superscripts. Thus, for example, U^n or $\boldsymbol{U}$ will denote a random n-vector $(U_1, ..., U_n)$, and $u^n \equiv \boldsymbol{u} = (u_1, ..., u_n)$ is a specific vector value in $\mathcal{U}^n$, the n-th Cartesian power of $\mathcal{U}$. The notations u_i^j and U_i^j, where i and j are integers and $i \leq j$, will designate segments $(u_i, \ldots, u_j)$ and $(U_i, \ldots, U_j)$, respectively, where for $i = 1$, the subscript will be omitted (as above). For $i > j$, u_i^j (or U_i^j) will be understood as the null string. Some of the vectors, in our setting, will have common notation but with different subscripts, for example, $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$, or $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$. In this case, the t-th coordinate of $\boldsymbol{X}_1$ will be denoted by $\boldsymbol{X}_{1,t}$, the substring $(y_{2,i}, y_{2,i+1}, \cdots, y_{2,j})$, for $j > i$, will be denoted $\boldsymbol{y}_{2,i}^j$, etc. The alphabets of all random variables, throughout most of this paper, will be assumed finite unless specified otherwise.

Sources and channels will be denoted generically by the letter P or Q, subscripted by the name of the RV and its conditioning, if applicable, e.g., $P_U(u)$ is the probability function of U at the point $U = u$, $P_{Z|S}(z|s)$ is the conditional probability of $Z = z$ given $S = s$, and so on. Whenever clear from the context, these subscripts will be omitted. The notation $\mathbb{E}$ will denote the expectation operator, and it will be subscripted by the probability measure w.r.t. which the expectation is taken, whenever this is may not be clear from the context. Information theoretic quantities like entropies, mutual informations and divergences will be denoted following the usual conventions of the information theory literature, e.g., $H(U^N)$, $I(Z^n;W^k)$, $D(P_{Y|XS}\|P_Y)$, and so on.

We refer to the communication system depicted in Fig. 1. Consider a source, $\{(U_i,V_{1,i},V_{2,i})\}_{i=1}^{\infty}$, of independent copies of a triplet of RVs, (U,V_1,V_2), taking values in $\mathcal{U}\times\mathcal{V}_1\times\mathcal{V}_2$, and drawn under a joint distribution $P_{UV_1V_2}$. The process $\{U_i\}$ is the part of the source to be encoded, whereas $\{V_{1,i}\}$ and $\{V_{2,i}\}$ will play the roles of SI streams (correlated to the source), available to decoder no. 1 and decoder no. 2, respectively. Consider next two memoryless channels. Channel no. i operates at the relative rate of ρ_i channel uses per source symbol (triplet), and is defined by

$$P_{\boldsymbol{Y}_i|\boldsymbol{X}_i}(\boldsymbol{y}_i|\boldsymbol{x}_i)=\prod_t P_{Y_i|X_i}(y_{i,t}|x_{i,t}),\quad i=1,2. \tag{1}$$

The number ρ_i is referred to as the *bandwidth expansion factor* of channel no. i. A block of length n, $U^n=(U_1,\ldots,U_n)$, generated by the source, is fed into two joint source–channel encoders: Encoder no. i produces a block of length $n_i=\rho_i n$, $\boldsymbol{X}_i=(X_{i,1},\ldots,X_{i,n_i})\in(\mathcal{X}_i)^{n_i}$, which is the input to channel no. i, $i=1,2$. Each transmitted block $\boldsymbol{X}_i$ must satisfy a transmission cost constraint (generalized power constraint):

$$\mathbb{E}\phi_i(\boldsymbol{X}_i)\le n_i\Gamma_i, \tag{2}$$

where $\phi_i(\boldsymbol{x}_i)=\sum_{t=1}^{n_i}\phi_i(x_{i,t})$ for some cost function $\phi_i:\mathcal{X}_i\to\mathbb{R}^+$. In the absence of a transmission cost constraint at channel no. i, we set $\Gamma_i=\infty$ (or define $\phi_i\equiv 0$). The output of channel no. 1, $\boldsymbol{Y}_1=(Y_{1,1},\ldots,Y_{1,n_1})\in(\mathcal{Y}_1)^{n_1}$ is fed into both decoders, whereas the output of channel no. 2, $\boldsymbol{Y}_2=(Y_{2,1},\ldots,Y_{2,n_2})\in(\mathcal{Y}_2)^{n_2}$ is fed into decoder no. 2 only. In the case of a system with feedback, $\boldsymbol{Y}_1$ is fed also into encoder no. 2 (in addition to U^n). Decoder no. i is also fed by the SI block $\boldsymbol{V}_i=(V_{i,1},\ldots,V_{i,n})\in(\mathcal{V}_i)^n$ and produces an estimate of the source $\hat{\boldsymbol{U}}_i=(\hat{U}_{i,1},\ldots,\hat{U}_{i,n})\in\hat{\mathcal{U}}^n$, $i=1,2$. The quality of each estimate is judged in terms of the expectation of an additive distortion measure:

$$d_i(\boldsymbol{U},\hat{\boldsymbol{U}}_i)=\sum_{l=1}^{n}d_i(U_l,\hat{U}_{i,l}) \tag{3}$$

where $d_i(u,\hat{u}_i)\ge 0$, $u\in\mathcal{U}$, $\hat{u}_i\in\hat{\mathcal{U}}$, $i=1,2$, is a given single–letter distortion measure. As usual in rate-distortion theory, we assume that for every $u\in\mathcal{U}$, there exist letters $\hat{u}_i\in\hat{\mathcal{U}}_i$, $i=1,2$, such that $d_i(u,\hat{u}_i)=0$.

Definition 1 *An $(n,n_1,n_2,D_1,\Gamma_1,D_2,\Gamma_2)$ joint source–channel code of successive refinement for the source $P_{UV_1V_2}$ and the channels $P_{Y_1|X_1}$, $P_{Y_2|X_2}$, consists of a first–stage encoder–decoder pair*

(f_1, g_1)

$$f_1 : \mathcal{U}^n \to (\mathcal{X}_1)^{n_1} \tag{4}$$

$$g_1 : (\mathcal{Y}_1)^{n_1} \times (\mathcal{V}_1)^n \to \hat{\mathcal{U}}^n \tag{5}$$

and a second–stage encoder–decoder pair (f_2, g_2)

$$f_2 : \mathcal{U}^n \to (\mathcal{X}_2)^{n_2} \qquad (f_2 : \mathcal{U}^n \times (\mathcal{Y}_1)^{n_1} \to (\mathcal{X}_2)^{n_2} \quad \text{in the case of feedback}) \tag{6}$$

$$g_2 : (\mathcal{Y}_1)^{n_1} \times (\mathcal{Y}_2)^{n_2} \times (\mathcal{V}_2)^n \to \hat{\mathcal{U}}^n \tag{7}$$

such that

$$\begin{aligned} \mathbb{E}d_1(\boldsymbol{U}, g_1(\boldsymbol{Y}_1, \boldsymbol{V}_1)) &\leq nD_1, & (8)\\ \mathbb{E}d_2(\boldsymbol{U}, g_2(\boldsymbol{Y}_1, \boldsymbol{Y}_2, \boldsymbol{V}_2)) &\leq nD_2, & (9)\\ \mathbb{E}\phi_1(f_1(\boldsymbol{U})) &\leq n_1\Gamma_1, & (10)\\ \mathbb{E}\phi_2(f_2(\boldsymbol{U})) &\leq n_2\Gamma_2, & (11) \end{aligned}$$

where, again, in the case of feedback, $f_2(\boldsymbol{U})$*, in the last inequality, is replaced by* $f_2(\boldsymbol{U}, \boldsymbol{Y}_1)$*.*

We will say that the code (f_1, g_1, f_2, g_2) *incurs a distortion-cost quadruple* $(D_1, \Gamma_1, D_2, \Gamma_2)$, when the last four inequalities are replaced by equalities.

Definition 2 *A distortion–cost quadruple* $(D_1, \Gamma_1, D_2, \Gamma_2)$ *is said to be* achievable *with bandwidth expansion factors* ρ_1 *and* ρ_2*, if for every* $\epsilon > 0$*, and sufficiently large blocklength* n *there exists an* $(n, \rho_1 n, \rho_2 n, D_1 + \epsilon, \Gamma_1, D_2 + \epsilon, \Gamma_2)$ *joint source-channel code of successive refinement for the source* $P_{UV_1V_2}$ *and the channels* $P_{Y_1|X_1}$*,* $P_{Y_2|X_2}$*. The* distortion–cost region*, denoted* $\mathcal{D}(\rho_1, \rho_2)$*, is the closure of the set of all quadruples* $(D_1, \Gamma_1, D_2, \Gamma_2)$ *that are achievable with bandwidth expansion factors* ρ_1 *and* ρ_2*.*

Since the bandwidth expansion factors, ρ_1 and ρ_2, will be fixed parameters throughout the sequel, we will use the shorthand terminology of "achievable distortion–cost quadruple", without explicit reference to the bandwidth expansion factors. By the same token, $\mathcal{D}(\rho_1, \rho_2)$ will occasionally be replaced by the shorthand notation $\mathcal{D}$, without denoting explicitely the dependence on ρ_1 and ρ_2.

From Definition 1, and similarly as in [15], it is clear that the distortion–cost performance depends on $P_{UV_1V_2}$ only via its marginals P_{UV_1} and P_{UV_2}. We say then, that a source $P_{UV_1V_2}$ is *stochastically degraded* if there exists a distribution $P_{\tilde{U}\tilde{V}_1\tilde{V}_2}$ such that $P_{U\tilde{V}_1} = P_{UV_1}$, $P_{U\tilde{V}_2} = P_{UV_2}$, and $U \multimap \tilde{V}_2 \multimap \tilde{V}_1$ is a Markov chain. Since the distortion–cost region depends on the source only via its marginals, no distinction has to be made between physically degraded (Markov structured) and stochastically degraded sources. As is the case in [15], we restrict attention in this work to degraded SI sources.

Our first objective is to provide a single–letter characterization of $\mathcal{D}$ and to propose strategies for (asymptotically) achieving any given point in $\mathcal{D}$. Since this characterization is strongly based on the main results of [15], we now pause to provide a brief description of these results. As mentioned in the Introduction, in [15], we considered the special case of the present model, where the channels are clean, namely, the problem of pure source coding with a successive refinement structure, in the presence of SI at the decoders. In particular, the main result of [15] is a characterization of the achievable region of rate–distortion quardruples $\{(R_1, D_1, \Delta R, D_2)\}$, where R_1 is the rate (in bits/symbol) at the first stage, $\Delta R \stackrel{\triangle}{=} R_2 - R_1$ is the rate in the second stage, R_2 is the total rate at both stages, and D_1 and D_2 are the distortion levels of the two reconstructions. The main result of [15] (in the notation of the present paper) is the following:

Theorem 1 *[15] A rate–distortion quadruple $(R_1, D_1, \Delta R, D_2)$ is achievable if and only if there exists a triple of random variables (W, S, Z), taking values in finite alphabets, $\mathcal{W}$, $\mathcal{S}$, $\mathcal{Z}$, respectively, such that the following conditions are satisfied:*

1. *$(W, S, Z) \multimap U \multimap V_2 \multimap V_1$ is a Markov chain.*

2. *There exist deterministic mappings $f_1 : \mathcal{W} \times \mathcal{V}_1 \to \hat{\mathcal{U}}$ and $f_2 : \mathcal{Z} \times \mathcal{V}_2 \to \hat{\mathcal{U}}$ such that*

$$\mathbb{E}d(X, f_1(W, V_1)) \leq D_1 \tag{12}$$
$$\mathbb{E}d(X, f_2(Z, V_2)) \leq D_2 \tag{13}$$

3. *The alphabets $\mathcal{W}, \mathcal{S}, \mathcal{Z}$ satisfy $|\mathcal{W}| \leq |\mathcal{U}| + 2$, $|\mathcal{S}| \leq (|\mathcal{U}| + 1)^2$, and $|\mathcal{Z}| \leq |\mathcal{U}|(|\mathcal{U}| + 2)(|\mathcal{U}| + 1)^2 + 1$.*

4. *The rates R_1 and R_2 satisfy*

$$I(U; W|V_1) + I(U; S|W, V_2) \leq R_1, \tag{14}$$
$$I(U; Z|W, S, V_2) \leq \Delta R. \tag{15}$$

The achievability of a given rate–distortion quadruple that satisfies these conditions is shown in [15] by random binning arguments, in the spirit of [18], but applied in an hierarchical structure of successive refinement. The details can be found in the proof of the direct part of the main theorem in [15].

