To Code, Or Not To Code: On the Optimality of Symbol-by-Symbol Communication*

Michael Gastpar,* Bixio Rimoldi,* and Martin Vetterli*†

May 23, 2001

*Communication Systems Department
Swiss Federal Institute of Technology, Lausanne, Switzerland
Email: {Michael.Gastpar, Bixio.Rimoldi, Martin.Vetterli}@epfl.ch

†Department of EECS
UC Berkeley, Berkeley, CA 94720, USA

Abstract

When is uncoded transmission optimal? This paper derives easy-to-check necessary and sufficient conditions that do not require finding the rate-distortion and the capacity-cost functions. We consider the symbol-by-symbol communication of discrete-time memoryless sources across discrete-time memoryless channels, using single-letter coding and decoding. This is an optimal communication system if and only if the channel input cost function and the distortion measure can be written in a form that we explicitly characterize. There are two well-known examples where uncoded transmission is optimal. The first example consists of a Gaussian source and a Gaussian channel. In the second example the source and the channel are binary. But these are just two out of infinitely many examples that one can construct in a straightforward way from our results. As a matter of fact, one can arbitrarily pick the source distribution, the single-letter encoder/decoder, and the channel conditional distribution, and make the system optimal by choosing the channel input cost function and the distortion measure according to the given closed-form expression. The paper also discusses the advantages of uncoded transmission for non-ergodic channels and multiuser communications. Finally, some results concerning M-block-length codes are obtained.

Keywords: uncoded transmission, joint source/channel coding, separation theorem, single-letter codes, single-source broadcast

*The material in this paper was presented in part at the 2000 IEEE International Symposium on Information Theory, Sorrento, Italy, ("To code or not to code," p. 236), and will be presented in part at the 2001 IEEE International Symposium on Information Theory, Washington DC ("On source/channel codes of finite block length").

1 Introduction

Communications engineers have a long acquaintance with the "separation principle," i.e., the strategy of splitting the coding into two stages, source compression and channel coding. This key strategy has been introduced and shown to be optimal by Shannon in 1948 [1, Thm. 21]. The result is of surprisingly wide validity in point-to-point communication [2]. Consequently, the separation idea has split the research community into two camps, those who examine source compression and those who investigate channel coding.

In parallel, many researchers have been aware of the central shortcomings of the separation principle: it disregards delay and complexity issues, and it does not generally hold in non-ergodic and multiuser communication. In fact, to prove the separation theorem, it is necessary to allow for infinite coding complexity and delay in general. A *joint* source/channel code may reduce both delay and complexity (which is illustrated below), but to design such a code is generally a more difficult optimization problem. In some cases, however, the source and the channel are already somewhat *matched* to each other. The separation strategy cannot exploit such a favorable situation to reduce complexity, but a joint source/channel code may very well do so. In fact, source and channel may be matched so well that *uncoded* transmission is already sufficient to achieve optimal performance: The source output is applied directly to the channel input, and the channel output is the estimate of the source. In that case, complexity and delay are reduced to their absolute minimum. One famous example is the transmission of a Gaussian source across an additive white Gaussian noise (AWGN) channel, another example the transmission of a binary uniform source across a binary symmetric channel. Clearly, uncoded transmission is meaningful only when the source and channel alphabets are the same. This constraint can be removed by allowing for a *single-letter* mapping, i.e., a rule that maps each source output symbol separately onto one channel input symbol. Such single-letter joint source/channel codes are the objects of study of the present paper.

The paper is organized as follows. After giving the definitions in Section 2, we develop in Section 3 a criterion to establish the optimality of a communication scheme that employs single-letter codes. To find such a criterion following standard textbook information theory, one could determine the end-to-end distortion Δ incurred and the power (or, more generally, *cost*) Γ used on the channel by the purported transmission strategy. Then, a necessary condition for optimality is that the rate needed to encode the source at distortion Δ has to be equal to the capacity of the channel at input cost Γ. To verify this, rate-distortion and capacity-cost functions have to be determined. Unfortunately, this problem has explicit solutions only in a handful of special cases. In general, numerical methods are required. In

contrast to this, following a slightly different approach, the key point of the present paper is that it is indeed possible to give an explicit answer.

In a nutshell, suppose that a memoryless source specified by the random variable S with distribution $p(s)$ is encoded (symbol-by-symbol) into $X = f(S)$. The symbol X is transmitted across a memoryless noisy channel specified by a conditional distribution $p_{Y|X}$. The channel output Y is decoded to yield the estimate of the source, $\hat{S} = g(Y)$. In this paper, we show that this is an optimal communication system if and only if the channel input cost function $\rho(x)$ and the distortion measure $d(s, \hat{s})$ are chosen (up to shifts and scaling) as

$$\begin{aligned} \rho(x) &= D\left(p_{Y|X}(\cdot|x)||p_Y(\cdot)\right) \\ d(s,\hat{s}) &= -\log_2 p(s|\hat{s}), \end{aligned}$$

where $D(\cdot||\cdot)$ denotes the Kullback-Leibler distance, $p_Y(\cdot)$ the distribution of the channel output Y, and $p(s|\hat{s})$ the distribution of S given the estimate $\hat{S}$. These arguments are made precise in Theorem 7. Our solution differs from the classical approach (where the input cost function and the distortion measure are the fixed quantities) in that we fix the source and channel distributions and pick the cost function and the distortion measure such as to make the system optimal. From the above formulae, it is clear that our criterion to establish optimality can be used directly to construct an arbitrarily large supply of communication systems in the spirit of the famous Gaussian example. Section 4 provides a few examples that illustrate this.

Section 5 presents and develops some applications of the theory. In Section 5.1, we tackle the question of the *existence* of single-letter codes that perform optimally: For a given source/channel pair, is there a single-letter code that performs optimally? We present explicit answers for some specific classes of source/channel pairs.

In Section 5.2, the results obtained in Section 3 are applied to longer codes. For a given source/channel pair, suppose that there is no single-letter code that performs optimally. Will there be a block code of length M that does? For a certain class of discrete memoryless source/channel pairs, we find that the answer is negative (as long as the length remains finite).

The significance of single-letter joint source/channel codes extends beyond the validity of the separation principle. This is discussed in Section 5.3. In fact, such codes feature a certain universality in that one and the same code may perform optimally for an entire set of source/channel pairs. This is relevant for instance for non-ergodic channels and multi-user communications. A final example illustrates the potential practical interest in single-letter source/channel codes: such simple codes may actually outperform any strategy

that is based on the (generally unjustified) application of Shannon's separation theorem to multi-user communication.

2 Definitions

The key elements of the problem studied in this paper are the discrete-time memoryless source and channel, and the single-letter code. In this section, we provide definitions of those entities. We denote random variables by capital letters, e.g. S, and their realizations by lower-case letters, e.g. s. The distribution of the random variable S is denoted by $p_S(s)$. For continuous alphabets, $p_S(s)$ is a probability density function (pdf); for discrete alphabets, a probability mass function (pmf). When the subscript is just the capitalized version of the argument in parentheses, we will often write simply $p(s)$.

Definition 1 (source). A (discrete-time memoryless) source (p_S, d) is specified by a pdf (or pmf, respectively) $p_S(s)$ on an alphabet $\mathcal{S}$ and a nonnegative function $d(s, \hat{s}) : \mathcal{S} \times \hat{\mathcal{S}} \to \mathbb{R}^+$ called the distortion measure. The rate-distortion function (see e.g. [3]) of the source (p_S, d) is denoted by $R(D)$.

Definition 2 (channel). A (discrete-time memoryless) channel $(p_{Y|X}, \rho)$ is specified by a conditional pdf (or pmf, respectively) $p_{Y|X}(y|x)$, where $X \in \mathcal{X}$ and $Y \in \mathcal{Y}$, and a nonnegative function $\rho(x) : \mathcal{X} \to \mathbb{R}^+$ called the channel input cost function. The capacity-cost function (see e.g. [4]) of the channel $(p_{Y|X}, \rho)$ is denoted by $C(P)$. This function is also called capacity-constraint function in [5].

In order to decide on the optimality of a communication system that employs single-letter codes, the unconstrained capacity of the channel turns out to be an important quantity:

Definition 3 (unconstrained capacity). The *unconstrained capacity* of the channel $(p_{Y|X}, \rho)$ is the capacity of the channel disregarding input costs, that is

$$C_0 \quad = \quad \max_{p_X} I(X;Y). \tag{1}$$

Hence, C_0 is independent of the choice of ρ; it is solely a property of $p_{Y|X}$. When $\rho(x) < \infty, \forall x \in \mathcal{X}$, an equivalent definition is $C_0 = C(P \to \infty)$. Note also that C_0 is infinite for some channels (e.g. the AWGN channel).

In this paper, we study communication by means of a single-letter code, defined as follows:

Definition 4 (single-letter source/channel code). A single-letter source/channel code (f, g) is specified by an encoding function $f(\cdot) : \mathcal{S} \to \mathcal{X}$ and a decoding function $g(\cdot) : \mathcal{Y} \to \hat{\mathcal{S}}$.

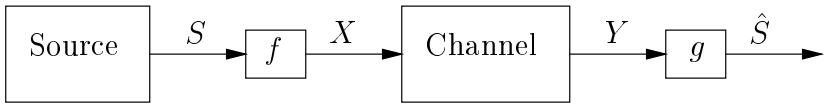

Figure 1: The considered system.

Suppose the source (p_S, d) is transmitted across the channel $(p_{Y|X}, \rho)$ using the single-letter code (f, g). The average input cost used on the channel is found to be $\Gamma = E\rho(X) = E\rho(f(S))$, and the average distortion achieved by the code (f, g) is $\Delta = Ed(S, \hat{S}) = Ed(S, g(Y))$. We will sometimes refer to (Γ, Δ) as the *cost-distortion* pair. The main goal of this paper is to determine necessary and sufficient conditions such that this communication scheme performs optimally according to the following definition:

Definition 5 (optimality). For the transmission of a source (p_S, d) across a channel $(p_{Y|X}, \rho)$, a single-letter source/channel code (f, g) is optimal if both

(i) the distortion $\Delta = Ed(S, \hat{S})$ incurred using (f, g) is the minimum distortion that can be achieved at input cost $\Gamma = E\rho(X)$ with the best possible joint source/channel code (regardless of complexity), and

(ii) the cost $\Gamma = E\rho(X)$ incurred using (f, g) is the minimum cost needed to achieve distortion $\Delta = Ed(S, \hat{S})$ with the best possible joint source/channel code (regardless of complexity). [1]

The scope of the present investigations is limited to communication using single-letter codes. Nevertheless, it is clear that longer source/channel codes are of interest, too. Such a code would map M source symbols onto N channel symbols. Let us point out that when all alphabets are discrete, any longer source/channel code can be interpreted as a single-letter code in appropriately extended alphabets. This point of view turns out to be useful in Section 5.2.

3 Single-Letter Codes That Perform Optimally

It is well-known that there are instances of source/channel pairs for which single-letter codes achieve the best possible performance. This result is particularly surprising since

[1] Note that the two conditions do not necessarily imply one another. In fact, in the literature, optimality of a transmission scheme is sometimes defined by one of the two conditions only. Our results can be modified to apply to that case as well.

such codes are extremely easy to implement and operate at zero delay. In this section, we derive necessary and sufficient conditions under which single-letter codes are optimal.

The first insight is that a point-to-point communication system is optimal essentially if and only if $R(\Delta) = C(\Gamma)$, where Δ is the incurred distortion and Γ is the channel input cost that is used. This follows straightforwardly from the separation principle, and it is also a simple corollary to the results of [6].

Clearly, if Δ can be decreased without changing $R(\Delta)$, then the condition $R(\Delta) = C(\Gamma)$ is *not* sufficient for optimality. Similarly, if Γ can be decreased without changing $C(\Gamma)$, the condition $R(\Delta) = C(\Gamma)$ is *not* sufficient for optimality, either. Thus, we can state the following basic lemma:

Lemma 1. *The transmission of the source (p_S, d) across the channel $(p_{Y|X}, \rho)$ by means of a single-letter code (f, g) is optimal if and only if*

(i) $R(\Delta) = C(\Gamma)$, *and*

(ii) *neither can Δ be lowered without changing $R(\Delta)$ nor can Γ be lowered without changing $C(\Gamma)$.*

Remark. Δ may be decreased without changing $R(\Delta)$ only if $R(\Delta) = 0$. Likewise, Γ may be decreased without changing $C(\Gamma)$ only if $C(\Gamma) = C_0$. This is developed in Section 3.2.

Proof. ($\Rightarrow$.) To prove the necessity of Lemma 1, we need Shannon's separation theorem [1, Thm. 21]. It states that there does not exist a communication strategy such that $R(\Delta) > C(\Gamma)$, and that if $R(\Delta) < C(\Gamma)$, then there exists a better communication system. Hence, if the system is optimal, $R(\Delta) = C(\Gamma)$.
($\Leftarrow$.) To prove the sufficiency of Lemma 1, first note that by assumption, $R(\Delta) = C(\Gamma)$. By the definition of the rate-distortion function, $R(\Delta)$ is the least rate needed to describe the source at distortion Δ. Since by assumption, Δ cannot be decreased, it is the smallest distortion achievable with $C(\Gamma)$ bits. By the definition of the capacity-cost function, $C(\Gamma)$ is the maximum rate at which communication is feasible at input cost ("power") Γ. Since by assumption, Γ cannot be decreased, it is the smallest cost to allow for a rate of $C(\Gamma)$, which completes the proof. □

Lemma 1 contains two conditions that together are necessary and sufficient to establish the optimality of a communication system that uses single-letter codes. These conditions will now be examined in detail. In Section 3.1, we elaborate on the first condition, i.e., $R(\Delta) = C(\Gamma)$. The second condition is somewhat subtler; it will be discussed in Section 3.2. In Section 3.3, the results are combined to yield a general criterion for the optimality of single-letter codes.

3.1 Condition *(i)* of Lemma 1

As a first step, we can reformulate the condition $R(\Delta) = C(\Gamma)$ more explicitly as follows:

Lemma 2. *$R(\Delta) = C(\Gamma)$ holds if and only if the following three conditions are simultaneously satisfied:*

(i) *the distribution p_X of $X = f(S)$ achieves capacity on the channel $(p_{Y|X}, \rho)$ at maximum input cost $\Gamma = E\rho(X)$, i.e., $I(X;Y) = C(\Gamma)$,*

(ii) *the conditional distribution $p_{\hat{S}|S}$ of $\hat{S} = g(Y)$ given S achieves the rate-distortion function of the source (p_S, d) at distortion $\Delta = Ed(S,\hat{S})$, i.e. $I(S;\hat{S}) = R(\Delta)$, and*

(iii) *$f(\cdot)$ and $g(\cdot)$ are such that $I(S;\hat{S}) = I(X;Y)$.*

Proof.

$$\begin{aligned} R(\Delta) &= \min_{q_{\hat{S}|S}: Ed(S,\hat{S}) \le \Delta} I(S;\hat{S}) \overset{(a)}{\le} I(S;\hat{S}) \overset{(b)}{\le} I(X;Y) \\ &\overset{(c)}{\le} \max_{q_X: E\rho(X) \le \Gamma} I(X;Y) = C(\Gamma), \end{aligned} \tag{2}$$

with equality in (a) if and only if $p_{\hat{S}|S}$ achieves the rate-distortion function of the source, in (b) if and only if $I(S;\hat{S}) = I(X;Y)$, and in (c) if and only if p_X achieves the capacity-cost function of the channel. Thus, $R(\Delta) = C(\Gamma)$ is satisfied if and only if all three conditions in Lemma 2 are satisfied, which completes the proof. □

There are four pairs of entities involved, namely the source (p_S, d), the channel $(p_{Y|X}, \rho)$, the code (f, g) and the cost-distortion pair (Γ, Δ). These four pairs are not independent of one another. For instance, the latter is completely determined by the first three. The corresponding communication system (as shown in Fig. 1) performs optimally if and only if these four pairs are selected in such a way as to fulfill all the requirements of Lemma 1.

There are various ways to verify whether the requirements are satisfied. Some of them lead to problems that notoriously do not admit analytical solutions. For example, following Lemma 2, we could compute the capacity-cost function $C(\cdot)$ of the channel $(p_{Y|X}, \rho)$ and evaluate it at Γ. This is known to be a problem that does not have a closed-form solution for all but a small set of channels. Similarly, one could compute the rate-distortion function $R(\cdot)$ of the source (p_S, d) and evaluate it at Δ. Again, closed-form solutions are known only for a handful of special cases. Once the rate-distortion and the capacity-cost functions are determined, we are ready to check the conditions of Lemma 1.

One of the main difficulties with this approach lies in the fact that for a given cost function ρ, there is no general closed-form expression for the channel input distribution

that achieves capacity; numerical solutions can be found via the Arimoto-Blahut algorithm. The key idea of the following theorem is to turn this game around: for any distribution q_X over the channel input alphabet $\mathcal{X}$, there exists a closed-form solution for the input cost function ρ such that the distribution q_X achieves capacity.

Theorem 3. *For fixed source distribution p_S, single-letter encoder f and channel conditional distribution $p_{Y|X}$:*

(i) If $I(X;Y) < C_0$, the first condition of Lemma 2 is satisfied if and only if the input cost function satisfies

$$\rho(x) \quad = \quad c_1 D(p_{Y|X}(\cdot|x)||p_Y(\cdot)) + \rho_0, \tag{3}$$

where $c_1 > 0$ and ρ_0 are constants, and $D(\cdot||\cdot)$ denotes the Kullback-Leibler distance between two distributions.

(ii) If $I(X;Y) = C_0$, the first condition of Lemma 2 is satisfied for any function $\rho(x)$.

To gain insight, let q_X be the channel input distribution induced by some source distribution through the encoder f. For any cost function ρ, one finds an expected cost and a set of admissible input distributions leading to the same (or smaller) average cost. The input distribution q_X lies in that set, but it does not necessarily maximize mutual information. The key is now to find the cost function, and thus the set of admissible input distributions, in such a way that the input distribution q_X maximizes mutual information within the set. In the special case where the input distribution q_X achieves C_0, it clearly maximizes mutual information among distributions in *any* set, regardless of ρ. Hence, in that case, the choice of the cost function ρ is unrestricted.

A formal proof follows. The reader interested in not interrupting the flow of the exposition is advised to skip to Theorem 4.

Proof. Let $p_{Y|X}$ be fixed. For any distribution p_X on $\mathcal{X}$, define

$$I'_{p_X}(x) \quad = \quad D(p_{Y|X}(\cdot|x)||p_Y), \tag{4}$$

where $p_Y(y) = Ep_{Y|X}(y|X)$ is the marginal distribution of Y when X is distributed according to p_X.

It is quickly verified that with this definition, $I_{p_X}(X;Y) = \langle p_X, I'_{p_X} \rangle$, where $\langle f, g \rangle$ denotes the standard inner product, i.e. for discrete alphabets, $\langle f, g \rangle = \sum_x f(x)g(x)$ and for continuous alphabets, $\langle f, g \rangle = \int f(x)g(x)dx$. With this notation, we may write $D(p_{Y|X}(\cdot|x)||p_Y) = \langle p_{Y|X}, \log_2 \frac{p_{Y|X}}{p_Y} \rangle_y$, where the subscript emphasizes that the inner product is taken in the variable y. The following auxiliary lemma is crucial for the proof:

Lemma: For any p_X and $\tilde{p}_X$, $I_{\tilde{p}_X}(X;Y) - I_{p_X}(X;Y) \;\leq\; \langle \tilde{p}_X - p_X, I'_{p_X} \rangle$.

To see this, note first that since $I_{p_X}(X;Y) = \langle p_X, I'_{p_X} \rangle$, we equivalently prove the inequality $\langle \tilde{p}_X, I'_{p_X} \rangle - I_{\tilde{p}_X}(X;Y) \;\geq\; 0$, for any $p_X, \tilde{p}_X$.

$$\begin{aligned}
\langle \tilde{p}_X, I'_{p_X} \rangle - I_{\tilde{p}_X}(X;Y) &= \langle \tilde{p}_X, I'_{p_X} \rangle - \langle \tilde{p}_X, I'_{\tilde{p}_X} \rangle \\
&= \langle \tilde{p}_X, I'_{p_X} - I'_{\tilde{p}_X} \rangle \\
&= \langle \tilde{p}_X, D(p_{Y|X}||p_Y) - D(p_{Y|X}||\tilde{p}_Y) \rangle \\
&= \langle \tilde{p}_X, \langle p_{Y|X}, \log_2 \frac{\tilde{p}_Y}{p_Y} \rangle_y \rangle_x \\
&\overset{(a)}{=} \langle \langle \tilde{p}_X, p_{Y|X} \rangle_x, \log_2 \frac{\tilde{p}_Y}{p_Y} \rangle_y \\
&= \langle \tilde{p}_Y, \log_2 \frac{\tilde{p}_Y}{p_Y} \rangle_y \\
&= D(\tilde{p}_Y||p_Y) \;\geq\; 0, \qquad (5)
\end{aligned}$$

where (a) is a change of summation (or integration) order and the inequality follows from the fact that the Kullback-Leibler distance is nonnegative. The theorem can then be proved as follows.