Remark 1 – Properties of the achievable region: Let us denote by $\mathcal{R}$ the set of all achievable rate-distortion quadruples $(R_1, D_1, \Delta R, D_2)$. By application of the time-sharing principle, $\mathcal{R}$ is convex. By the properties of the distortion measures d_i, $i = 1, 2$, reproduction of the source with zero distortion at both decoders is possible with finite (although high) rates. It follows that if $(R_1, D_1, \Delta R, D_2) \in \mathcal{R}$, then for every $\delta > 0$ there exists $\delta' > 0$ such that $(R_1 + \delta', D_1 - \delta, \Delta R, D_2) \in \mathcal{R}$ provided that $D_1 \geq \delta$, and $(R_1, D_1, \Delta R + \delta', D_2 - \delta) \in \mathcal{R}$ provided that $D_2 \geq \delta$, and in addition, $\delta' \to 0$ as $\delta \to 0$. We will make use of this observation in Section 4 (see Remark 2 there).

3 Main Results

In this section, we give a single–letter characterization of $\mathcal{D}$ for a given degraded source $P_{UV_1V_2}$ and channels $P_{Y_1|X_1}$, $P_{Y_2|X_2}$, with bandwidth expansion factors ρ_1 and ρ_2. Define $\mathcal{D}^*$ to be the set of all distortion–cost quadruples $\{(D_1, \Gamma_1, D_2, \Gamma_2)\}$ for which there exists a triple of random variables (W, S, Z), taking values in finite sets $\mathcal{W}$, $\mathcal{S}$, and $\mathcal{Z}$, respectively, such that the following conditions are satisfied:

1. $(W, S, Z) \multimap U \multimap V_2 \multimap V_1$ is a Markov chain.

2. There exist deterministic maps $\psi_1 : \mathcal{W} \times \mathcal{V}_1 \to \hat{\mathcal{U}}$ and $\psi_2 : \mathcal{Z} \times \mathcal{V}_2 \to \hat{\mathcal{U}}$ such that

$$\mathbb{E}d_1(U, \psi_1(W, V_1)) \leq D_1 \tag{16}$$

$$\mathbb{E}d_2(U, \psi_2(Z, V_2)) \leq D_2 \tag{17}$$

3. The alphabets $\mathcal{W}$, $\mathcal{S}$, $\mathcal{Z}$ satisfy

$$|\mathcal{W}| \leq |\mathcal{U}| + 2, \tag{18}$$

$$|\mathcal{S}| \leq (|\mathcal{U}| + 1)^2, \tag{19}$$

$$|\mathcal{Z}| \leq |\mathcal{U}|(|\mathcal{U}| + 2)(|\mathcal{U}| + 1)^2 + 1. \tag{20}$$

4. The capacity–cost functions, $C_i(\Gamma_i) \stackrel{\Delta}{=} \max_{X_i:\ \mathbb{E}\phi_i(X_i)\leq\Gamma_i} I(X_i;Y_i)$, $i=1,2$, satisfy

$$
\begin{aligned}
I(U;W|V_1)+I(U;S|W,V_2) &\leq \rho_1 C_1(\Gamma_1) &(21)\\
I(U;Z|W,S,V_2) &\leq \rho_2 C_2(\Gamma_2). &(22)
\end{aligned}
$$

Our main result is the following:

Theorem 2 *For any discrete, memoryless, stochastically degraded source, and a pair of independent discrete memoryless channels with bandwidth expansion factors ρ_1 and ρ_2, $\mathcal{D}=\mathcal{D}^*$.*

The proof appears in Section 6.1.

The similarity between the characterization of the region of achievable rate–distortion quadruples of [15], as described at the end of Section 2, and the characterization of $\mathcal{D}=\mathcal{D}^*$, is self evident. In fact, the only difference is that the first–stage coding rate R_1 and the second–stage incremental rate, R_2-R_1, of the former, are replaced, in the latter, by $\rho_1C_1(\Gamma_1)$ and $\rho_2C_2(\Gamma_2)$, respectively. The immediate conclusion from this observation is that the separation principle applies to our model. In other words, given a stochastically degraded memoryless source and a pair of independent memoryless channels, joint source–channel coding with successive refinement, at the presence of SI at the decoders, can be implememted, without loss of asymptotic optimality, by first applying an optimal successive refinement source code, as in [15], regardless of the channels, and then using separate capacity–achieving channel codes for transmission over the channels.

It is interesting to examine the case of feedback. It turns out, as we mentioned in the Introduction, that feedback does not increase the distortion–cost region $\mathcal{D}$. This is in analogy to the well known result that feedback cannot increase the capacity of a memoryless channel. Formally, let us denote by $\mathcal{D}^f$ the distortion–cost region for successive refinement with feedback. Then, we have the following result.

Theorem 3 *For any discrete, memoryless, stochastically degraded source, and a pair of independent discrete memoryless channels, $\mathcal{D}^f=\mathcal{D}=\mathcal{D}^*$. That is, feedback from the output of the first–stage channel to the encoder of the second stage does not increase the distortion–cost region.*

The proof is identical to the proof of Theorem 2. In particular, the proof of converse part of Theorem 2 is general enough to allow feedback (whereas the direct part is without feedback).

An important special case is that of identical SI, i.e., when $V_1 \equiv V_2$. In this case, the achievable region admits a simpler characterization, as follows. Define $\mathcal{D}_i^*$ similarly as $\mathcal{D}^*$, where $V_1 \equiv V_2 \stackrel{\Delta}{=} V$, the bounds on the alphabets sizes (18), (19), and (20), are replaced by

$$|\mathcal{W}| \leq |\mathcal{U}| + 2, \tag{23}$$

$$|\mathcal{Z}| \leq |\mathcal{U}|(|\mathcal{U}| + 2) + 1. \tag{24}$$

and the inequalities (21), (22), are replaced by

$$I(U;W|V) \leq \rho_1 C_1(\Gamma_1) \tag{25}$$

$$I(U;Z|W,V) \leq \rho_2 C_2(\Gamma_2). \tag{26}$$

Defining now $\mathcal{D}$ (resp. $\mathcal{D}_i^f$) as $\mathcal{D}$ (resp. $\mathcal{D}^f$) with the restriction of identical SI, we have the following corollary to Theorem 3.

Corollary 1 *For any discrete memoryless joint source (U,V) and a pair of independent discrete memoryless channels, $\mathcal{D}_i = \mathcal{D}_i^f = \mathcal{D}_i^*$.*

Proof. In view of Theorem 3, we have to show that when $V_1 \equiv V_2 \equiv V$, the characterization of $\mathcal{D}^*$ reduces to that of $\mathcal{D}_i^*$. Indeed, in that case, by (21), (22), we can write

$$\rho_1 C_1(\Gamma_1) \geq I(U;W|V) + I(U;S|W,V) = I(U;W,S|V) \tag{27}$$

$$\rho_2 C_2(\Gamma_2) \geq I(U;Z|W,S,V). \tag{28}$$

Observe that W and S appear in the mutual information functions of (27) and (28) always together. The functions ψ_1 and ψ_2, by definition, do not depend on S. Thus, we can define a new auxiliary random variable $W' = (W,S)$, and a new function $\psi_1'(W',V) = \psi_1(W,V)$, without altering the distortions or violating the inequalities (21) or (22). This proves the corollary with the only exception that the alphabet size of Z is given by (20) instead of (24), and that of W' is given by $|\mathcal{W}'| = |\mathcal{W}| \cdot |\mathcal{S}|$ instead of the right hand side of (23). The fact that the alphabet sizes can be restricted to (23) and (24), follows from Carathéodory's theorem. The details are omitted. $\square$

Although feedback does not improve performance in terms of achievable distortion–cost trade-offs, it turns out that feedback can significantly reduce the complexity of optimum (or nearly optimum) communication schemes. We next give two examples of situations where the presence of feedback enables to achieve optimal performance with the use of simple single–letter codes ($n = 1$) when $\rho_1 = \rho_2 = 1$, as an alternative to separate source and channel coding in long blocks. Without feedback, in the first example (binary–Hamming), it will be obviously impossible to achieve optimality using single–letter codes, whereas in the second example, which is Gaussian–quadratic, it will become evident from the more general theory developed in Section 5. The common idea in both examples is to use the second stage in order to transmit the estimation error of the first stage (as described in the Introduction). In both examples, there is no SI correlated to the source which is available to the decoders, and the distortion measures at both stages are identical, i.e., $d_1 = d_2$. Finally, in both examples, the optimality is in the sense that the joint source–channel bound is attained at both stages at the same time, i.e., $R(D_1) = C_1(\Gamma_1)$ and $R(D_2) = C_1(\Gamma_1) + C_2(\Gamma_2)$.

Example 1 (Binary–Hamming). Let U be the binary symmetric source (BSS) with $\mathcal{U} = \hat{\mathcal{U}} = \{0, 1\}$, and let the first channel be the binary symmetric channel (BSC) with crossover probability $p < 1/2$, input and output alphabets $\mathcal{X}_1 = \mathcal{Y}_1 = \{0, 1\}$, and no transmission constraint ($\Gamma_1 = \infty$). Let the second channel be a clean binary channel, with $\mathcal{X}_2 = \mathcal{Y}_2 = \{0, 1\}$, and with a transmission cost constraint $\mathbb{E}X_2 \leq \Gamma_2$, where $p \leq \Gamma_2 < 0.5$. Let $d_1 = d_2$ be the Hamming distortion measure. Consider the following single–letter code with feedback: The first stage encoder is $X_1 = f_1(U) = U$ (no coding) and the first stage decoder is $\hat{U}_1 = g_1(Y_1) = Y_1$. This obviously achieves Hamming distortion $D_1 = p$ which is well–known to be optimum as $R(D_1) = 1 - h(D_1) = 1 - h(p) = C_1(\infty)$. As for the second stage, the encoder would be $X_2 = f_2(U, Y_1) = U \oplus Y_1$, where $\oplus$ denotes addition modulo 2. This encoder satifies the power constraint since $\mathbb{E}X_2 = \Pr\{Y_1 \neq U\} = p \leq \Gamma_2$, and the decoder is $\hat{U}_2 = g_2(Y_1, Y_2) = Y_1 \oplus Y_2$. Since

$$Y_1 \oplus Y_2 = Y_1 \oplus X_2 = Y_1 \oplus U \oplus Y_1 = U,$$

the distortion at the second stage is $D_2 = 0$. Notably, for $p = \Gamma_2$ the system is, moreover, optimal in the sense that there is no 'waste' on unused channel resources since $R(D_2 = 0) = 1 = [1 - h(p)] + h(p) = C_1(\infty) + C_2(\Gamma_2)$. Note that due to the power constraint, a single–letter code without feedback cannot be optimal: Since both X_2 and U are binary, there are only four possible functions $X_2 = f_2(U)$, two of which cause information loss ($f_2(u) \equiv 0$ and $f_2(u) \equiv 1$) and the other two violate the transmission cost constraint since $\Gamma_2 < 0.5$.

Example 2 (Gaussian–Quadratic). Let $U \sim \mathcal{N}(0, \sigma_U^2)$, $N_1 \sim \mathcal{N}(0, \sigma_1^2)$, and $N_2 \sim \mathcal{N}(0, \sigma_2^2)$ be independent, where U respresents the source, N_1 – the additive noise of the first channel, $Y_1 = X_1 + N_1$, and N_2 – the additive noise of the second channel, $Y_2 = X_2 + N_2$. The model is depicted in Figure 2.