($\Leftarrow$.) (Sufficiency of the formula.) Fix a distribution p_X over the channel input alphabet. Let ρ be arbitrary and let $\tilde{p}_X$ be any channel input distribution such that

$$E_{\tilde{p}_X}\rho(X) \;\leq\; E_{p_X}\rho(X). \qquad (6)$$

For any $\lambda \geq 0$,

$$\begin{aligned}
I_{p_X}(X;Y) - I_{\tilde{p}_X}(X;Y) &\geq \langle p_X - \tilde{p}_X, I'_{p_X} \rangle \\
&\geq \langle p_X - \tilde{p}_X, I'_{p_X} - \lambda\rho \rangle, \qquad (7)
\end{aligned}$$

where the first inequality is the last lemma, and the second follows by assumption on $\tilde{p}_X$. If $\lambda\rho(x) = I'_{p_X}(x) + c$, then the last expression is zero, proving that $I_{p_X}(X;Y)$ indeed maximizes mutual information.

When $I_{p_X}(X;Y) = C_0$, then the input distribution p_X maximizes $I(X;Y)$ regardless of $\rho(x)$ and trivially fulfills the expected cost constraint.

($\Rightarrow$.) (Necessity of the formula.) In order to establish the necessity of formula (3), we need the derivative of the mutual information $I(X;Y)$ with respect to the distribution of X. In the finite-dimensional case, we can argue as follows: As long as $I(X;Y) < C_0$, we are in the *strictly* increasing region of $C(P)$, and therefore

$$p_X \;\in\; \arg \max_{\tilde{p}_X : E_{\tilde{p}_X}\rho(X) \leq E_{p_X}\rho(X)} I_{\tilde{p}_X}(X;Y) \;\Longrightarrow\; p_X \;\in\; \arg \max_{\tilde{p}_X : E_{\tilde{p}_X}\rho(X) = E_{p_X}\rho(X)} I_{\tilde{p}_X}(X;Y).$$

But a first-order necessary condition for p_X to be a maximizer in the latter maximization problem is that the corresponding Lagrange functional, evaluated at p_X, vanishes. That is,

$$\frac{d}{d\tilde{p}_X(x)} L^{(\rho)}(\tilde{p}_X, \lambda, \mu)\Big|_{\tilde{p}_X = p_X} \quad = \quad 0, \tag{8}$$

for every $x \in \mathcal{X}$, where

$$L^{(\rho)}(\tilde{p}_X, \lambda, \mu) \quad = \quad I_{\tilde{p}_X}(X;Y) - \lambda(\sum_x \tilde{p}_X(x)\rho(x) - \Gamma) - \mu(\sum_x \tilde{p}_X(x) - 1). \tag{9}$$

By evaluating the derivatives, this indeed implies the claimed formula (3), as long as $\lambda \neq 0$. When does $\lambda = 0$ occur? Writing out the formula for λ,

$$\lambda \quad = \quad \frac{1}{\rho(s)} \left(\frac{d}{dp_X(x)} I(X;Y) - \mu \right), \tag{10}$$

we see that $\lambda = 0$ if and only if $\frac{d}{dp_X(x)} I(X;Y) = \text{const.}$, for all x. But this can only arise if $I(X;Y) = C_0$.

This argument may be extended to infinite dimensions by considering the Gateaux differential of $I_p(X;Y)$ and using Thm. 2, p. 188, in Luenberger [7]. □

Theorem 3 gives an explicit formula to select the input cost function ρ for given channel conditional and input distributions. By analogy, the next theorem gives a similar condition for the distortion measure.

Theorem 4. *For fixed source distribution p_S, channel conditional distribution $p_{Y|X}$ and single-letter code (f, g):*

(i) If $0 < I(S;\hat{S})$, the second condition of Lemma 2 is satisfied if and only if the distortion measure satisfies

$$d(s, \hat{s}) \quad = \quad -c_2 \log_2 p(s|\hat{s}) + d_0(s), \tag{11}$$

where $c_2 > 0$ and $d_0(\cdot)$ is an arbitrary function.

(ii) If $I(S;\hat{S}) = 0$, the second condition of Lemma 2 is satisfied for any function $d(s, \hat{s})$.

This theorem should be understood by complete analogy to Theorem 3. That is, let $q_{\hat{S}|S}$ be the conditional distribution induced by some channel conditional distribution through the encoder f and the decoder g. For any distortion measure d, an average distortion $\Delta = E_{q_{\hat{S}|S}} d(S, \hat{S})$ can be computed, which implies a set of alternative conditional distributions that also yield distortion Δ. The key is to find d in such a way that the chosen $q_{\hat{S}|S}$ minimizes $I(S;\hat{S})$ among all conditional distributions in the set. This argument is made precise in the following proof.

Proof. To simplify the notation, we will use the symbol W in place of $p_{\hat{S}|S}$ in the proof. Define

$$I'_W(s,\hat{s}) \;=\; \log_2 \frac{W(\hat{s}|s)}{p_{\hat{S}}(\hat{s})}, \tag{12}$$

where $p_{\hat{S}}$ is the marginal distribution of $\hat{S}$.

In particular, note that with this definition, $I_W(S;\hat{S}) = \langle p_S W, I'_W\rangle$, where with slight abuse of notation, we have used $\langle p_S W, I'_W\rangle$ to mean $\int\int p_S(s)W(\hat{s}|s)I'_W(s,\hat{s})ds d\hat{s}$. In the proof, we use the following auxiliary lemma:

Lemma: For any W and $\tilde{W}$, $I_{\tilde{W}}(S;\hat{S}) - I_W(S;\hat{S}) \geq \langle p_S\tilde{W} - p_S W, I'_W\rangle$.

Using the fact that $I_W(S;\hat{S}) = \langle p_S W, I'_W\rangle$, we consider

$$\begin{aligned} I_{\tilde{W}}(S;\hat{S}) - \langle p_S\tilde{W}, I'_W\rangle &= \langle p_S\tilde{W}, \log_2 \frac{\tilde{W}}{\tilde{p}_{\hat{S}}}\rangle - \langle p_S\tilde{W}, \log_2 \frac{W}{p_{\hat{S}}}\rangle \\ &= \langle p_S\tilde{W}, \log_2 \frac{\tilde{V}}{p_S}\rangle - \langle p_S\tilde{W}, \log_2 \frac{V}{p_S}\rangle \\ &= \langle p_{\hat{S}}\tilde{V}, \log_2 \frac{\tilde{V}}{V}\rangle \;=\; \langle p_{\hat{S}}, D(\tilde{V}||V)\rangle \;\geq\; 0, \end{aligned} \tag{13}$$

where we have used V to denote the conditional distribution of S given $\hat{S}$ under W, i.e. $V(s|\hat{s}) = W(\hat{s}|s)p(s)/p(\hat{s})$, and correspondingly $\tilde{V}$ to denote the same distribution, but under $\tilde{W}$, i.e. $\tilde{V}(s|\hat{s}) = \tilde{W}(\hat{s}|s)p(s)/\tilde{p}(\hat{s})$. $D(\tilde{V}||V)$ denotes the Kullback-Leibler distance between $\tilde{V}$ and V in the variable s, hence it is a function of $\hat{s}$. The last inner product is thus one-dimensional in the variable $\hat{s}$. The inequality follows from the fact that the Kullback-Leibler distance is nonnegative.

With this, we are ready to prove the theorem.

($\Leftarrow$.) (Sufficiency of the formula.) Let d be arbitrary, let $\tilde{W}$ be an arbitrary conditional distribution such that

$$E_{p_S\tilde{W}}d(S,\hat{S}) \;\leq\; E_{p_S W}d(S,\hat{S}). \tag{14}$$

For any $\lambda > 0$,

$$\begin{aligned} I_{\tilde{W}}(S;\hat{S}) - I_W(S;\hat{S}) &\geq \langle p_S\tilde{W} - p_S W, I'_W(s,\hat{s})\rangle \\ &\geq \langle p_S\tilde{W} - p_S W, I'_W + \lambda d\rangle, \end{aligned} \tag{15}$$

where the first inequality is the last lemma, and the second follows by assumption on $\tilde{W}$. If $\lambda d(s,\hat{s}) = -I'_W(s,\hat{s}) + \tilde{d}_0(s)$, then the last expression is zero, proving that $I_W(S;\hat{S})$ indeed minimizes mutual information. Setting $\tilde{d}_0(s) = -\log_2 p(s) + \lambda d_0(s)$ gives the claimed formula (11).

When $I_W(S;\hat{S}) = 0$, then trivially W achieves the minimum mutual information $I(S;\hat{S})$ over all $\tilde{W}$ that satisfy $E_{\tilde{W}} d(S,\hat{S}) \leq E_W d(S,\hat{S})$, regardless of d.