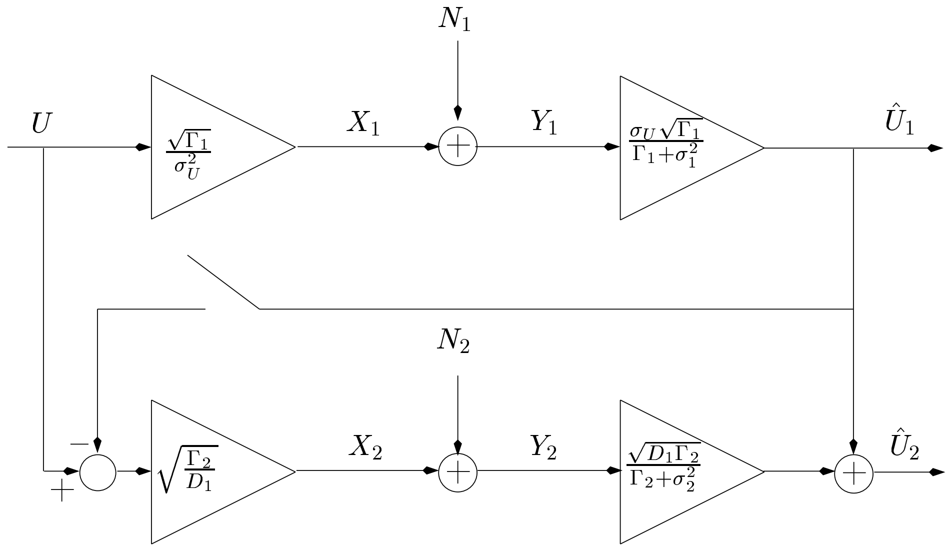

Figure 2: Hierarchical joint source-channel coding for the Gaussian channel.

Both channels are subjected to power constraints $\mathbb{E}X_i^2 \leq \Gamma_i$, $i = 1, 2$, and the distortion measure at both stages is quadratic. The first–stage encoder is

$$X_1 = f_1(U) = \frac{\sqrt{\Gamma_1}}{\sigma_U} \cdot U \tag{29}$$

and the first–stage decoder is

$$\hat{U}_1 = \frac{\sigma_U\sqrt{\Gamma_1}}{\Gamma_1 + \sigma_1^2} \cdot Y_1. \tag{30}$$

The resulting distortion level is

$$D_1 = \mathbb{E}(U - \hat{U}_1)^2 = \frac{\sigma_U^2}{1 + \Gamma_1/\sigma_1^2} \tag{31}$$

which satisfies, as is well–known, the equation:

$$R(D_1) = \frac{1}{2}\log\frac{\sigma_U^2}{D_1} = \frac{1}{2}\log\left(1 + \frac{\Gamma_1}{\sigma_1^2}\right) = C_1(\Gamma_1). \tag{32}$$

As for the second stage, let $X_2 = \alpha E$ where $E = U - \hat{U}_1$ and where α is chosen to match the second power constraint, i.e., $\alpha = \sqrt{\Gamma_2 / D_1}$. Let the second–stage decoder be defined as

$$\hat{U}_2 = \hat{U}_1 + \hat{E} \stackrel{\Delta}{=} \hat{U}_1 + \beta Y_2, \tag{33}$$

where

$$\beta = \frac{\sqrt{D_1 \Gamma_2}}{\Gamma_2 + \sigma_2^2}. \tag{34}$$

The resulting distortion is

$$\begin{aligned} D_2 &= \mathbb{E}(U - \hat{U}_2)^2 \\ &= \mathbb{E}(E + \hat{U}_1 - \hat{E} - \hat{U}_1)^2 \\ &= \mathbb{E}(E - \hat{E})^2 \\ &= \frac{D_1}{1 + \Gamma_2/\sigma_2^2} \\ &= \frac{\sigma_U^2}{(1 + \Gamma_1/\sigma_1^2)(1 + \Gamma_2/\sigma_2^2)} \end{aligned} \tag{35}$$

which satisfies the equation

$$\begin{aligned} R(D_2) &= \frac{1}{2} \log \frac{\sigma_U^2}{D_2} \\ &= \frac{1}{2} \log \left(1 + \frac{\Gamma_1}{\sigma_1^2}\right) + \frac{1}{2} \log \left(1 + \frac{\Gamma_2}{\sigma_2^2}\right) \\ &= C_1(\Gamma_1) + C_2(\Gamma_2). \end{aligned} \tag{36}$$

These two simple examples motivate us to investigate single–letter codes in more generality. The remaining part of this paper is dedicated to this goal. Necessary and sufficient conditions for the optimality of single–letter codes will be established in Section 5 (with and without feedback) in the spirit of the techniques of [7] (cf. also [4]). In particular, the idea of using the second stage for transmission of error signal of the first stage (when feedback is present) will be examined more generally.

4 Notions of Optimality

This section is devoted to a preparatory step before we can turn to derive conditions for optimality of single–letter codes in the next section. In particular, we give a formal definition of *optimaility* of a rate–distortion quadruple (R_1, D_1, R_2, D_2), or more precisely, a quadruple $(R_1, D_1, \Delta R, D_2)$, where $\Delta R = R_2 - R_1$, for the pure source coding problem of [15], and a similar formal definition of optimality of a distortion–cost quadruple $(D_1, \Gamma_1, D_2, \Gamma_2)$. We then characterize the relationship between them. Generally speaking, this characterization is related to, and strongly based upon the separation principle that we have shown in Section 3: Given a distortion–cost quadruple $(D_1, \Gamma_1, D_2, \Gamma_2) \in \mathcal{D}^*$, one can always find two numbers R_1 and ΔR such that

$$I(U;W|V_1) + I(U;S|W_1,V_2) \leq R_1 \leq \rho_1 C_1(\Gamma_1) \quad (37)$$

$$I(U;Z|W,S,V_2) \leq \Delta R \leq \rho_2 C_2(\Gamma_2). \quad (38)$$

A good combination of a source encoder and a channel encoder is characterized by tightness of these inequalities. Obviously, tightness of (at least one of) the above two left–hand inequalities (of upper bounding the mutual information expressions by R_1 and ΔR) are necessary conditions for tightness of the corresponding inequalities in the definition of $\mathcal{D}^*$ (of upper bounding the same expressions by $\rho_1 C_1(\Gamma_1)$ and $\rho_2 C_2(\Gamma_2)$, respectively). The additional condition needed is, of course, that R_1 and/or ΔR would be close to $\rho_1 C_1(\Gamma_1)$ and $\rho_2 C_2(\Gamma_2)$, respectively. This simple observation will help us to decompose the problem of optimality to a subproblem related to the source and a subproblem related to the channels, i.e., to handle separately conditions for optimality w.r.t. the source and the distortion measures, on the one hand, and optimality w.r.t. the channels and the transmission cost functions, on the other hand (similarly as in [7]).

As mentioned in Section 2, in [15], it is shown that for the pure source coding problem, a rate–distortion quadruple $(R_1, D_1, \Delta R, D_2)$ is achievable if and only if there exist random variables (W, S, Z) and a pair of mappings (ψ_1, ψ_2), such that $(W,S,Z) \multimap U \multimap V_2 \multimap V_1$ is a Markov chain, and the distortions and rates satisfy

$$\begin{aligned} \mathbb{E} d_1(U, \psi_1(W, V_1)) &\leq D_1 \\ \mathbb{E} d_2(U, \psi_2(Z, V_2)) &\leq D_2 \\ I(U;W|V_1) + I(U;S|W,V_2) &\leq R_1 \\ I(U;Z|W,S,V_2) &\leq \Delta R. \end{aligned}$$

We proceed to the definition of optimality of a rate–distortion quadruple.

Definition 3 *An achievable rate–distortion quadruple* $(R_1, D_1, \Delta R, D_2) \in \mathcal{R}$ *is said to be* optimal *if* $(R_1-\delta_1, D_1-\delta_2, \Delta R-\delta_3, D_2-\delta_4) \notin \mathcal{R}$ *whenever* δ_i, $i = 1, 2, 3, 4$, *are all non–negative and at least one of them is strictly positive. It is said to be* distortion-optimal *if* $(R_1, D_1 - \delta_1, \Delta R, D_2 - \delta_2) \notin \mathcal{R}$ *whenever* δ_i, $i = 1, 2$, *are non-negative and at least one of them is strictly positive.*

Loosely speaking, optimality means that $(R_1, D_1, \Delta R, D_2)$ lies on the boundary of the achievable region in at least one of the four dimensions of the space of quadruples. In simple words, any improvement in one of the components of this vector must come at the expense of degradation in some other component.

Remark 2 – Properties of optimal rate-distortion quadruples: We will explore here some properties of optimal points, in particular, the relations between optimal and distortion-optimal quadruples will be highlighted. The following can be deduced from the properties of $\mathcal{R}$ (see Remark 1). Let

$$(R_1, D_1, \Delta R, D_2) \in \mathcal{R}. \tag{39}$$

If $D_1 > 0$, and

$$(R_1, D_1 - \delta, \Delta R, D_2) \notin \mathcal{R} \quad \forall \delta > 0, \tag{40}$$

then necessarily

$$(R_1 - \delta, D_1, \Delta R, D_2) \notin \mathcal{R} \quad \forall \delta > 0. \tag{41}$$

To see this, assume that (40) holds, yet

$$(R_1 - \delta_1, D_1, \Delta R, D_2) \in \mathcal{R} \tag{42}$$

for some $\delta_1 > 0$. By the properties of $\mathcal{R}$ (see Remark 1), for any $\delta_2 \in (0, D_1)$

$$(R_1 - \delta_1 + \delta_2', D_1 - \delta_2, \Delta R, D_2) \in \mathcal{R}, \tag{43}$$

where $\delta_2' \to 0$ as $\delta_2 \to 0$. Thus, choose δ_2 small enough so that

$$R_1 - \delta_1 + \delta_2' \leq R_1. \tag{44}$$

Then, a convex combination of the left hand side of (43) with the left hand side of (39), with weights λ and $1 - \lambda$, resp., yields that

$$(R_1, D_1 - \lambda\delta_2, \Delta R, D_2) \in \mathcal{R} \tag{45}$$

for some $\lambda > 0$, contradicting (40). A similar conclusion holds for the second stage, with the differential rate: if $D_2 > 0$ and $(R_1, D_1, \Delta R, D_2 - \delta) \notin \mathcal{R}$ for all $\delta > 0$, then necessarily $(R_1, D_1, \Delta R - \delta, D_2) \notin \mathcal{R}$ for all $\delta > 0$. In view of the above discussion, we conclude that an achievable rate-distortion quadruple with strictly positive distortions is optimal if and only if it is distortion-optimal.

We proceed to the definition of optimality of a distortion–cost quadruple, which is in the same spirit.

Definition 4 *A distortion–cost quadruple* $(D_1, \Gamma_1, D_2, \Gamma_2) \in \mathcal{D}$ *is said to be* optimal *if* $(D_1 - \delta_1, \Gamma_1 - \delta_2, D_2 - \delta_3, \Gamma_2 - \delta_4) \notin \mathcal{D}$ *whenever* δ_i, $i = 1, 2, 3, 4$, *are all non-negative and at least one of them is strictly positive. A code* (f_1, g_1, f_2, g_2) *is optimal if it incurs an optimal distortion–cost quadruple.*

The next lemma gives necessary and sufficient conditions for a quadruple of strictly positive distortions and costs to be optimal. Note that since $\mathcal{D} = \mathcal{D}^f$ (Theorem 3), Lemma 1 below holds in the presence, as well as in the absence, of feedback.

Lemma 1 *A quadruple* $(D_1, \Gamma_1, D_2, \Gamma_2)$ *with* $\Gamma_i > 0$ *and* $D_i > 0$, $i = 1, 2$, *is optimal if and only if the following three conditions hold:*

1. *There exist random variables* (W, S, Z) *and mappings* $\psi_1 : \mathcal{W} \times \mathcal{V}_1 \to \hat{\mathcal{U}}$, $\psi_2 : \mathcal{Z} \times \mathcal{V}_2 \to \hat{\mathcal{U}}$, *such that the conditions defining* $\mathcal{D}^*$ *hold, with the distortion inequalities and the mutual information inequalities all holding with equalities.*
2. *The rate–distortion quadruple*

$$(I(U; W|V_1) + I(U; S|W, V_2), D_1, I(U; Z|W, S, V_2), D_2)$$

is optimal (in the sense of Definition 3).

3. $\Gamma_i' < \Gamma_i$ *implies* $C_i(\Gamma_i') < C_i(\Gamma_i)$, $i = 1, 2$.

Proof. We start with the necessity part. Let an optimal distortion–cost quadruple $(D_1, \Gamma_1, D_2, \Gamma_2) \in \mathcal{D}$ be given. By Theorem 3, there must exist random variables (W, S, Z) and functions ψ_i, $i = 1, 2$, satisfying the conditions in the definition of $\mathcal{D}$, namely, $(W, S, Z) \multimap U \multimap V_2 \multimap V_1$ is Markov, and eqs. (16) to (22) hold. For these random variables, the Markovity condition of the theorem clearly holds. Assume, conversely, first that condition 3 of the lemma does not hold, i.e., Γ_1 (or Γ_2) can be reduced without reducing C_1 (or C_2). Thus, Theorem 3 implies that, by separately coding for the source and the channels, we can reduce Γ_1 (Γ_2) without altering the distortions, contradicting the assumption that $(D_1, \Gamma_1, D_2, \Gamma_2)$ is optimal. Hence condition 3 is necessary. Assume next that condition 2 of the lemma is violated. In particular, say, that $(I(U; W|V_1) + I(U; S|W, V_2) - \delta, D_1, I(U; Z|W, S, V_2), D_2)$ is achievable for some $\delta > 0$. Then by the results of [15], we can separately code the for source (using a successive–refinement code for source, as in [15]) and the channels, with smaller Γ_1, contradicting, again, the optimality of $(D_1, \Gamma_1, D_2, \Gamma_2)$, where we have used the continuity of $C_1(\cdot)$, and the fact that $\Gamma_1 > 0$. Similar considerations show that

$$(I(U; W|V_1) + I(U; S|W, V_2), D_1, I(U; Z|W, S, V_2) - \delta, D_2) \tag{46}$$

$$(I(U; W|V_1) + I(U; S|W, V_2), D_1 - \delta, I(U; Z|W, S, V_2), D_2) \tag{47}$$

$$(I(U; W|V_1) + I(U; S|W, V_2), D_1, I(U; Z|W, S, V_2), D_2 - \delta) \tag{48}$$

are inachievable for any $\delta > 0$. Hence condition 2 of the lemma is necessary. It remains to prove necessity of condition 1. By Theorem 2, these equations should hold with inequality sign '$\leq$'. Assume that

$$\mathbb{E}d(U, \psi_1(W, V_1)) = D_1' < D_1. \tag{49}$$

Then, by Theorem 3, the quadruple $(D_1', \Gamma_1, D_2, \Gamma_2)$ is achievable, again in contradiction to the assumption that $(D_1, \Gamma_1, D_2, \Gamma_2)$ is optimal. The necessity of equalities in the other distortion inequality and in the mutual information inequalities are proved similarly.

Turning now to the sufficiency part, let $(D_1, \Gamma_1, D_2, \Gamma_2)$ satisfy the conditions of Lemma 1. Assume, conversely, that it is not optimal. Thus,

$$(D_1 - \delta_1, \Gamma_1 - \delta_2, D_2 - \delta_3, \Gamma_2 - \delta_4) \in \mathcal{D} \tag{50}$$

for some non–negative δ_i, $i = 1, 2, 3, 4$, at least one of which is strictly positive. By Theorem 3, there exist random variables (W', S', Z'), and maps ψ_1', ψ_2', such that

$$\begin{aligned}
\mathbb{E}d_1(U, \psi_1'(W', V_1)) &\leq D_1 - \delta_1 && (51)\\
\mathbb{E}d_2(U, \psi_2'(Z', V_2)) &\leq D_2 - \delta_3 && (52)\\
I(U; W'|V_1) + I(U; S'|Z', V_2) &\leq \rho_1 C_1(\Gamma_1 - \delta_2) && (53)\\
I(U; Z'|W', S', V_2) &\leq \rho_2 C_2(\Gamma_2 - \delta_4), && (54)
\end{aligned}$$

implying that the rate–disotrtion quadruple

$$(I(U;W|V_1)+I(U;S|W,V_2),D_1,I(U;Z|W,S,V_2),D_2)$$

is not optimal, contradicting condition 2 of Lemma 1. $\square$

In the case of identical SI, $V_1=V_2=V$, the conditions of Lemma 1 can be simplified. We state these as a corollary.

Corollary 2 *In the case of identical SI, a quadruple* $(D_1,\Gamma_1,D_2,\Gamma_2)$*, with* $\Gamma_i>0$ *and* $D_i>0$*,* $i=1,2$*, is optimal if and only if there exist random variables* (W,Z) *and mappings* $\psi_1:\mathcal{W}\times\mathcal{V}\to\hat{\mathcal{U}}$ *and* $\psi_2:\mathcal{Z}\times\mathcal{V}\to\hat{\mathcal{U}}$*, such that the following conditions simultaneously hold:*

1. $(W,Z)\multimap U\multimap V$ *form a Markov chain.*
2. $\mathbb{E}d_1(U,\psi_1(W,V))=D_1$, $\mathbb{E}d_2(U,\psi_2(Z,V))=D_2$,
3. *(a)* $I(U;W|V)=\rho_1C_1(\Gamma_1)$
   (b) $I(U;Z|W,V)=\rho_2C_2(\Gamma_2)$
4. *The rate–distortion quadruple* $(I(U;W|V),D_1,I(U;Z|W,V),D_2)$ *is optimal.*
5. $\Gamma_i'<\Gamma_i$ *implies* $C_i(\Gamma_i')<C_i(\Gamma_i)$*,* $i=1,2$.

The proof follows the lines of the proof of Lemma 1 (making use of Corollary 1 instead of Theorem 3), and is therefore omitted.

5 Single–Letter Codes

Equipped with the results of Sections 3 and 4, we are now ready to turn to the derivation of necessary and sufficient conditions for the optimality of single–letter joint source–channel codes (namely, codes of block length $n=1$) for channels with bandwidth expansion factors $\rho_1=\rho_2=1$, in the presence and in the absence of feedback. For the sake of simplicity, we assume, in this section, identical SI, i.e., $V_1=V_2=V$, and, of course, $\mathcal{V}_1=\mathcal{V}_2=\mathcal{V}$.

The outline of the derivations in this section is as follows. Lemma 2 below gives generic necessary and sufficient conditions for the optimality of a given single–letter code in terms of the random variables induced by this single–letter code. In Lemmas 3 and 4 that follow, these conditions are translated into equivalent, but more explicit conditions that can be checked relatively easily. In particular, Lemma 3 focuses on the relation between capacity–achieving channel inputs and the form of the transmission cost functions, whereas Lemma 4 is, analogously, about the relation between optimum test channels and the form of the distortion measures. Lemma 5 then fills in the missing link of conditions for equality between the relevant source–related mutual informations ("source coding rates") and the corresponding channel–related mutual informations (capacities), so as to make the corresponding data processing inequalities tight. Finally, Theorem 4, which is the main result of this section, puts it all together and gives the full set of explicit necessary and sufficient conditions for the optimality of a given single–letter code. An immediate consequence of Theorem 4, stated in Corollary 3, is that under the quadratic distortion measure, and in the absence of SI, feedback is necessary to achieve optimality with single-letter codes, provided that the distortion at the first stage, D_1, is less than the a priori variance of the source. This conclusion holds in full generality - i.e., it is not restricted to Gaussian sources, nor to additive channels. The proofs of Lemma 2 and Lemma 4 appear in Subsections 6.2 and 6.3, respectively, and the proofs of Theorem 4 and Corollary 3 are provided in this section. The rest are omitted.

Lemma 2 *Let a single–letter code* (f_1, g_1, f_2, g_2)*, incurring a distortion–cost quadruple* $(D_1, \Gamma_1, D_2, \Gamma_2)$*,* $D_i > 0$*,* $\Gamma_i > 0$*,* $i = 1, 2$*, be given. There exists random variables* (W, Z) *and functions* (ψ_1, ψ_2)*, satisfying conditions 1–4 of Corollary 2, if and only if the following conditions are simultaneously satisfied:*

1. *For* $i = 1, 2$*,* $I(X_i; Y_i) \equiv I(f_i(U); Y_i) = C_i(\Gamma_i)$*. If feedback is included, then for* $i = 2$*, we set* $X_2 = f_2(U, Y_1)$*.*
2. *The rate–distortion quadruple* $(I(U; Y_1|V), D_1, I(U; Y_2|Y_1, V), D_2)$ *is optimal.*
3. *(a)* $I(U; Y_1|V) = I(X_1; Y_1)$
   (b) $I(U; Y_1, Y_2|Y_1, V) = I(X_2; Y_2)$

As explained earlier, and following the methodology of [7], we now proceed to decompose the requirements of Lemma 2 into separate, but more explicit conditions regarding the transmission cost functions and the distortion measures. The next result was stated and proved in [7].

Lemma 3 *For a given source distribution P_U, single–letter encoders f_1, f_2, and conditional distributions $P_{Y_1|X_1}$, $P_{Y_2|X_2}$:*

1. *For $i = 1, 2$, if $I(X_i; Y_i) \equiv I(f_i(U); Y_i) < C_i(\infty)$, then $I(f_i(U); Y_i) = C_i(\Gamma_i)$ if and only if for all $x \in \mathcal{X}$:*

$$\phi_i(x) \begin{cases} = c_i D(P_{Y_i|X_i}(\cdot|x) \| P_{Y_i}) + \phi_{0,i} & \text{if } P_{X_i}(x) > 0, \\ \geq c_i D(P_{Y_i|X_i}(\cdot|x) \| P_{Y_i}) + \phi_{0,i} & \text{otherwise} \end{cases} \tag{55}$$

where $c_i > 0$ and $\phi_{0,i}$ are constants.

2. *For $i = 1, 2$, if $I(X_i; Y_i) \equiv I(f_i(U); Y_i) = C_i(\infty)$, then $I(f_i(U); Y_i) = C_i(\Gamma_i)$ for any function ϕ_i.*

As the channels in our model are independent, each one is treated separately, exactly as the single–user channel model of [7]. In particular, note that Lemma 3 is exactly [7, Lemma 3]. Thus, the proof is omitted.

Analogously to Lemma 3, we next state the conditions under which the distortion measures "match" the source, channels, and code (f_1, g_1, f_2, g_2), yielding an optimal coding scheme. In contrast to the conditions on the transmission cost functions ϕ_i, $i = 1, 2$, stated in Lemma 3, the conditions on the distortion measures are not the same as those stated in [7, Lemma 4]. Moreover, the condition on d_2 depends on whether or not feedback is present.

Lemma 4 *Let a source distribution P_{UV}, channel conditional distributions $P_{Y_1|X_1}$, $P_{Y_2|X_2}$, and a single–letter code (f_1, g_1, f_2, g_2), be given. Assume that $I(U; Y_1|V) > 0$, $I(U; Y_2|Y_1, V) > 0$, and $D_i > 0$, $i = 1, 2$.*

1. *In the absence of feedback, condition 2 of Lemma 2 is satisfied if and only if the distortion measures satisfy, for every $u \in \mathcal{U}$, $y_1 \in \mathcal{Y}_1$, and $y_2 \in \mathcal{Y}_2$:*

$$\begin{aligned} \mathbb{E}\left\{d_1(u, g_1(y_1, V)) | U = u\right\} &= \lambda_1 \mathbb{E}\left\{\log P_{Y_1|V}(y_1|V) | U = u\right\} \\ &\quad - \lambda_1 \log P_{Y_1|U}(y_1|u) + k_1(u), \end{aligned} \tag{56}$$

$$\begin{aligned} \mathbb{E}\left\{d_2(u, g_2(y_1, y_2, V)) | U = u\right\} &= \lambda_2 \mathbb{E}\left\{\log P_{Y_2|Y_1,V}(y_2|y_1, V) | U = u\right\} \\ &\quad - \lambda_2 \log P_{Y_2|U}(y_2|u) + k_2(u) \end{aligned} \tag{57}$$

for some positive constants λ_1, λ_2, and functions $k_1(u)$, $k_2(u)$, where the expectations are taken w.r.t. $P_{V|U=u}$.