($\Rightarrow$.) (Necessity of the formula.) In order to establish the necessity of formula (11), we need the derivative of the mutual information $I(S;\hat{S})$ with respect to the conditional distribution of $\hat{S}$ given S. In the finite-dimensional case, we can argue as follows: As long as $I(S;\hat{S}) > 0$, we are in the *strictly* decreasing region of $R(D)$, and therefore

$$W \in \arg \min_{\tilde{W}: Ed(S,\hat{S}) \leq E_{p_S W} d(S,\hat{S})} I(S;\hat{S}) \implies W \in \arg \min_{\tilde{W}: Ed(S,\hat{S}) = E_{p_S W} d(S,\hat{S})} I(S;\hat{S}).$$

But a first-order necessary condition for W to be a minimizer in the latter minimization problem is that the corresponding Lagrange functional, evaluated at W, vanishes. That is,

$$\frac{d}{d\tilde{W}(\hat{s}|s)} L^{(d)}(\tilde{W}, \lambda, \mu_0, \mu_1, \dots) \Big|_{\tilde{W}=W} = 0, \tag{16}$$

for every pair $(s,\hat{s}) \in \mathcal{S} \times \hat{\mathcal{S}}$, where

$$\begin{aligned} L^{(d)}(\tilde{W}, \lambda, \mu_0, \mu_1, \dots) &= \\ &= I_{\tilde{W}}(S;\hat{S}) - \lambda(\sum_r \sum_z p_S(r)\tilde{W}(z|r)d(r,z) - \Delta) - \sum_r \mu_r (\sum_t \tilde{W}(t|r) - 1). \end{aligned} \tag{17}$$

By evaluating the derivatives, this indeed implies the claimed formula (11), as long as $\lambda \neq 0$. When does $\lambda = 0$ occur? Writing out the formula for λ,

$$\lambda = \frac{1}{d(s,\hat{s}) p_S(s)} \left(\frac{d}{dW(\hat{s}|s)} I(S;\hat{S}) - \mu_s \right), \tag{18}$$

we see that $\lambda = 0$ if and only if $\frac{d}{dW(\hat{s}|s)} I(S;\hat{S}) = \mu_s$, for all $\hat{s}$. But this can only arise if $I(S;\hat{S}) = 0$.

Again, this argument may be extended to infinite dimensions by considering the Gateaux differential of $I_W(S;\hat{S})$ and using Thm. 2, p. 188, in Luenberger [7]. □

Theorems 3 and 4 correspond to the first and second requirement in Lemma 2. The third requirement of that lemma is $I(X;Y) = I(S;\hat{S})$. This condition rules out certain codes (f,g). Essentially, the code must be a deterministic mapping; if on the contrary, the code adds randomness, it acts like a cascaded channel, and hence, $I(X;Y)$ and $I(S;\hat{S})$ will not be equal. The following sufficient (but not necessary) condition ensures that $I(X;Y) = I(S;\hat{S})$.

Lemma 5. *If the encoder is a deterministic mapping and the inverse of the decoder is also a deterministic mapping,*[2] *then $I(X;Y) = I(S;\hat{S})$, i.e., the third condition of Lemma 2 is satisfied.*

[2] That is, for every $\hat{s}$ with $p(\hat{s}) > 0$, there is exactly one y such that $\hat{s} = g(y)$.

Proof. For the encoder, consider

$$\begin{aligned} I(S,X;Y) &= I(S;Y)+I(X;Y|S) \\ &= I(X;Y)+I(S;Y|X), \end{aligned} \tag{19}$$

where $I(S;Y|X)=0$ since $S \to X \to Y$ is a Markov chain, and hence $I(X;Y) = I(S;Y) + I(X;Y|S)$. If the encoder is deterministic, then $H(X|S)=0$ and hence $I(X;Y)=I(S;Y)$. (Note however that in certain cases, $H(X|S)=H(X|Y,S)\neq 0$; hence it is not in general a necessary condition that the encoder be deterministic.)

To complete the proof, we show that $I(S;\hat{S})=I(S;Y)$. Consider

$$\begin{aligned} I(S;Y,\hat{S}) &= I(S;\hat{S})+I(S;Y|\hat{S}) \\ &= I(S;Y)+I(S;\hat{S}|Y), \end{aligned} \tag{20}$$

where $I(S;\hat{S}|Y)=0$ since $S \to Y \to \hat{S}$ is a Markov chain. But if the inverse of the decoder is deterministic, then $H(Y|\hat{S})=0$ and hence $I(S;Y)=I(S;\hat{S})$. □

In summary, our discussion of the requirement $R(\Delta)=C(\Gamma)$ produced a set of explicitly verifiable conditions that together ensure $R(\Delta)=C(\Gamma)$. However, to obtain an explicit criterion that can establish the optimality of a single-letter code, it still remains to scrutinize the second requirement of Lemma 1. This is the goal of the next section.

3.2 Condition *(ii)* of Lemma 1

Lemma 1 contains two simultaneous requirements to ensure the optimality of a communication system that employs single-letter codes. The first requirement, $R(\Delta)=C(\Gamma)$, is studied and developed in detail in Section 3.1; in this section, we examine the second condition, namely when it is impossible to lower Δ without changing $R(\Delta)$, and when it is impossible to lower Γ without changing $C(\Gamma)$. This permits to give a general criterion to establish the optimality of any communication system that uses single-letter codes.

The crux of the problem is illustrated in Figure 2. It shows simultaneously the capacity-cost function of the channel (left) and the rate-distortion function of the source (right). Problematic cases may only occur if either $R(\cdot)$ or $C(\cdot)$ are horizontal, i.e. when they have reached their asymptotic values $R(D\to\infty)$ and $C(P\to\infty)$. This only happens when the mutual information is zero or C_0. For example, both the cost-distortion pair (Γ_1,Δ) and the cost-distortion pair (Γ_2,Δ) satisfy the condition $R(\Delta)=C(\Gamma)$; however, only the pair (Γ_2,Δ) corresponds to an optimal transmission strategy. By analogy, an example can be given involving two different distortions. A concrete example of a system where the condition $R(\Delta)=C(\Gamma)$ is not sufficient is given in Appendix A.

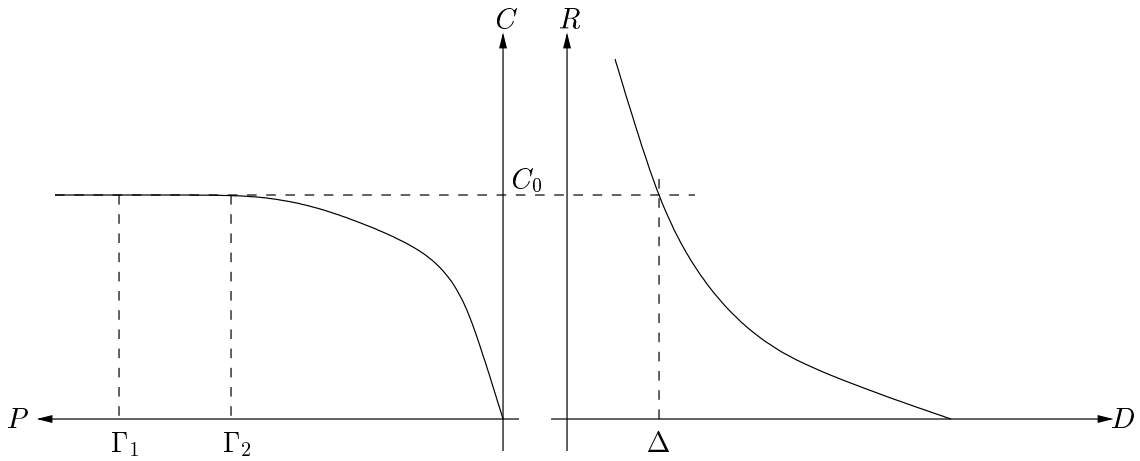

Figure 2: When $R(\Delta) = C(\Gamma)$ is not sufficient to guarantee optimality.

Continuing in this line of thought, we obtain the following proposition.

Proposition 6. *Suppose that the transmission of the source (p_S, d) across the channel $(p_{Y|X}, \rho)$ using the single-letter code (f, g) satisfies $R(\Delta) = C(\Gamma)$. Then,*

(i) *Γ cannot be lowered without changing $C(\Gamma)$ if and only if one of the following two conditions is satisfied:*

  (a) *$I(X;Y) < C_0$, or*

  (b) *$I(X;Y) = C_0$ and among the distributions that achieve C_0, p_X belongs to the ones with lowest cost. In particular, the last condition is trivially satisfied whenever p_X is the unique channel input distribution achieving C_0.*

(ii) *Δ cannot be lowered without changing $R(\Delta)$ if and only if one of the following two conditions is satisfied:*

  (a) *$I(S;\hat{S}) > 0$, or*

  (b) *$I(S;\hat{S}) = 0$ and among the conditional distributions for which $I(S;\hat{S}) = 0$, $p_{\hat{S}|S}$ belongs to the ones with lowest distortion. In particular, the last condition is trivially satisfied if $p_{\hat{S}|S}$ is the unique conditional distribution achieving $I(S;\hat{S}) = 0$.*

Proof. Part *(i)*: To see that condition *(a)* is sufficient, define $\Gamma_{max} = \min\{P : C(P) = C_0\}$. For every $\Gamma < \Gamma_{max}$, the value $C(\Gamma)$ uniquely specifies Γ. This follows from the fact that $C(\cdot)$ is convex and nondecreasing. From Lemma 2, $R(\Delta) = C(\Gamma)$ implies $C(\Gamma) = I(X;Y)$. Hence, $I(X;Y) < C_0$ implies $C(\Gamma) < C_0$, which in turn implies that it is not possible to

change Γ without changing $C(\Gamma)$. To see that condition *(b)* is sufficient, note that if among the achievers of C_0, p_X belongs to the ones with lowest cost, then it is indeed impossible to lower Γ without changing $C(\Gamma)$. In particular, if p_X is the only achiever of C_0, then there cannot be another p_X that achieves the same rate, namely C_0, but with smaller cost, simply because there is no other p_X that achieves C_0.

It remains to show that if neither *(a)* nor *(b)* are satisfied, then Γ can indeed be lowered. In that case, $I(X;Y) = C_0$ (it cannot be larger than C_0). Moreover, there must be multiple achievers of C_0, and p_X is not the one minimizing Γ. In other words, Γ can indeed be lowered without changing $C(\Gamma) = C_0$.

The proof of part *(ii)* of the proposition goes along the same lines. To see that condition *(a)* is sufficient, define $\Delta_{max} = \min\{D : R(D) = 0\}$. For every $\Delta < \Delta_{max}$, the value $R(\Delta)$ uniquely specifies Δ. This follows from the fact that $R(\cdot)$ is convex and non-increasing. From Lemma 2, $R(\Delta) = C(\Gamma)$ implies $R(\Delta) = I(S;\hat{S})$. Hence, $0 < I(S;\hat{S})$ implies $0 < R(\Delta)$, which in turn implies that it is not possible to change Δ without changing $R(\Delta)$. For condition *(b)*, note that if among the achievers of zero mutual information, $p_{\hat{S}|S}$ belongs to the ones with lowest distortion, then it is indeed impossible to lower Δ without changing $R(\Delta)$. In particular, if $p_{\hat{S}|S}$ is the unique conditional distribution achieving zero mutual information, then there may not be another conditional distribution achieving the same rate (zero) but with smaller distortion, simply because by assumption, there is no other conditional distribution achieving zero mutual information.

It remains to show that if neither *(a)* nor *(b)* are satisfied, then Δ can indeed be lowered. In that case, $I(S;\hat{S}) = 0$ (it cannot be smaller than 0). Moreover, there must be multiple achievers of zero mutual information, and $p_{\hat{S}|S}$ does not minimize the distortion among them. In other words, Δ can indeed be lowered without changing $R(\Delta) = 0$. □

Remark. In the most general case of Proposition 6, it is necessary to specify the cost function and the distortion measure before the conditions can be verified. Let us point out, however, that in many cases of practical interest, this is *not* necessary. In particular, if $I(X;Y) < C_0$, or if $I(X;Y) = C_0$ but p_X is the unique distribution that achieves C_0, then Part *(i)* is satisfied irrespective of the choice of the cost function. By analogy, if $0 < I(S;\hat{S})$, or if $I(S;\hat{S}) = 0$ but $p_{\hat{S}|S}$ is the unique conditional distribution for which $I(S;\hat{S}) = 0$, then Part *(ii)* is satisfied irrespective of the choice of the distortion measure.

In summary, our discussion of Condition *(ii)* of Lemma 1 supplied a set of explicitly verifiable criteria. The main result of this paper is obtained by combining this with the results of Section 3.1.

3.3 The Main Result

The main result of this paper is a simple criterion to check whether a given single-letter code performs optimally for a given source/channel pair. Lemma 1 showed that on the one hand, the system has to satisfy $R(\Delta) = C(\Gamma)$. The choice of the cost function ρ as in Theorem 3 ensures that the channel input distribution achieves capacity. Similarly, the choice of the distortion measure according to Theorem 4 ensures that the conditional distribution of $\hat{S}$ given S achieves the rate-distortion function of the source. Together with the condition that $I(S;\hat{S}) = I(X;Y)$, this ensures that $R(\Delta) = C(\Gamma)$. But Lemma 1 required on the other hand that Γ may not be lowered without changing $C(\Gamma)$, and that Δ may not be lowered without changing $R(\Delta)$. Recall that this is *not* ensured by Theorems 3 and 4. Rather, it was discussed in Section 3.2 and led to Proposition 6. It is now a simple matter to combine the insight gained in the latter proposition with the statements from Theorems 3 and 4. This leads to a quite simple criterion to establish the optimality of a large class of communication systems that employ single-letter codes:

Theorem 7. *Consider the transmission of the source (p_S, d) across the channel $(p_{Y|X}, \rho)$ using the single-letter code (f, g). The following statements hold:*

(o) *If $I(S;\hat{S}) \neq I(X;Y)$, then the system does not perform optimally.*

(i) *If $0 < I(S;\hat{S}) = I(X;Y) < C_0$, the system is optimal if and only if $\rho(x)$ and $d(s,\hat{s})$ are chosen according to Theorems 3 and 4, respectively.*

(ii) *If $0 < I(S;\hat{S}) = I(X;Y) = C_0$, the system is optimal if and only if $d(s,\hat{s})$ is chosen according to Theorem 4, and $\rho(x)$ is such that $E\rho(X) \leq E_{\tilde{p}_X}\rho(X)$ for all other achievers $\tilde{p}_X$ of C_0. In particular, the last condition is trivially satisfied if p_X is the unique channel input distribution achieving C_0.*

(iii) *If $0 = I(S;\hat{S}) = I(X;Y) < C_0$, the system is optimal if and only if $\rho(x)$ is chosen according to Theorem 3, and $d(s,\hat{s})$ is such that $Ed(S,\hat{S}) \leq E_{\tilde{p}_{\hat{S}|S}} d(S,\hat{S})$ for all other achievers $\tilde{p}_{\hat{S}|S}$ of $I(S;\hat{S}) = 0$. In particular, the last condition is trivially satisfied if $p_{\hat{S}|S}$ is the unique conditional distribution for which $I(S;\hat{S}) = 0$.*

(iv) *If $C_0 = 0$, then the system is optimal if and only if $E\rho(X) \leq E_{\tilde{p}_X}\rho(X)$ for all channel input distributions $\tilde{p}_X$, and $Ed(S,\hat{S}) \leq E_{\tilde{p}_{\hat{S}|S}} d(S,\hat{S})$ for all conditional distributions $\tilde{p}_{\hat{S}|S}$.*

Proof. Part *(o)*. From the Data Processing Theorem (e.g. [8, Thm. 2.8.1]), $I(S;\hat{S}) \neq I(X;Y)$ implies $I(S;\hat{S}) < I(X;Y)$. Moreover, $I(S;\hat{S}) < I(X;Y)$ implies $R(\Delta) < C(\Gamma)$ (see

also the proof of Lemma 2). But then, by Lemma 1, the system does not perform optimally.
Part *(i)*. If $0 < I(S;\hat{S})$ and $I(X;Y) < C_0$, the system is optimal *if and only if* $R(\Delta) = C(\Gamma)$ (Lemma 1 with Proposition 6). We have shown that this is equivalent to requiring the three conditions of Lemma 2 to be satisfied. The third of these conditions, $I(S;\hat{S}) = I(X;Y)$, is satisfied by assumption. As long as $0 < I(S;\hat{S})$ and $I(X;Y) < C_0$, Theorems 3 and 4 establish that the first two are satisfied *if and only if* ρ and d are chosen according to Formulae (3) and (11), respectively.
Part *(ii)*. If $I(X;Y) = C_0$, the system is optimal *if and only if* $R(\Delta) = C(\Gamma)$ and among the achievers of C_0, p_X belongs to the ones with lowest cost (Lemma 1 with Proposition 6). The condition $R(\Delta) = C(\Gamma)$ is satisfied *if and only if* the three conditions of Lemma 2 are satisfied. The third of these conditions, $I(S;\hat{S}) = I(X;Y)$, is satisfied by assumption. When $0 < I(S;\hat{S})$ but $I(X;Y) = C_0$, Theorems 3 and 4 establish that the first two are satisfied *if and only if* d is chosen according to Formula (11).
Part *(iii)*. If $0 = I(S;\hat{S})$, the system optimal *if and only if* $R(\Delta) = C(\Gamma)$ and among the conditional distributions for which $I(S;\hat{S}) = 0$, $p_{\hat{S}|S}$ belongs to the ones with lowest distortion (Lemma 1 with Proposition 6). The condition $R(\Delta) = C(\Gamma)$ is satisfied *if and only if* three conditions of Lemma 2 are satisfied. The third of these conditions, $I(S;\hat{S}) = I(X;Y)$, is satisfied by assumption. When $I(X;Y) < C_0$ but $I(S;\hat{S}) = 0$, Theorems 3 and 4 establish that the first two are satisfied *if and only if* ρ is chosen according to Formula (3).
Part *(iv)* has been added for completeness only. It should be clear that if $C_0 = 0$, then automatically, all the mutual information conditions are satisfied since all mutual informations must be zero, and all that has to be checked is that the cost and the distortion are minimal. Obviously, this case is of limited practical interest. □

As pointed out earlier, one attractive issue with this criterion is that it permits to construct an arbitrarily large supply of examples. The next section illustrates this point.

4 Examples

Example 1 (Gaussian). The goal of this example is to illustrate and verify the findings of Section 3 for the Gaussian example. Let the (memoryless) source be zero-mean Gaussian of variance σ_S^2 with distortion measure $d(s,\hat{s}) = (s-\hat{s})^2$ (i.e. mean-square error). Let the (memoryless) channel be an additive noise channel, where the noise process, denoted by Z, is zero-mean Gaussian of variance σ^2, and the input cost function is $\rho(x) = x^2$. Finally, let the code be

$$f(s) \;=\; \sqrt{\frac{P}{\sigma_S^2}}\, s \;=\; \alpha s \tag{21}$$

and

$$g(y) \;=\; \sqrt{\frac{\sigma_S^2}{P}}\frac{P}{P+\sigma^2}\; y \;=\; \beta y. \tag{22}$$

This setup appears in several places in the literature, e.g. in [3].

The cost-distortion pair is found to be $\Gamma = P$ and $\Delta = \sigma_S^2\sigma^2/(P+\sigma^2)$. Using the rate-distortion and the capacity-cost functions, we can directly verify that Lemma 1 is satisfied. For the AWGN channel, the capacity-cost function is $C(P) = 1/2\log_2(1+P/\sigma^2)$, and for the iid Gaussian source of variance σ_S^2, the rate-distortion function is $R(D) = 1/2\log_2(\sigma_S^2/D)$ (see e.g. [8]). Plugging in the cost-distortion pair (Γ,Δ) as found above, we find both times $1/2\log_2(1+P/\sigma^2)$, confirming that the first condition of Lemma 1 is satisfied. Moreover, since $0 < I(S;\hat{S})$ and $I(X;Y) < C_0$, Proposition 6 implies that neither Γ nor Δ can be decreased (leaving the other fixed). Therefore, we conclude that the purported communications scheme performs optimally.