2. If the presence of feedback, condition 2 of Lemma 2 is satisfied if and only if d_1 satisfies (56), and d_2 satisfies

$$\begin{aligned}\mathbb{E}\left\{d_2(u, g_2(y_1, y_2, V))|U = u\right\} &= \lambda_2 \mathbb{E}\left\{\log P_{Y_2|Y_1,V}(y_2|y_1, V)|U = u\right\} \\ &\quad -\lambda_2 \log P_{Y_2|Y_1,U}(y_2|y_1, u) + k_2(y_1, u) \qquad (58)\end{aligned}$$

for some positive constant λ_2 and a function $k_2(y_1, u)$, where, again, expectations are taken w.r.t. $P_{V|U=u}$.

As a final step before stating our main result, we characterize the situations under which condition 3 of Lemma 2 holds.

Lemma 5 *Condition 3 of Lemma 2 holds if and only if the random variables Y_1, Y_2, and V are independent and the encoding functions f_1 and f_2 are information lossless, that is, if and only if, in the absence of feedback*

$$\begin{aligned}P_{Y_1,Y_2,V} &= P_{Y_1}P_{Y_2}P_V &(59)\\ I(U;Y_1) &= I(X_1;Y_1) &(60)\\ I(U;Y_2) &= I(X_2;Y_2), &(61)\end{aligned}$$

and, in the presence of feedback, (61) is replaced by

$$I(U,Y_1;Y_2) = I(X_2;Y_2). \qquad (62)$$

The proof of Lemma 5 follows by simple applications of the chain rule for mutual information, and is omitted.

We are now in a position to state the main result concerning single–letter source–channel codes for successive refinement with SI. For simplicity of the exposition, we assume here that condition 5 of Corollary 2 holds, that is, the transmission costs cannot be reduced without decreasing the capacities of the channels. We also make the assumptions that $I(U;Y_1|V)$, $I(U;Y_2|Y_1,V)$, and the distortions at both stages, are strictly positive.

Theorem 4 *Assume that the following hold:*

- *Assumption 1: Γ_i cannot be reduced without reducing $C_i(\Gamma_i)$, and $I(X_i;Y_i) < C_i(\infty)$, $i = 1, 2$.*

- *Assumption 2:* $I(U;Y_1|V)$, $I(U;Y_2|Y_1,V)$, *and* D_i, $i=1,2$, *are strictly positive.*

Then, a single–letter code is optimal if and only if the following conditions simultaneously hold:

1. *The random variables* Y_1, Y_2, *and* V *are independent.*

2. *The encoding functions* f_1 *and* f_2 *are information lossless, i.e.,*

$$I(X_1;Y_1) = I(U;Y_1) \tag{63}$$

and

$$I(X_2;Y_2) = \begin{cases} I(U;Y_2) & \textit{when feedback is absent} \\ I(U,Y_1;Y_2) & \textit{when feedback is present} \end{cases} \tag{64}$$

3. *The transmission cost functions satisfy*

$$\phi_i(x) \begin{cases} = c_i D(P_{Y_i|X_i}(\cdot|x)\|P_{Y_i}) + \phi_{0,i} & \textit{if } \ P_{X_i}(x) > 0, \\ \geq c_i D(P_{Y_i|X_i}(\cdot|x)\|P_{Y_i}) + \phi_{0,i} & \textit{otherwise} \end{cases} \tag{65}$$

for $i=1,2$, *where* $c_i > 0$ *and* $\phi_{0,i}$ *are constants.*

4. *The distortion measures satisfy*

$$\mathbb{E}\left\{d_1(u, g_1(y_1, V))|U=u\right\} = \lambda_1 \log P_{Y_1}(y_1) - \lambda_1 \log P_{Y_1|U}(y_1|u) + k_1(u), \tag{66}$$

and

$$\mathbb{E}\left\{d_2(u, g_2(y_1, y_2, V))|U=u\right\} = \begin{cases} \lambda_2 \log \frac{P_{Y_2}(y_2)}{P_{Y_2|U}(y_2|u)} + k_2(u) & \textit{without feedback} \\ \lambda_2 \log \frac{P_{Y_2}(y_2)}{P_{Y_2|Y_1,U}(y_2|y_1,u)} + k_2(y_1,u) & \textit{with feedback} \end{cases} \tag{67}$$

for some positive constants λ_1, λ_2, *and functions* k_1, k_2.

Proof. The proof is a combination of Corollary 2, and Lemmas 2, 3, 4, and 5. In view of Assumption 1 of Theorem 4, we only have to guarantee the satisfaction of the conditions stated in Lemma 2. Thus:

- First condition of Lemma 2 – due to Lemma 3, this is satisfied by (65) above.

- Third condition of Lemma 2 – due to Lemma 5, this amounts to Conditions 1 and 2 of Theorem 4.

- Second condition of Lemma 2 – is satisfied due to the independence of Y_1, Y_2, and V, and Lemma 4.

This completes the proof of the Theorem. $\square$

This result provides a tool to examine optimality of successive refinement schemes, operating over noisy channels, possibly with the presence of SI correlated with the source, at the decoders. It is easy to verify, that Example 2 of Section 3 satisfy the conditions of Theorem 4. In addition, it enables us to examine when a feedback channel is necessary in order to achieve optimality with relatively simple, single-letter codes. In this context, let us return now to discuss the Gaussian example that was given in Section 3, and to re-examine it in view of the results stated in Theorem 4.

Example 2 (cont'd). We now show that feedback is necessary for the given system to be optimal. That is, while the (asymptotically) achievable region $\mathcal{D}$ of distortion-cost quadruples is independent of whether feedback is present, it cannot be achieved using single-letter codes without feedback. To see this, observe that the right hand side of (67) for the case of no feedback, does not depend on y_1. On the other hand, the left hand side there may depend on y_1, via the decoding function of the second stage, $g_2(y_1, y_2)$ (no side information is present here, therefore we dropped the dependence on v, and the conditional expectation). Therefore, if the second stage decoder g_2 really utilizes y_1, equation (67) cannot hold for the case of no feedback. Note that the linear decoder of the second stage in the Gaussian-quadratic example makes use of the output of the first channel, y_1. We can further examine whether a choice of another decoder facilitates optimality without feedback. Note, however, that once the encoder of the first stage f_1 is linear, and the noise random variables N_1, N_2, are independent, the optimal second stage decoder g_2 makes use of Y_1, and therefore feedback is necessary for optimality. This is because Y_1 is an additional observation on U with noise N_2 that is independent of the noise N_1 of Y_1.

Clearly, this observation goes beyond the Gaussian-quadratic regime with linear encoders. Specifically, we claim that under the quadratic distortion measure, feedback is necessary to achieve optimality with single-letter codes, whenever the distortion of the first stage, D_1, is strictly less than the variance of the source U. This is the subject of the next corollary. Note that it is not restricted to the Gaussian model, nor to additive channels.

Corollary 3 *Let U be an arbitrary zero mean random variable with variance σ_U^2, and $P_{Y_1|X_1}$, $P_{Y_2|X_2}$, be arbitrary channels. In the absence of SI and under quadratic distortion measure, feedback* *is necessary to achieve optimality with single-letter codes, whenever* $D_1 < \sigma_U^2$.

Proof. Let $g_1(Y_1)$ be the estimator of U at the output of the first stage. The mean square error of this estimator is given by

$$D_1 = \mathbb{E}[(U - g_1(Y_1))^2] = \mathbb{E}(U^2) + \mathbb{E}g_1^2(Y_1) - 2\mathbb{E}[Ug_1(Y_1)], \tag{68}$$

therefore we must have

$$\mathbb{E}[Ug_1(Y_1)] \neq 0, \tag{69}$$

as otherwise, choosing $g_1(Y_1) \equiv 0$ will yield a better estimator, with MSE equal to σ_U^2. Next, observe that $\mathbb{E}(U|Y_2)$ is the best estimator that the second encoder can produce without using Y_1. Thus, in view of the discussion in Example 2 above, it is enough to show that

$$\arg\min_{\alpha} \mathbb{E}(U - \mathbb{E}(U|Y_2) - \alpha \cdot g_1(Y_1))^2 \neq 0 \qquad \text{whenever } \mathbb{E}[Ug_1(Y_1)] \neq 0, \tag{70}$$

as this implies that the use of Y_1 improves on the best estimation that is based on Y_2 only. Hence the best estimator g_2 must depend on y_1, the left hand side of (67) depends on y_1, implying, in turn, that (67) cannot hold for the case of no feedback.

Differentiating (70) with respect to α and equating to 0, we obtain for the optimal α

$$\alpha^* = \frac{\mathbb{E}(Ug_1(Y_1)) - \mathbb{E}\left\{g_1(Y_1)\mathbb{E}(U|Y_2)\right\}}{\mathbb{E}(g_1^2(Y_1))}. \tag{71}$$

However

$$\mathbb{E}\left[g_1(Y_1)\mathbb{E}(U|Y_2)\right] = \mathbb{E}\left[\mathbb{E}\left\{g_1(Y_1)\mathbb{E}(U|Y_2)|Y_2\right\}\right] = \mathbb{E}\left[\mathbb{E}(g_1(Y_1)|Y_2)\mathbb{E}(U|Y_2)\right] = 0, \tag{72}$$

where we have used, in the last equality, the fact that in optimal systems Y_1 and Y_2 are independent (first condition of Theorem 4), and that

$$\mathbb{E}g_1(Y_1) = \mathbb{E}U = 0, \tag{73}$$

as otherwise, distortion can be reduced by subtracting a constant from $g_1(y_1)$. Hence the optimal α is

$$\alpha^* = \frac{\mathbb{E}(Ug_1(Y_1))}{\mathbb{E}(g_1^2(Y_1))}, \tag{74}$$

yielding (70). □

To summarize, optimality cannot be achieved without feedback, whenever the encoders $f_1(U)$ and $f_2(U)$ are such, that the optimal second stage decoder $g_2(Y_1, Y_2)$ makes use of Y_1.

6 Proofs

6.1 Proof of Theorem 2

The direct part follows by applying Wyner–Ziv successive rate–distortion coding, as in [15, Theorem 1], followed by classical, single–user channel coding for each stage (i.e., separate source and channel coding, without feedback). Proceeding to the converse part, consider first stage no. 1, for which we can write the chain of inequalities:

$$
\begin{aligned}
\rho_1 n C_1(\Gamma_1) &\geq I(\boldsymbol{X}_1;\boldsymbol{Y}_1) \\
&\geq I(\boldsymbol{U};\boldsymbol{Y}_1) \\
&= I(\boldsymbol{U},\boldsymbol{V}_1;\boldsymbol{Y}_1) \\
&= I(\boldsymbol{U};\boldsymbol{Y}_1|\boldsymbol{V}_1) \\
&= I(\boldsymbol{U};\boldsymbol{Y}_1,\boldsymbol{V}_2|\boldsymbol{V}_1) - I(\boldsymbol{U};\boldsymbol{V}_2|\boldsymbol{Y}_1,\boldsymbol{V}_1) \\
&= \sum_{i=1}^{n}\Big[I(U_i;\boldsymbol{Y}_1,\boldsymbol{V}_2|U^{i-1},\boldsymbol{V}_1) - I(\boldsymbol{U};V_{2,i}|\boldsymbol{Y}_1,\boldsymbol{V}_1,V_{2,1}^{i-1})\Big] \qquad (75)
\end{aligned}
$$

where the first inequality in (75) is due to the fact that the input constraint for the first channel is satisfied. For convenience, we use the notation $U^{n\setminus i}$ to denote $U^{i-1}U_{i+1}^n$, and similarly for all other vectors of random variables, e.g., $V_1^{n\setminus i} = (V_{1,1}^{i-1}, V_{1,i+1}^n)$, etc. Since $(U_i, V_{1,i})$ is independent of $(U^{i-1}, V_1^{n\setminus i})$, we have, for the first term in the summand of (75):

$$
\begin{aligned}
I(U_i;\boldsymbol{Y}_1,\boldsymbol{V}_2|U^{i-1},\boldsymbol{V}_1) &= H(U_i|V_{1,i},U^{i-1},V_1^{n\setminus i}) - H(U_i|V_{1,i},\boldsymbol{Y}_1,\boldsymbol{V}_2,U^{i-1},V_1^{n\setminus i}) \\
&= H(U_i|V_{1,i}) - H(U_i|V_{1,i},\boldsymbol{Y}_1,\boldsymbol{V}_2,U^{i-1},V_1^{n\setminus i}) \\
&= I(U_i;U^{i-1},V_1^{n\setminus i},\boldsymbol{Y}_1,\boldsymbol{V}_2|V_{1,i}). \qquad (76)
\end{aligned}
$$