Let us apply Theorem 7 to this scenario. It is clear that $I(X;Y) = I(S;\hat{S})$ since both $f(\cdot)$ and $g(\cdot)$ are bijective maps. Moreover, since C_0 is infinite for the AWGN channel, Case *(i)* of Theorem 7 applies. Hence, $\rho(\cdot)$ and $d(\cdot,\cdot)$ have to be selected as in the formulae of Theorems 3 and 4, respectively. For the cost function, we first determine

$$\begin{aligned} D(p_{Y|X}(\cdot|x)||p_Y(\cdot)) &= D\left(p_Z\left(\frac{\cdot-\alpha x}{\alpha}\right)\,\middle\|\, p_Y\right) \\ &= -h\left(p_Z\left(\frac{\cdot-\alpha x}{\alpha}\right)\right) - \int p_Z\left(\frac{y-\alpha x}{\alpha}\right)\log_2 p_Y(y)dy. \end{aligned} \tag{23}$$

Since the entropy of a Gaussian is independent of its mean, the first term is a constant, say a. Hence,

$$\begin{aligned} &D(p_{Y|X}(\cdot|x)||p_Y(\cdot)) = \\ &= a - \int \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(y-\alpha x)^2}{2\alpha^2\sigma^2}}\left(\log_2\frac{1}{\sqrt{2\pi\alpha^2(\sigma_S^2+\sigma^2)}} - \frac{y^2}{2\alpha^2(\sigma_S^2+\sigma^2)}\right)dy \\ &= a_1 + a_2\int \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(y-\alpha x)^2}{2\alpha^2\sigma^2}}y^2dy = a_1 + a_2(\alpha^2\sigma^2 + (\alpha x)^2) = b_1x^2 + b_2, \end{aligned} \tag{24}$$

where the a_i and b_i are appropriate constants. Since the formula of Theorem 3 only specifies $\rho(x)$ up to an affine transform, their precise value is irrelevant. For example, by choosing (in Theorem 3) $c_1 = 1/b_1$ and $\rho_0 = -b_2/b_1$, Eqn. (3) reads $\rho(x) = x^2$. For the distortion measure, we have to determine $p(s|\hat{s})$. Since both f and g are linear mappings, $p(\hat{s}|s)$ is found to be

$$p(\hat{s}|s) \;=\; \frac{1}{\beta}p_{Y|X}\left(\hat{s}/\beta\,|\,\alpha s\right) \;=\; \frac{1}{\sqrt{2\pi}\sigma\beta}e^{-\frac{1}{2\sigma^2\beta^2}(\hat{s}-\alpha\beta s)^2}. \tag{25}$$

The marginal of $\hat{S}$ can be determined by recalling that Y is Gaussian with variance $P+\sigma^2$. Hence, $\hat{S}$ is Gaussian with variance $\beta^2(P+\sigma^2)$. Plugging in, we find

$$\log_2 \frac{p(\hat{s}|s)}{p(\hat{s})} = \log_2 \frac{\beta\sqrt{P+\sigma^2}}{\sigma} e^{-\frac{1}{2\sigma^2\beta^2}(\hat{s}-\alpha\beta s)^2+\frac{1}{2\beta^2(P+\sigma^2)}\hat{s}^2} \tag{26}$$

which gives (by defining c_2 and $d_0(s)$ in Theorem 4 appropriately)

$$d(s,\hat{s}) = c_2\left(\hat{s} - \frac{\alpha\beta(P+\sigma^2)}{P}\, s\right)^2 + d_0(s). \tag{27}$$

It is quickly verified that plugging in the definitions of α and β yields the standard mean-square error distortion. Hence, Theorem 7 allows to conclude that the suggested communications scheme performs optimally.

As a side note, suppose that the coefficients α and β are chosen differently, which means that the single-letter code is "mismatched." Then, the above derivation shows that the code performs optimally with respect to a "weighted" MSE distortion, with weighting as given by the last equation.

Example 2 (binary). Let the source be binary and uniform with Hamming distortion measure, and let the channel be binary and symmetric (with $\epsilon < 1/2$) without an input cost constraint (i.e. $\rho(x) = const.$, $\forall x$). Let f and g be the identity maps, i.e. $f(s) = s$ and $g(y) = y$. This setup is also considered in e.g. in [4] and [5].

For this channel, the capacity is $C(\Gamma) = C_0 = 1 - H_b(\epsilon)$, where $H_b(\cdot)$ denotes the binary entropy function. The rate-distortion function for the binary source is $R(D) = 1 - H_b(D)$ (see e.g. [8]). In the present example, the distortion is found to be $\Delta = Ed(S,\hat{S}) = \epsilon$, from which $R(\Delta) = 1 - H_b(\epsilon)$. Thus, $R(\Delta) = C(\Gamma)$ is satisfied. For $\epsilon < 1/2$, there is a unique achiever of C_0, and hence, from Proposition 6, neither Δ nor Γ can be decreased (leaving the other fixed). Thus, by Lemma 1, the considered communications scheme performs optimally.

Let us establish the same fact using Theorem 7. Trivially, $I(X;Y) = I(S;\hat{S})$, and we find

$$\frac{p_{\hat{S}|S}(\hat{s}|s)}{p_{\hat{S}}(\hat{s})} = \frac{p_{Y|X}(\hat{s}|s)}{p_Y(\hat{s})} = \frac{1}{2}\, p_{Y|X}(\hat{s}|s) = \begin{cases} \frac{1}{2}(1-\epsilon), & \text{if } \hat{s} = s, \\ \frac{1}{2}\epsilon, & \text{otherwise.} \end{cases} \tag{28}$$

Taking $d_0(s) = \frac{\log_2(1-\epsilon)/2}{\log_2(1-\epsilon)/\epsilon}$ and $c_2 = \frac{1}{\log_2(1-\epsilon)/\epsilon}$ reveals that one of the distortion measures that satisfy the requirement in Theorem 7 is indeed the Hamming distance.

Example 3 (Laplacian). This example studies the transmission of a Laplacian source across an additive white Laplacian noise (AWLN) channel, defined as follows:

$$p_S(s) = \frac{\alpha_0}{2} e^{-\alpha_0|s|} \tag{29}$$

$$p_{Y|X}(y|x) = \frac{\alpha}{2} e^{-\alpha|y-x|}. \tag{30}$$

We also use $Z = Y - X$ to denote the additive noise. Hence, Z is Laplacian with parameter α. Assume that $\alpha_0 < \alpha$ (which implies $ES^2 > EZ^2$). Note that with trivial changes, the derivations can be altered for the case $\alpha_0 \geq \alpha$. Moreover, let the encoding and the decoding function be simply the identity (in other words, we consider uncoded transmission). The corresponding output distribution $p_Y(y)$ is found to be

$$p_Y(y) \quad = \quad \frac{\alpha_0 \alpha}{2} \frac{\alpha e^{-\alpha_0 |y|} - \alpha_0 e^{-\alpha |y|}}{\alpha^2 - \alpha_0^2}. \tag{31}$$

Since the channel is an independent additive noise channel, the formula in Theorem 3 can be rewritten as

$$\rho(x) \quad = \quad - \int_z p_Z(z) \log_2 p_Y(x+z) dz. \tag{32}$$

A numerical approximation to this is illustrated in Fig. 3 for a particular choice of the parameters: $\alpha_0 = 3$ and $\alpha = 9$, hence the signal-to-noise ratio in the example is $\alpha^2/\alpha_0^2 = 9$. Note that $\rho(s)$ as in Eqn. (32) is negative for some values of s. For the figure, we have added a suitable constant. The figure reveals that $\rho(s)$ is similar to the magnitude function (at least for our choice of the parameters). The next step is to compute the distortion measure

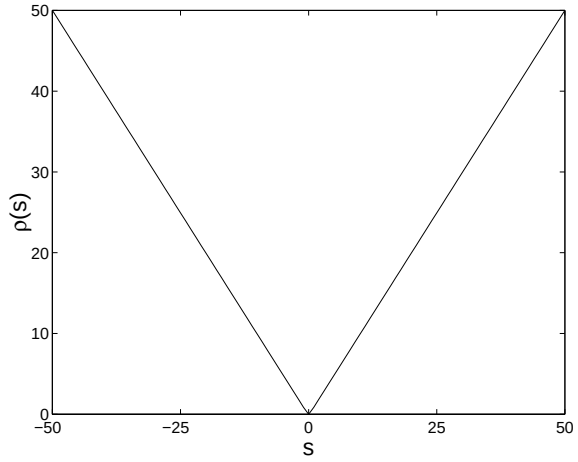

Figure 3: Channel input cost function $\rho(s)$ according to Eqn. (32).

that makes the system optimal. According to Theorem 4, we need to determine

$$-\log_2 p(s|\hat{s}) \quad = \quad -\log_2 \frac{p_{Y|X}(\hat{s}|s) p_S(s)}{p_Y(\hat{s})} \quad = \quad -\log_2 \frac{\alpha^2 - \alpha_0^2}{2} \frac{e^{-\alpha|\hat{s}-s| - \alpha_0 |s|}}{\alpha e^{-\alpha_0 |\hat{s}|} - \alpha_0 e^{-\alpha |\hat{s}|}}. \tag{33}$$

However, this function is negative for some $(s, \hat{s})$. To make it nonnegative, we add, for each s, an appropriate constant, namely the $\log_2$ of

$$\max_{\hat{s}} p(s|\hat{s}) \quad = \quad \frac{\alpha^2 - \alpha_0^2}{2} e^{-\alpha_0 |s|} \max_{\hat{s}} \frac{e^{-\alpha|\hat{s}-s|}}{\alpha e^{-\alpha_0 |\hat{s}|} - \alpha_0 e^{-\alpha |\hat{s}|}} = \frac{\alpha^2 - \alpha_0^2}{2} \frac{1}{\alpha - \alpha_0 e^{-(\alpha - \alpha_0)|s|}}.$$

Plugging in, we obtain

$$d(s,\hat{s}) \quad = \quad |\hat{s}-s| + \frac{1}{\alpha}\log_2 \frac{\alpha e^{-\alpha_0|\hat{s}|} - \alpha_0 e^{-\alpha|\hat{s}|}}{\alpha e^{-\alpha_0|s|} - \alpha_0 e^{-\alpha|s|}}. \tag{34}$$

This is illustrated in Fig. 4 for the above choice of the parameters ($\alpha_0 = 3$ and $\alpha = 9$).[3] To conclude this example, let us point out that there is no straightforward answer to the question whether this distortion measure is practically meaningful. To judge on that, the physical objectives have to be taken into consideration.

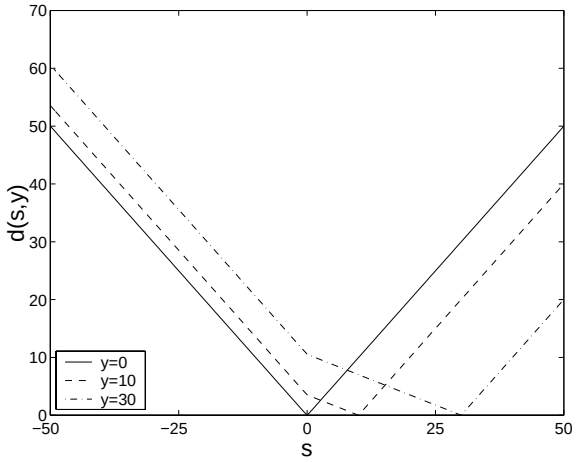

Figure 4: Distortion measure $d(s,y)$ according to Eqn. (34) for fixed y.