Next, due to the Markov structure

$$
V_{2,i} \multimap (U_i,V_{1,i}) \multimap (U^{n\setminus i},\boldsymbol{Y}_1,V_{2,1}^{i-1},V_1^{n\setminus i}) \qquad (77)
$$

we have, for the second term in the summand of (75)

$$
\begin{aligned}
I(\boldsymbol{U};V_{2,i}|\boldsymbol{Y}_1,\boldsymbol{V}_1,V_{2,1}^{i-1}) &= H(V_{2,i}|\boldsymbol{Y}_1,\boldsymbol{V}_1,V_{2,1}^{i-1}) - H(V_{2,i}|\boldsymbol{U},\boldsymbol{Y}_1,\boldsymbol{V}_1,V_{2,1}^{i-1}) \\
&= H(V_{2,i}|\boldsymbol{Y}_1,\boldsymbol{V}_1,V_{2,1}^{i-1}) - H(V_{2,i}|U_i,\boldsymbol{Y}_1,\boldsymbol{V}_1,V_{2,1}^{i-1}) \\
&= I(U_i;V_{2,i}|\boldsymbol{Y}_1,\boldsymbol{V}_1,V_{2,1}^{i-1}). \qquad (78)
\end{aligned}
$$

Substituting (76) and (78) into (75) results in

$$
\begin{aligned}
&\rho_1 n C_1(\Gamma_1) \geq \\
&\geq \sum_{i=1}^{n} \Big[I(U_i; U^{i-1}, V_1^{n\backslash i}, \boldsymbol{Y}_1, \boldsymbol{V}_2 | V_{1,i}) - I(U_i; V_{2,i} | \boldsymbol{Y}_1, \boldsymbol{V}_1, V_{2,1}^{i-1}) \Big] \\
&= \sum_{i=1}^{n} \Big[I(U_i; V_1^{n\backslash i}, \boldsymbol{Y}_1, V_{2,1}^{i-1} | V_{1,i}) + I(U_i; U^{i-1}, V_{2,i}^{n} | V_1^{n\backslash i}, \boldsymbol{Y}_1, V_{2,1}^{i-1}, V_{1,i}) \\
&\quad - I(U_i; V_{2,i} | \boldsymbol{Y}_1, \boldsymbol{V}_1, V_{2,1}^{i-1}) \Big] \\
&= \sum_{i=1}^{n} \Big[I(U_i; V_1^{n\backslash i}, \boldsymbol{Y}_1, V_{2,1}^{i-1} | V_{1,i}) + I(U_i; U^{i-1}, V_{2,i+1}^{n} | V_{2,i}, V_{1,i}, V_1^{n\backslash i}, \boldsymbol{Y}_1, V_{2,1}^{i-1}) \Big]. \qquad (79)
\end{aligned}
$$

The Markovity $V_1 \multimap V_2 \multimap U$ implies also

$$
V_{1,i} \multimap V_{2,i} \multimap (U_i, \boldsymbol{Y}_1, V_1^{n\backslash i}, V_{2,1}^{i-1}), \qquad (80)
$$

and the second term in the summand of (79) can be written as

$$
\begin{aligned}
& I(U_i; U^{i-1}, V_{2,i+1}^{n} | \boldsymbol{Y}_1, \boldsymbol{V}_1, V_{2,1}^{i}) \\
&= H(U_i | \boldsymbol{Y}_1, \boldsymbol{V}_1, V_{2,1}^{i}) - H(U_i | U^{i-1}, \boldsymbol{Y}_1, \boldsymbol{V}_1, \boldsymbol{V}_2) \\
&= H(U_i, V_{1,i} | \boldsymbol{Y}_1, V_1^{n\backslash i}, V_{2,1}^{i}) - H(V_{1,i} | \boldsymbol{Y}_1, V_1^{n\backslash i}, V_{2,1}^{i}) - H(U_i | U^{i-1}, \boldsymbol{Y}_1, \boldsymbol{V}_1, \boldsymbol{V}_2) \\
&= H(V_{1,i} | U_i, \boldsymbol{Y}_1, V_1^{n\backslash i}, V_{2,1}^{i}) + H(U_i | \boldsymbol{Y}_1, V_1^{n\backslash i}, V_{2,1}^{i}) \\
&\quad - H(V_{1,i} | \boldsymbol{Y}_1, V_1^{n\backslash i}, V_{2,1}^{i}) - H(U_i | U^{i-1}, \boldsymbol{Y}_1, \boldsymbol{V}_1, \boldsymbol{V}_2) \\
&\stackrel{(a)}{=} H(U_i | \boldsymbol{Y}_1, V_1^{n\backslash i}, V_{2,1}^{i}) - H(U_i | U^{i-1}, \boldsymbol{Y}_1, \boldsymbol{V}_1, \boldsymbol{V}_2) \\
&= I(U_i; U^{i-1}, V_{1,i}, V_{2,i+1}^{n} | \boldsymbol{Y}_1, V_1^{n\backslash i}, V_{2,1}^{i}) \\
&\geq I(U_i; U^{i-1}, V_{2,i+1}^{n} | \boldsymbol{Y}_1, V_1^{n\backslash i}, V_{2,1}^{i}) \qquad (81)
\end{aligned}
$$

where (80) was used in(a) above. Substituting (81) in (79), we obtain the following bound on the capacity of the first channel:

$$
\rho_1 n C_1(\Gamma_1) \geq \sum_{i=1}^{n} \Big[I(U_i; \boldsymbol{Y}_1, V_1^{n\backslash i}, V_{2,1}^{i-1} | V_{1,i}) + I(U_i; U^{i-1}, V_{2,i+1}^{n} | \boldsymbol{Y}_1, V_1^{n\backslash i}, V_{2,1}^{i}) \Big]. \qquad (82)
$$

Proceeding to the second channel, observe that given $\boldsymbol{X}_2$, the output vector $\boldsymbol{Y}_2$ is independent of all other random vectors, that is,

$$(\boldsymbol{X}_1, \boldsymbol{Y}_1, \boldsymbol{V}_1, \boldsymbol{V}_2, \boldsymbol{U}) \multimap \boldsymbol{X}_2 \multimap \boldsymbol{Y}_2 \tag{83}$$

Hence

$$\begin{aligned} I(\boldsymbol{U}; \boldsymbol{Y}_2 | \boldsymbol{Y}_1, \boldsymbol{V}_1, \boldsymbol{V}_2) &= H(\boldsymbol{Y}_2 | \boldsymbol{Y}_1, \boldsymbol{V}_1, \boldsymbol{V}_2) - H(\boldsymbol{Y}_2 | \boldsymbol{Y}_1, \boldsymbol{V}_1, \boldsymbol{V}_2, \boldsymbol{U}) \\ &\leq H(\boldsymbol{Y}_2) - H(\boldsymbol{Y}_2 | \boldsymbol{X}_2, \boldsymbol{Y}_1, \boldsymbol{V}_1, \boldsymbol{V}_2, \boldsymbol{U}) \\ &\stackrel{(a)}{=} H(\boldsymbol{Y}_2) - H(\boldsymbol{Y}_2 | \boldsymbol{X}_2) = I(\boldsymbol{X}_2; \boldsymbol{Y}_2) \\ &\leq n\rho_2 C_2(\Gamma_2), \end{aligned} \tag{84}$$

where (a) holds by (83). Therefore

$$n\rho_2 C_2(\Gamma_2) \geq I(\boldsymbol{U}; \boldsymbol{Y}_2 | \boldsymbol{Y}_1, \boldsymbol{V}_1, \boldsymbol{V}_2) = \sum_{i=1}^{n} I(U_i; \boldsymbol{Y}_2 | U^{i-1}, \boldsymbol{Y}_1, \boldsymbol{V}_1, \boldsymbol{V}_2). \tag{85}$$

We now define the random variables

$$W_i = (\boldsymbol{Y}_1, V_1^{n \setminus i}, V_{2,1}^{i-1}) \tag{86}$$

$$S_i = (U^{i-1}, V_{2,i+1}^{n}, W_i) \tag{87}$$

$$Z_i = (\boldsymbol{Y}_2, S_i). \tag{88}$$

With these definitions, the following Markov relations hold:

$$W_i \multimap S_i \multimap Z_i \multimap U_i \multimap V_{2,i} \multimap V_{1,i}, \tag{89}$$

and the bounds (82) and (85) become

$$\rho_1 C_1(\Gamma_1) \geq \frac{1}{n} \sum_{i=1}^{n} [I(U_i; W_i | V_{1,i}) + I(U_i; S_i | V_{2,i}, Z_i)] \tag{90}$$

$$\rho_2 C_2(\Gamma_2) \geq \frac{1}{n} \sum_{i=1}^{n} I(U_i; Z_i | S_i, W_i, V_{2,i}). \tag{91}$$

It remains to show that the sums in (90) and (91), as well as the required distortion inequalities, can be replaced by single–letter expressions with variables W,S, and Z, satisfying the Markov structure and the conditions stated in eqs. (12) to (22) in the definition of the region $\mathcal{D}^*$. These steps follow exactly the parallel derivations in [15, Proof of Theorem 1], and are therefore omitted.

6.2 Proof of Lemma 2

Starting with the sufficiency part, assume that conditions 1, 2, and 3 above hold. Let the variables W and Z, and the mappings ψ_1 and ψ_2, be defined as follows:

$$W = Y_1 \tag{92}$$

$$Z = (Y_1, Y_2) \tag{93}$$

$$\psi_1(W, V) = g_1(Y_1, V) \tag{94}$$

$$\psi_2(Z, V) = g_2(Y_1, Y_2, V). \tag{95}$$

We now show that conditions 1 to 4 of Corollary 2 are satisfied. For the first condition, observe that

$$P_{Y_1Y_2V|U} = P_{V|UY_1Y_2}P_{Y_1Y_2|U} = P_{V|U}P_{Y_1Y_2|U} \tag{96}$$

where the second equality holds regardless of whether feedback is present or absent. Thus, we have the Markov structure $(Y_1, Y_2) \multimap U \multimap V$ and condition 1 of Corollary 2 holds. Next, since the code (f_1, g_1, f_2, g_2) incurs $(D_1, \Gamma_1, D_2, \Gamma_2)$, we must have

$$\mathbb{E}d_1(U, g_1(Y_1, V)) = D_1 \tag{97}$$

$$\mathbb{E}d_2(U, g_2(Y_1, Y_2, V)) = D_2, \tag{98}$$

which, together with eqs. (92)–(95), imply condition 2 of Corollary 2. Condition 3 of Corollary 2 is satisfied due to conditions 3 and 1 of the lemma. Condition 4 of Corollary 2 is satisfied due to condition 2 above, again with the substitutions given by eqs. (92)–(95). This completes the proof of the sufficiency part.

Turning to the necessity part, assume that for a given code (f_1, g_1, f_2, g_2), incurring strictly positive distortions and costs $(D_1, \Gamma_1, D_2, \Gamma_2)$, there exist random variables (W, Z) and mappings (ψ_1, ψ_2), satisfying conditions 1–4 of Corollary 2. We have

$$\begin{aligned} I(X_1; Y_1) &\overset{(a)}{\geq} I(U; Y_1) \\ &\overset{(b)}{\geq} I(U; Y_1|V) \\ &\overset{(c)}{\geq} I(U; W|V) \\ &\overset{(d)}{=} C_1(\Gamma_1) \end{aligned}$$

$$= \max_{\mathbb{E}\phi_1(X_1')\leq\Gamma_1} I(X_1';Y_1') \tag{99}$$

where Y_1' is the output of the first channel due to input X_1', and where (a) is implied by the data processing inequality, (b) due to the Markov chain $V \multimap U \multimap X_1 \multimap Y_1$, (c) is due to condition 4 of Corollary 2 (as W minimizes $I(U;\cdot|V)$ over a set of RV's that includes Y_1), and (d) is due to condition 3 of Corollary 2, with $\rho_1 = 1$. Similarly,

$$\begin{aligned} I(X_2;Y_2) &\overset{(a)}{\geq} I(U,Y_1;Y_2) \\ &\geq I(U,Y_1;Y_2|Y_1,V) \\ &= I(U;Y_2|Y_1,V) \\ &\geq I(U;Z|W,V) \\ &= C_2(\Gamma_2) \\ &= \max_{\mathbb{E}\phi_2(X_2')\leq\Gamma_2} I(X_2';Y_2') \end{aligned} \tag{100}$$

where here note that (a) holds whether or not feedback is included. Hence we conclude that eqs. (99) and (100) must hold with equalities. This implies all the conditions of Lemma 2.

6.3 Proof of Lemma 4

Part 1 – no feedback. Beginning with the conditions on d_1, we establish first the sufficiency part. Assume that the distortion measures satisfy (56) and (57). Let $Q_{\tilde{Y}_1,\tilde{Y}_2,U,V}$ be a distribution satisfying

$$\sum_{y_1,y_2} Q_{\tilde{Y}_1,\tilde{Y}_2,U,V}(y_1,y_2,u,v) = P_{U,V}(u,v) \quad \forall u,v \tag{101}$$

and the following Markov structure holds

$$(\tilde{Y}_1,\tilde{Y}_2) \multimap U \multimap V. \tag{102}$$

To establish sufficiency, it is enough to show that if

$$\mathbb{E}_Q d_1(U, g_1(\tilde{Y}_1,V)) \leq \mathbb{E}_P d_2(U, g_1(Y_1,V)) \tag{103}$$

then

$$I(U;\tilde{Y}_1|V) \geq I(U;Y_1|V), \tag{104}$$

and if the following simultaneously hold

$$\begin{aligned}
\mathbb{E}_Q d_2(U, g_2(\tilde{Y}_1, \tilde{Y}_2, V)) &\leq \mathbb{E}_P d_2(U, g_2(Y_1, Y_2, V)) && (105)\\
\mathbb{E}_Q d_1(U, g_1(\tilde{Y}_1, V)) &= \mathbb{E}_P d_2(U, g_1(Y_1, V)), && (106)\\
I(U; \tilde{Y}_1|V) &= I(U; Y_1|V), && (107)
\end{aligned}$$

then

$$I(U; \tilde{Y}_2|\tilde{Y}_1, V) \geq I(U; Y_2|Y_1, V), \tag{108}$$

and moreover, whenever the inequality in (103) is strict, so is the inequality in (104), and whenever the inequality in (105) is strict, so is (108). This will imply that $(I(U;Y_1|V), D_1, I(U;Y_2|Y_1,V), D_2)$ is distortion optimal, and hence also optimal (see Remark 2).

Let us examine the following quantity

$$\begin{aligned}
&I(U; \tilde{Y}_1|V) - \mathbb{E}_Q \log \frac{P_{Y_1|U}(\tilde{Y}_1|U)}{P_{Y_1|V}(\tilde{Y}_1|V)} = \\
&= \sum_{u,y_1,v} P_{UV}(u,v) Q_{\tilde{Y}_1|U}(y_1|u) \left[\log \frac{Q_{\tilde{Y}_1|U}(y_1|u)}{Q_{\tilde{Y}_1|V}(y_1|v)} - \log \frac{P_{Y_1|U}(y_1|u)}{P_{Y_1|V}(y_1|v)}\right] \\
&\overset{(a)}{=} \sum_{u,y_1,v} P_{UV}(u,v) Q_{\tilde{Y}_1|U}(y_1|u) \left[\log \frac{Q_{\tilde{Y}_1,U,V}(y_1,u,v)}{Q_{\tilde{Y}_1,V}(y_1,v) P_{U|V}(u|v)} - \log \frac{P_{Y_1,U,V}(y_1,u,v)}{P_{Y_1,V}(y_1,v) P_{U|V}(u|v)}\right] \\
&= \sum_{u,y_1,v} P_{UV}(u,v) Q_{\tilde{Y}_1|U}(y_1|u) \left[\log \frac{Q_{U|\tilde{Y}_1,V}(u|y_1,v)}{P_{UV}(u,v)} - \log \frac{P_{U|Y_1,V}(u|y_1,v)}{P_{UV}(u,v)}\right] \\
&= D(Q_{U|\tilde{Y}_1,V} || P_{U|Y_1,V}) \\
&\geq 0, \qquad (109)
\end{aligned}$$

with equality if and only if $Q_{\tilde{Y}_1,U,V} = P_{Y_1,U,V}$. (We have used the Markov structure (102), and $Y \multimap U \multimap V$, in (a).) Thus, we can deduce that, for every $\lambda_1 > 0$

$$\begin{aligned}
&\lambda_1 [I(U; \tilde{Y}_1|V) - I(U; Y_1|V)] \\
&= \lambda_1 \sum_{u,y-1,v} P_{U,V}(u,v) \left[Q_{\tilde{Y}_1|U}(y_1|u) \log \frac{Q_{\tilde{Y}_1|U}(y_1|u)}{Q_{\tilde{Y}_1|V}(y_1|v)} - Q_{\tilde{Y}_1|U}(y_1|u) \log \frac{P_{Y_1|U}(y_1|u)}{P_{Y_1|V}(y_1|v)}\right. \\
&\quad \left. + Q_{\tilde{Y}_1|U}(y_1|u) \log \frac{P_{Y_1|U}(y_1|u)}{P_{Y_1|V}(y_1|v)} - P_{Y_1|U}(y_1|u) \log \frac{P_{Y_1|U}(y_1|u)}{P_{Y_1|V}(y_1|v)}\right]
\end{aligned}$$

$$
\begin{aligned}
&\stackrel{(a)}{\geq} \lambda_1 \sum_{u,y_1,v} P_{UV}(u,v)\left[Q_{\tilde{Y}_1|U}(y_1|u) - P_{Y_1|U}(y_1|u)\right] \log \frac{P_{Y_1|U}(y_1|u)}{P_{Y_1|V}(y_1|v)} \\
&\stackrel{(b)}{\geq} \sum_{u,y_1,v} P_{UV}(u,v)\left[Q_{\tilde{Y}_1|U}(y_1|u) - P_{Y_1|U}(y_1|u)\right]\left(\lambda_1 \log \frac{P_{Y_1|U}(y_1|u)}{P_{Y_1|V}(y_1|v)} - d_1(u, g_1(y_1,v))\right) \\
&= \sum_{u,y_1} P_U(u)\left[Q_{\tilde{Y}_1|U}(y_1|u) - P_{Y_1|U}(y_1|u)\right] \sum_v P_{V|U}(v|u)\left(\lambda_1 \log \frac{P_{Y_1|U}(y_1|u)}{P_{Y_1|V}(y_1|v)} - d_1(u, g_1(y_1,v))\right) \\
&\stackrel{(c)}{=} 0,
\end{aligned}
\tag{110}
$$

where (a) follows from (109), (b) follows from (103), and (c) upon using (56). If the inequality in (103) is strict, so is the inequality in (b). This indeed proves that under (56), eqn. (103) implies (104).

We now proceed to the conditions on d_2, namely, that under (57), eqs. (105), (106), and (107), imply (108). To this end, note first that under (56), a necessary condition for (107) to hold is that inequality (a) in (110) is satisfied with equality, which in turn holds if and only if $Q_{\tilde{Y}_1,U,V} = P_{Y_1,U,V}$. Clearly, this will imply also (106). Hence, a necessary and sufficient condition for (106), (107) to hold is that $Q_{\tilde{Y}_1,U,V} = P_{Y_1,U,V}$. Thus, instead of a general distribution $Q_{\tilde{Y}_1,\tilde{Y}_2.U,V}$ satisfying (101), (102), we examine now a distribution $Q_{Y_1,\tilde{Y}_2,U,V}$, such that

$$
\sum_{y_2} Q_{Y_1,\tilde{Y}_2,U,V}(y_1,y_2,u,v) = P_{Y_1,U,V}(y_1,u,v) \quad \forall\; y_1,u,v \tag{111}
$$

and such that the following holds

$$
(Y_1,\tilde{Y}_2) \multimap U \multimap V \tag{112}
$$

(see also eq. (93)). Note that, although we have, due to the two–channel structure, the Markov conditions

$$
Y_2 \multimap U \multimap (V, Y_1), \tag{113}
$$

we do not impose

$$
\tilde{Y}_2 \multimap U \multimap (V, Y_1) \tag{114}
$$

We cannot impose such a structure, since for the quadruple $(I(U;Y_1|V), D_1, I(U;Y_2|Y_1,V), D_2)$ to be rate-distortion optimal, it has to compete with the rates and distortions implied by distributions $Q_{\tilde{Y}_1,\tilde{Y}_2,U,V}$ that satisfy the conditions stated in Theorem 1, where $\tilde{Y}_1$ and $\tilde{Y}_2$ play the role of the auxiliary random variables W and Z, respectively. A structure like (114) is not implied by the conditions stated in Theorem 1.

Assume that eq. (57) holds. We have to show that if

$$\mathbb{E}_Q d_2(U, g_2(Y_1, \tilde{Y}_2, V)) \le \mathbb{E}_P d_2(U, g_2(Y_1, Y_2, V)) \tag{115}$$

then necessarily

$$I(U; \tilde{Y}_2|Y_1, V) \ge I(U; Y_2|Y_1, V). \tag{116}$$

The proof now follows quite closely the lines of the proof of the sufficiency part for d_1. We give here few details, as we will refer to it in the case of feedback. Note first that

$$\begin{aligned}
& I(U; \tilde{Y}_2|Y_1, V) - \mathbb{E}_Q \log \frac{P_{Y_2|U}(\tilde{Y}_2|U)}{P_{Y_2|Y_1,V}(\tilde{Y}_2|Y_1, V)} = \\
&= \sum_{u,v,y_1,y_2} P_{UV}(u,v) Q_{Y_1,\tilde{Y}_2|U}(y_1, y_2|u) \left[\log \frac{Q_{\tilde{Y}_2|U,V,Y_1}(y_2|u,v,y_1)}{Q_{\tilde{Y}_2|V,Y_1}(y_2|v,y_1)} - \log \frac{P_{Y_2|U}(y_2|u)}{P_{Y_2|V,Y_1}(y_2|v,y_1)}\right] \\
&\overset{(a)}{=} D(Q_{U|Y_1,\tilde{Y}_2,V} || P_{U|Y_1,Y_2,V}) \\
&\ge 0,
\end{aligned} \tag{117}$$

where in (a) we have used (112) and (113). Thus, we can deduce that for any $\lambda_2 > 0$

$$\begin{aligned}
& \lambda_2 \left[I(U; \tilde{Y}_2|Y_1, V) - I(U; Y_2|Y_1, V)\right] \ge \\
&\overset{(a)}{\ge} \lambda_2 \sum_{u,v,y_1,y_2} P_{UV}(u,v) \left[Q_{Y_1,\tilde{Y}_2|U}(y_1, y_2|u) - P_{Y_1,Y_2|U}(y_1, y_2|u)\right] \log \frac{P_{Y_2|U}(y_2|u)}{P_{Y_2|Y_1,V}(y_2|y_1, v)} \\
&\overset{(b)}{\ge} \sum_{u,v,y_1,y_2} P_{UV}(u,v) \left[Q_{Y_1,\tilde{Y}_2|U}(y_1, y_2|u) - P_{Y_1,Y_2|U}(y_1, y_2|u)\right] \\
&\quad \times \left(\lambda_2 \log \frac{P_{Y_2|U}(y_2|u)}{P_{Y_2|Y_1,V}(y_2|y_1, v)} - d_2(u, g_2(y_1, y_2, v))\right) \\
&\overset{(c)}{=} 0
\end{aligned} \tag{118}$$

where (a) is due to (117), (b) is due to (115), and (c) follows from (57). This completes the proof of the sufficiency part.