Remark. From the last example, it should be clear that using the formulae given in Lemmata 3 and 4, an arbitrary supply of examples can be constructed that feature the same optimal behavior as the well-known example of a Gaussian source over a Gaussian channel.

5 Some Applications Of The Theory

5.1 Existence Of Single-Letter Codes With Optimal Performance

In Section 3, we characterized the relationship between source, channel and code such that the corresponding communication system performs optimally. The result can be applied directly if the source distribution, the channel distribution and the code are fixed. In this section, we fix the source (p_S, d) and the channel $(p_{Y|X}, \rho)$. What conditions have to satisfied such that there *exists* a single-letter code (f, g) that makes the overall system an optimal transmission scheme? We were able to find partial answers to this question.

[3]Note that the figure does not exactly depict Eqn. (34); rather, additive and multiplicative constants have been selected to get a clearer picture.

For the case $\mathcal{S} = \mathcal{X}$ and $\mathcal{Y} = \hat{\mathcal{S}}$, the answer can be phrased as follows: Given (p_S, d) and $(p_{Y|X}, \rho)$, assume that the encoder and the decoder are the identity maps. Determine $\tilde{\rho}(s)$ and $\tilde{d}(s, y)$ according to the formulae of Theorems 3 and 4, respectively. Then, find a function $f(s)$ such that $\tilde{\rho}(s) = \rho(f(s))$. If this is feasible, then find a function $g(y)$ such that $\tilde{d}(s, y) = d(s, g(y))$. If this is also feasible, then the single-letter code (f, g) performs optimally.

This constructive way of determining the existence of single-letter codes that perform optimally does not seem to lead to a concise general answer to the question. In the sequel, we present answers for certain particular scenarios.

Lemma 8 (binary). *Let $\mathcal{S} = \mathcal{X} = \mathcal{Y} = \hat{\mathcal{S}} = \{0, 1\}$, $\rho(x) = const.$, and $d(s, \hat{s}) = 1$ if $s \neq \hat{s}$, and $d(s, \hat{s}) = 0$ otherwise (Hamming distortion). Suppose that the channel has nonzero capacity. Then, there exists a single-letter code with optimal performance if and only if the source pmf p_S is uniform and the channel conditional pmf $p_{Y|X}$ is symmetric.*

Remark. The case $C_0 = 0$ can be handled separately using Part*(iv)* of Theorem 7.

Proof. Assume that $X = S$ and $\hat{S} = Y$. This is without loss of generality, since the only two alternatives are *(i)* that the encoder permutes the source symbols, which is equivalent to swapping the channel transition probabilities (by the symmetry of the problem), and *(ii)* that the encoder maps both source symbols onto one channel input symbol, which is always suboptimal except when the channel has capacity zero. We will use the following notation: $\epsilon = p_{Y|X}(1|0)$, $\delta = p_{Y|X}(0|1)$, $p_X(x = 0) = \bar{\pi}$ and $p_X(x = 1) = \pi$. For the system to be optimal, since the channel is left unconstrained, it is necessary that $I(X; Y) = C_0$. Therefore, Case *(ii)* of Theorem 7 applies. Hence, it is *necessary* that $d(s, \hat{s})$ be chosen according to Eqn. (11); i.e., we require that $-\log_2 p(s|\hat{s}) = -\log_2 p(x|y)$ be equivalent to the Hamming distortion. This is the same as requiring that $p_{X|Y}(0|1) = p_{X|Y}(1|0)$. Expressing $p(x|y)$ as a function of $\epsilon, \delta, \bar{\pi}$ and π, the latter implies that $\pi = \sqrt{(\epsilon(1-\epsilon))/(\delta(1-\delta))}\bar{\pi}$. Since moreover, $\pi + \bar{\pi} = 1$, we find

$$\pi \;=\; \frac{1}{1 + \sqrt{(\delta(1-\delta))/(\epsilon(1-\epsilon))}}. \tag{35}$$

We show that for channel of nonzero capacity, this is the capacity-achieving distribution if and only if $\epsilon = \delta$, which completes the proof. The capacity-achieving π satisfies the following condition:

$$\frac{d}{d\pi} I(X; Y) \;=\; (\epsilon + \delta - 1) \log_2 \frac{1 - ((1-\pi)(1-\epsilon) + \pi\delta)}{(1-\pi)(1-\epsilon) + \pi\delta} + H_b(\epsilon) - H_b(\delta) \;=\; 0. \tag{36}$$

Plugging in π from above yields

$$2^{\frac{H_b(\delta) - H_b(\epsilon)}{1-\delta-\epsilon}} \;=\; \frac{(1-\epsilon)\sqrt{\delta(1-\delta)} + \delta\sqrt{\epsilon(1-\epsilon)}}{\epsilon\sqrt{\delta(1-\delta)} + (1-\delta)\sqrt{\epsilon(1-\epsilon)}}. \tag{37}$$

Clearly, equality holds if $\epsilon = \delta$ (and thus $\bar{\pi} = \pi$), but also if $\epsilon = 1 - \delta$. In the latter case, the channel has zero capacity. To see that there are no more values of ϵ and δ for which equality holds, fix (for instance) δ and consider the curves defined by the right side and the left side of Eqn. (37), respectively. The left side is convex and decreasing in ϵ. For $0 \leq \epsilon \leq 1 - \delta$, the right side is also convex and decreasing. Hence, at most 2 intersections can occur in this interval, and we already know them both. By continuing in this fashion, or by upper and lower bounds, one can establish that there are no more intersections. □

Lemma 9 (L-ary uniform). *Let $\mathcal{S}, \mathcal{X}, \mathcal{Y}$ and $\hat{\mathcal{S}}$ be L-ary, $\rho(x) = const.$, for all x, $d(s, \hat{s}) = 1$ if $s \neq \hat{s}$, and $d(s, \hat{s}) = 0$ otherwise (Hamming distortion), and p_S be uniform. Moreover, let the channel have nonzero capacity C_0. Then, there exists a single-letter code with optimal performance if and only if the channel conditional pmf is $p_{Y|X}(y|x) = const.$, for $y \neq x$ (or a permutation thereof).*

Proof. Pick an arbitrary channel conditional distribution $p_{Y|X}$ for which there exists a single-letter code (f, g) that makes the overall system optimal. From Lemma 2, this implies that $I(X;Y) = C(\Gamma)$. Since the channel is unconstrained here, $C(\Gamma) = C_0$. Therefore, Case *(ii)* of Theorem 7 applies. That is, to perform optimally, the distortion measure must be chosen as a scaled and shifted version of $-\log_2 p(s|\hat{s})$. But since by assumption, the distortion measure must be the Hamming distance, we must have that $-\log_2 p(s|\hat{s}) = c_2(1 - \delta(s - \hat{s})) + d_0(s)$, where $\delta(\cdot)$ denotes the Kronecker delta function (i.e. it is one if the argument is zero, and zero otherwise). Equivalently, $p(s|\hat{s})$ must satisfy

$$p(s|\hat{s}) = \begin{cases} 2^{-d_0(s)}, & s = \hat{s}, \\ 2^{-c_2 - d_0(s)}, & s \neq \hat{s}. \end{cases} \tag{38}$$

The L simultaneous equations $\sum_s p(s|\hat{s}) = 1$ imply a full-rank linear system of equations in the variables $2^{-d_0(s)}$, from which it immediately follows that $d_0(s) = const.$ But this means that $p(s|\hat{s})$ must satisfy

$$p(s|\hat{s}) = \begin{cases} \alpha, & s = \hat{s}, \\ \frac{1-\alpha}{L-1}, & s \neq \hat{s}. \end{cases} \tag{39}$$

By assumption, $p(s)$ is uniform, which implies that $p(\hat{s})$ is also uniform. But since all alphabets are of the same size, the condition that $I(S;\hat{S}) = I(X;Y)$ implies that $p(x)$ and $p(y)$ are also uniform, and that $p(x|y)$ is a permutation of

$$p(x|y) = \begin{cases} \alpha, & y = x, \\ \frac{1-\alpha}{L-1}, & y \neq x. \end{cases} \tag{40}$$

But this implies that the channel $p(y|x)$ has to be symmetric with $p(y|x) = \alpha$ for $y = x$, and $p(y|x) = (1 - \alpha)/(L - 1)$ for $y \neq x$, or a permutation thereof. □

There is a nice intuition going along with the last result: Suppose that the channel is symmetric ([8, p. 190]) and that the probabilities of erroneous transition are $\{\epsilon_1, \ldots, \epsilon_{L-1}\}$ for every channel input. The distortion achieved by uncoded transmission is simply the sum of these probabilities. However, the distortion achieved by coded transmission depends on the capacity of the channel. Therefore, if uncoded transmission should have a chance to be optimal, we have to minimize the capacity of the channel subject to a fixed sum $\sum_{i=1}^{L-1} \epsilon_i$. But this is equivalent to maximizing the entropy of the "noise" $Z = Y - X$ subject to a fixed probability $p_Z(z = 0)$. Clearly, this maximum occurs when all the ϵ_i are equal.

Remark. Suppose that all alphabets are the real numbers, the distortion measure is the mean-square error and the input cost function is the square. Under these constraints, we believe that the only discrete-time memoryless source/channel pairs for which there exists a single-letter code that performs optimally consist of an iid Gaussian source and an AWGN channel.

5.2 Source/Channel Codes Of Finite Block Length

A natural extension of the analysis performed in this paper is the quest for source/channel codes of (finite) block length M that perform optimally. More precisely, attention shall still be restricted to discrete-time memoryless sources and channels as defined in Definitions 1 and 2, but the code is now of (finite) length M: it maps M source symbols onto M channel symbols,[4] using an arbitrary function. One of the interesting questions is the following: for a given memoryless source (p_S, d) and a given memoryless channel $(p_{Y|X}, \rho)$, is there a source/channel code of finite block length M with optimal performance?

Suppose that all alphabets are discrete, and consider the length-M extension source and channel. These extensions are also discrete, but for them, the M-letter code is a single-letter code, and hence we can use Theorems 3 and 4 to give the cost function and the distortion measure on length-M blocks that are necessary for optimal performance. However, the underlying source and channel are *memoryless*. Therefore, by definition, it must be possible to express the cost function on length-M blocks as a sum of M individual terms, and the same must be true for the distortion measure. This excludes certain M-letter codes. Our conjecture is that a finite-length code with optimal performance exists if and only if there exists also a single-letter code with optimal performance for the same source/channel pair. Here, we prove this conjecture under some additional assumptions:

[4] Clearly, a more general extension would be to study codes that map N source symbols onto M channel symbols. We do not have results for that case yet.