We proceed to the necessity part. Starting with d_1, recall that P_{UV} is fixed, and in addition, from the classical Wyner–Ziv results [18], we must have the Markov structure $Y_1 \multimap U \multimap V$. Thus, we have to minimize the conditional mutual information $I(U;Y_1|V)$ only over $P_{Y_1|U}$ satisfying the distortion constraint

$$\sum_{y_1,u,v} P_{UV}(u,v)P_{Y_1|U}(y_1|u)d_1(u,g_1(y_1,v)) = D_1 \tag{119}$$

with the reproduction function given by the decoder g_1. Note also that $I(U;Y_1|V)$ is a convex function of $P_{Y_1|U}$. Consider the Lagrange functional

$$\begin{aligned}
L &= I(U;Y_1|V) + \sum_u \mu_u \left[\sum_{y_1} P_{Y_1|U}(y_1|u) - 1\right] \\
&\quad + \lambda \left[\sum_{y_1,u,v} P_{UV}(u,v)P_{Y_1|U}(y_1|u)d_1(u,g_1(y_1,v)) - D_1\right] \\
&= \sum_{y_1,u,v} P_{UV}(u,v)P_{Y_1|U}(y_1|u) \log \frac{P_{Y_1|U}(y_1|u)}{\sum_{u'} P_{Y_1|U}(y_1|u')P_{U|V}(u'|v)} + \sum_u \mu_u \left[\sum_{y_1} P_{Y_1|U}(y_1|u) - 1\right] \\
&\quad + \lambda \left[\sum_{y_1,u,v} P_{UV}(u,v)P_{Y_1|U}(y_1|u)d_1(u,g_1(y_1,v)) - D_1\right].
\end{aligned} \tag{120}$$

(Using common optimization techniques, the positivity constraints on $P_{Y_1|U}(y_1|u)$ are taken care of later, when we consider positive (resp. negative) derivatives of L if the optimality point occurs at $P_{Y_1|U}(y_1|u) = 0$ (resp. $P_{Y_1|U}(y_1|u) = 1$). See (125).) Introducing the new variables

$$-\log r(u,v) = \frac{\mu_u}{P_{U|V}(u,v)} \tag{121}$$

we can write

$$L = \sum_{y_1,u,v} P_{UV}(u,v)\ell(u,v,y_1) + \sum_{u,v} P_{UV}(u,v)\log r(u,v) - \lambda D_1 \tag{122}$$

where

$$\begin{aligned}
\ell(u,v,y_1) &= P_{Y_1|U}(y_1|u) \log \frac{P_{Y_1|U}(y_1|u)}{r(u,v)\sum_{u'} P_{Y_1|U}(y_1|u')P_{U|V}(u'|v)} \\
&\quad + \lambda P_{Y_1|U}(y_1|u)d_1(u,g_1(y_1,v)).
\end{aligned} \tag{123}$$

Differentiating with respect to $P_{Y_1|U}(y_1'|u')$, we obtain, after some algebraic manipulations:

$$\frac{\partial L}{\partial P_{Y_1|U}(y_1'|u')} = \sum_{y_1,u,v} P_{UV}(u,v) \frac{\partial \ell(u,v,y_1)}{\partial P_{Y_1|U}(y_1'|u')}$$

$$= \sum_v P_{UV}(u',v)\left[\log\frac{P_{Y_1|U}(y_1'|u')}{r(u',v)P_{Y_1|V}(y_1'|v)} + \lambda d(u',g_1(y_1',v))\right], \tag{124}$$

implying that the optimal distribution $P_{Y_1|U}$ satisfies:

$$\lambda\mathbb{E}\left\{d_1(u,g_1(y_1,V))|U=u\right\}\begin{cases} \geq \mathbb{E}\left\{\log P_{Y_1|V}(y_1|V)|U=u\right\} \\ \quad -\log P_{Y_1|U}(y_1|u)+k(u) & \text{if } P_{Y_1|U}(y_1|u)=0, \\ \leq \mathbb{E}\left\{\log P_{Y_1|V}(y_1|V)|U=u\right\} \\ \quad -\log P_{Y_1|U}(y_1|u)+k(u) & \text{if } P_{Y_1|U}(y_1|u)=1, \\ = \mathbb{E}\left\{\log P_{Y_1|V}(y_1|V)|U=u\right\} \\ \quad -\log P_{Y_1|U}(y_1|u)+k(u) & \text{otherwise.} \end{cases} \tag{125}$$

For the pairs (u,y_1) satisfying $P_{Y_1|U}(y_1|u)=0$, the exact value of the right hand side of (125) is inconsequencial, as these pairs occure with probability 0. For those pairs (u,y_1) satisfying $P_{Y_1|U}(y_1|u)=1$, the value of the right hand side of (125) depends only on u, and hence can be set to any desired function by proper choice of $k(u)$. Therefore, (125) implies that we can set altogether

$$\lambda\mathbb{E}\left\{d_1(u,g_1(y_1,V))|U=u\right\} = \mathbb{E}\left\{\log P_{Y_1|V}(y_1|V)|U=u\right\} - \log P_{Y_1|U}(y_1|u)+k(u) \tag{126}$$

in accordance with the assertion of the lemma.

Regarding the necessity part on d_2, we minimize $I(U;Y_2|Y_1,V)$ with respect to $P_{Y_2|Y_1,U}$. Note again that $I(U;Y_2|Y_1,V)$ is a convex function of $P_{Y_2|Y_1,U}$ for fixed $P_{Y_1|U}$. Moreover, due to the Markov structure (113), $P_{Y_2|Y_1,U}(y_2|y_1,u)$ is independent of y_1. Thus we can take the same approach as above, and consider the Lagrange functional

$$\begin{aligned} L &= I(U;Y_2|Y_1,V) + \sum_u \mu_u\left[\sum_{y_2} P_{Y_2|U}(y_2|u) - 1\right] \\ &\quad + \lambda\left[\sum_{y_1,y_2,u,v} P_{U,V}(u,v)P_{Y_1|U}(y_1|u)P_{Y_2|U}(y_2|u)d_2(u,g_2(y_1,y_2,v)) - D_2\right]. \end{aligned} \tag{127}$$

From this point, the proof of the necessity condition on d_2 proceeds along the lines of the proof of the necessity condition on d_1, and is therefore omitted.

Part 2 – with feedback. Note that the feedback has no effect on the achievable distortions and costs, and no effect on the distribution of the random variables involved in the first stage. Therefore, the condition on d_1 remains as in the case of no feedback. The presence of feedback comes into account in the condition on d_2, since (113), which was used in (117), no longer holds. Thus, instead of proceeding along the lines of (117) and (118), we examine now the following quantity

$$
\begin{aligned}
& I(U;\tilde{Y}_2|Y_1,V) - \mathbb{E}_Q \log \frac{P_{Y_2|U,V,Y_1}(\tilde{Y}_2|U,V,Y_1)}{P_{Y_2|Y_1,V}(\tilde{Y}_2|Y_1,V)} = \\
&= \sum_{u,v,y_1,y_2} P_{UV}(u,v) Q_{Y_1,\tilde{Y}_2|U}(y_1,y_2|u) \left[\log \frac{Q_{\tilde{Y}_2|U,V,Y_1}(y_2|u,v,y_1)}{Q_{\tilde{Y}_2|V,Y_1}(y_2|v,y_1)} - \log \frac{P_{Y_2|U,V,Y_1}(y_2|u,v,y_1)}{P_{Y_2|V,Y_1}(y_2|v,y_1)}\right] \\
&\stackrel{(a)}{=} D(Q_{U|Y_1,\tilde{Y}_2,V}||P_{U|Y_1,Y_2,V}) \\
&\geq 0, \qquad (128)
\end{aligned}
$$

where in (a) we have used (112). Thus, we can deduce that for any $\lambda_2 > 0$

$$
\begin{aligned}
& \lambda_2 \left[I(U;\tilde{Y}_2|Y_1,V) - I(U;Y_2|Y_1,V)\right] \geq \\
& \stackrel{(a)}{\geq} \lambda_2 \sum_{u,v,y_1,y_2} P_{UV}(u,v) \left[Q_{Y_1,\tilde{Y}_2|U}(y_1,y_2|u) - P_{Y_1,Y_2|U}(y_1,y_2|u)\right] \log \frac{P_{Y_2|U,V,Y_1}(y_2|u,v,y_1)}{P_{Y_2|Y_1,V}(y_2|y_1,v)} \\
& \stackrel{(b)}{\geq} \sum_{u,v,y_1,y_2} P_{UV}(u,v) \left[Q_{Y_1,\tilde{Y}_2|U}(y_1,y_2|u) - P_{Y_1,Y_2|U}(y_1,y_2|u)\right] \\
& \quad \times \left(\lambda_2 \log \frac{P_{Y_2|U,V,Y_1}(y_2|u,v,y_1)}{P_{Y_2|Y_1,V}(y_2|y_1,v)} - d_2(u, g_2(y_1,y_2,v))\right) \\
& \stackrel{(c)}{=} 0 \qquad (129)
\end{aligned}
$$

where (a) is due to (128), (b) due to (115), and (c) due to (58). Hence the sufficiency follows.

To establish the necessity part in case of feedback, observe that $I(U;Y_2|Y_1,V)$ is still a convex function of $P_{Y_2|Y_1,U}$, but the dependence on y_1 cannot be dropped, i.e., (113) does not hold. Therefore we consider the Lagrange functional

$$
\begin{aligned}
L &= I(U;Y_2|Y_1,V) + \sum_{u,y_1} \mu_{u,y_1} \left[\sum_{y_2} P_{Y_2|Y_1,U}(y_2|y_1,u) - 1\right] \\
&\quad + \lambda \left[\sum_{y_1,y_2,u,v} P_{U,V}(u,v) P_{Y_1|U}(y_1|u) P_{Y_2|Y_1,U}(y_2|y_1,u) d_2(u, g_2(y_1,y_2,v)) - D_2\right]. \qquad (130)
\end{aligned}
$$

From this point, the proof proceeds along the lines of the proof on the necessity condition on d_1, and is omitted. We just note that the dependence of the Lagrange mutipliers μ_{u,y_1} on y_1 results in the dependence of the function k_2 in (58) on y_1, in addition to its dependence on u. □

References

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

[2] J. Chow and T. Berger, "Failure of successive refinement for symmetric Gaussian mixtures," *IEEE Trans. Inform. Theory*, vol. 43, no. 1, pp. 350–352, January 1997.

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

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

[5] W. H. R. Equitz, "Successive refinement of information," Ph.D. dissertation, Stanford University, June 1989.

[6] W. H. R. Equitz and T. M. Cover, "Successive refinement of information," *IEEE Trans. Inform. Theory*, vol. IT–37, no. 2, pp. 269–275, March 1991.

[7] M. Gastpar, B. Rimoldi, and M. Vetterli, "To code, or not to code: Lossy source-channel communication revisited," *IEEE Trans. Inform. Theory*, vol. 49, no. 5, pp. 1147–1158, May 2003.

[8] A. Kanlis and P. Narayan, "Error exponents for successive refinement by partitioning," *IEEE Trans. Inform. Theory*, vol. IT–42, no. 1, January 1996.

[9] V. N. Koshelev, "Hierarchical coding of discrete sources," *Problems of Information Transmission*, pp. 186-203, 1981.

[10] V. N. Koshelev, "On the divisibility of discrete sources with an additive single-letter distortion measure," *Problems of Information Transmission (IPPI)*, vol. 30, no. 1, pp. 27–43, 1994.

[11] N. Merhav, R. M. Roth, and E. Arikan, "Hierarchical guessing with a fidelity criterion," *IEEE Trans. Inform. Theory*, vol. 45, no. 1, pp. 330–337, January 1999.

[12] N. Merhav and S. Shamai (Shitz), "On joint source–channel coding for the Wyner–Ziv source and the Gel'fand–Pinsker channel," to appear in *IEEE Trans. Inform. Theory*, November 2003.

[13] B. Rimoldi, "Successive refinement of information: characterization of achievable rates," *IEEE Trans. Inform. Theory*, vol. IT–40, no. 1, pp. 253–259, January 1994.

[14] S. Shamai (Shitz), S. Verdú and R. Zamir, "Systematic lossy source/channel coding," *IEEE Trans. Inform. Theory*, vol. 44, no. 2, pp. 564–579, March 1998.

[15] Y. Steinberg and N. Merhav, "On successive refinement for the Wyner–Ziv problem," accepted to *IEEE Trans. Inform. Theory*.

[16] E. Tuncel and K. Rose, "Additive successive refinement," *IEEE Trans. Inform. Theory*, vol. 49, no. 8, pp. 1983–1991, August 2003.

[17] E. Tuncel, J. Nayak, and K. Rose, "On hierarchical type covering," preprint 2003.

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