Theorem 10. *Let (p_S, d) and $(p_{Y|X}, \rho)$ be a discrete memoryless source and a discrete memoryless channel, respectively. Suppose that all alphabets are of the same size, that $p(s) > 0$ for all $s \in \mathcal{S}$, that the distortion measure has the property that the matrix $\{2^{-d(s,\hat{s})}\}_{s,\hat{s}}$ is invertible and that the channel transition probability matrix is invertible. Then, there exists a source/channel code of finite block length that performs optimally if and only if, for the same source/channel pair, there exists also a single-letter source/channel code that performs optimally.*

Proof. See Appendix B. □

Among the restrictions imposed by the last theorem, the one on the distortion measure may seem somewhat unusual. Note however that the standard distortion measures like the Hamming distance and the squared-error distortion satisfy that restriction. In fact, any distortion measure under which the mapping $T(s) = \arg\min_{\hat{s}} d(s, \hat{s})$ is one-to-one satisfies the requirement.

5.3 Universality Of Single-Letter Source/Channel Codes

Optimal transmission systems designed according to the separation principle may be quite sensitive to parameter mismatch. Suppose e.g. that the capacity of the channel turns out to be smaller than the rate of the channel code that is used. The effect of this parameter mismatch on the final reconstruction of the data may be catastrophic.

Single-letter source/channel codes feature a graceful degradation as a function of mismatched parameters. In fact, in some cases, one and the same single-letter code achieves *optimal* performance for *multiple* source/channel pairs. In this sense, single-letter codes have a certain universality property. The following example illustrates this.

Example 1, continued (fading). Let the source be the Gaussian source from Example 1. The channel is slightly different from Example 1: It adds white Gaussian noise of variance σ_i^2 and scales the resulting signal by $P/(P + \sigma_i^2)$, but the value of σ_i^2 varies during transmission. The channel input signal X has to satisfy $EX^2 \leq P$. Take as the encoder a scaling by $\sqrt{P/\sigma_S^2}$ and as the decoder a scaling by $\sqrt{\sigma_S^2/P}$. From Example 1, it is clear that this code performs optimally irrespective of the value of σ_i^2.

In this example, the suggested code is universal for the transmission of a Gaussian source across any one out of an entire class of channels. In the spirit of the example, we introduce the following definition:

Definition 6 (universality). The single-letter code (f, g) is called universal for the source (p_S, d) and the class of channels given by $\mathcal{W} = \{(p_{Y|X}^{(0)}, \rho^{(0)}), (p_{Y|X}^{(1)}, \rho^{(1)}), \dots\}$ if, for all i,

the transmission of the source (p_S, d) across the channel $(p_{Y|X}^{(i)}, \rho^{(i)})$ using the code (f, g) is optimal.

Note that by complete analogy, one can define the universality of a code with respect to a *class* of sources and a class of channels. In order to keep notation simple, we leave this as an exercise to the reader. Instances of universality can be characterized by direct application of Theorem 7 to the present scenario. For example, from Theorem 7, Part *(i)*, we obtain the following corollary:

Corollary 11. *Consider a source (p_S, d) and a class of channels $\mathcal{W}$. Suppose that for all channels in the class, $0 < I(S;\hat{S}^{(i)})$ and $I(X;Y^{(i)}) < C_0^{(i)}$. Then, the single-letter code (f, g) is universal for the given source (p_S, d) and the given class of channels $\mathcal{W}$ if and only if for all i,*

$$\rho^{(i)}(x) = c_1^{(i)} D(p_{Y|X}^{(i)}(\cdot|x)||p_Y(\cdot)) + \rho_0^{(i)}, \tag{41}$$

$$d(s,\hat{s}) = -c_2^{(i)} \log_2 p^{(i)}(s|\hat{s}) + d_0^{(i)}(s), \tag{42}$$

$$I(S;\hat{S}^{(i)}) = I(X;Y^{(i)}), \tag{43}$$

where $c_1^{(i)} > 0$, $c_2^{(i)} > 0$ and $\rho_0^{(i)}$ are constants and $d_0^{(i)}(s)$ is an arbitrary function.

Proof. Follows directly from Theorem 7. □

By analogy, one can again include all the special cases of Theorem 7. This is left to the reader. The main reason for studying this particular property of memoryless source/channel codes lies in its practical implications. One implication is to time-varying (fading) channels, as illustrated by the above example: The channel varies over time, but it always remains inside the class $\mathcal{W}$. For that case, it is immediate that single-letter codes achieve the performance of the best source compression followed by the best channel code. However, the significance of single-letter codes extends beyond the validity of the separation theorem. Two scenarios under which single-letter codes outperform any code designed according to the separation paradigm are mentioned and illustrated explicitly in the sequel.

Implication 1 (non-ergodic channels). *Let the single-letter code (f, g) be universal for the source (p_S, d) and the class of channels $\mathcal{W}$. Let the channel be in $\mathcal{W}$, but not determined at the time of code design. Then, transmission using the single-letter code (f, g) achieves optimal performance, regardless of which particular channel is selected.*

Implication 2 (single-source broadcast). *Let the single-letter code (f, g) be universal for the source (p_S, d) and the class of channels $\mathcal{W}$. In the particular broadcast scenario where the single source (p_S, d) is transmitted across multiple channels $(p_{Y|X}^{(i)}, \rho^{(i)}) \in \mathcal{W}$,*

transmission using the single-letter code (f, g) achieves optimal performance on each channel individually.

Example 4 (single-source Gaussian broadcast). Let the source be i.i.d. Gaussian of variance P. Let the broadcast channel be Gaussian with two users. More specifically, the channel operation consists in adding white Gaussian noise of variance σ_1^2 and σ_2^2, respectively, and subsequent scaling by a factor of $\beta_1 = P/(P + \sigma_1^2)$ and $\beta_2 = P/P + \sigma_2^2)$, respectively. Assume w.l.o.g. $\sigma_1^2 < \sigma_2^2$. This is illustrated in Fig. 5. It is well-known (see

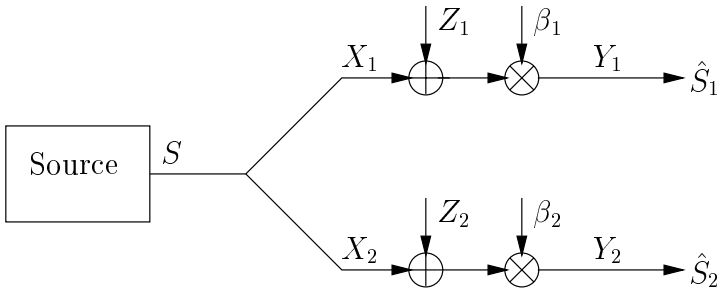

Figure 5: Single-source Gaussian broadcast.

also Example 1) that uncoded transmission is optimal on each of these channels individually, i.e. the distortion pair achieved by uncoded transmission is $\Delta_{u,1} = P\sigma_1^2/(P + \sigma_1^2)$ and $\Delta_{u,2} = P\sigma_2^2/(P + \sigma_2^2)$.

What is the achievable performance for a strategy based on the concept of separation? The source would have to be described by a coarse version and a refinement thereof. This problem has been studied in [9, 10]. For a Gaussian source, such a two-part description can be accomplished without loss. This means that if R_2 bits are used for the coarse version and R_1 bits for the refinement, then the reconstruction based on the coarse version only incurs a distortion of $D(R_2)$, while the reconstruction based on both the coarse version and the refinement incurs a distortion of $D(R_1 + R_2)$. Here, $D(\cdot)$ denotes the distortion-rate function of the source [3]. The rates that are available for these two descriptions are the pairs (R_1, R_2) in the capacity region of the Gaussian broadcast channel at hand. Since it is a degraded broadcast channel, the better receiver (the one at the end of the channel with σ_1^2) can also decode the information destined to the worse receiver [8]. Therefore, for the separation-based approach the distortion region is bounded by $\Delta_{c,1} = D(R_1 + R_2)$ and $\Delta_{c,2} = D(R_2)$, where R_1 and R_2 are on the boundary of the capacity region of the Gaussian broadcast channel. This is illustrated in Fig. 6 for a particular choice of the parameters. We observe that the distortion pair achieved by uncoded transmission lies strictly outside the distortion region for the separation-based approach that was described above.

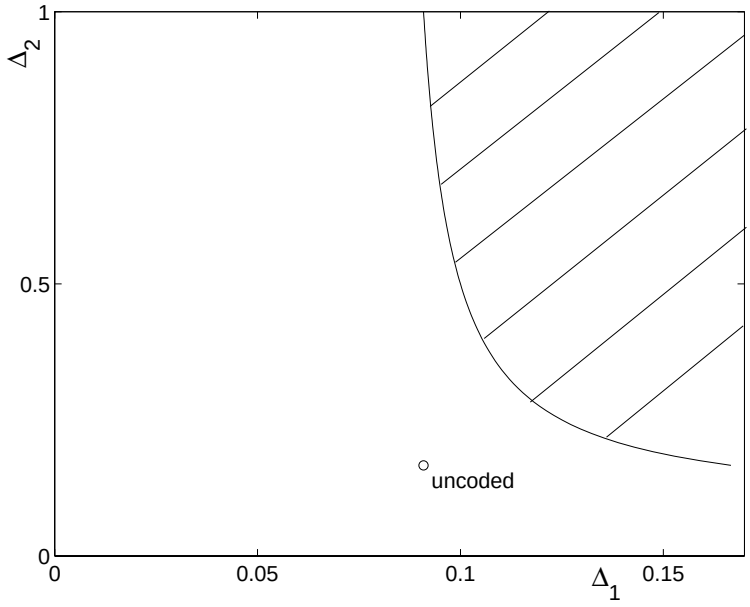

Figure 6: The distortion achievable by uncoded transmission (circle) versus the distortion region achievable by a transmission scheme based on the separation principle for Example 4. Parameters are $P = 1$, $\sigma_1^2 = 0.1$ and $\sigma_2^2 = 0.2$.

6 Concluding Remarks

To code, or not to code: that is the question. Undoubtedly, "not to code" is very appealing since it involves the smallest possible delay and complexity, but it can also involve a loss in transmission quality. However, for given source and channel (conditional) distributions, it is always possible to select the channel input cost function ρ and the distortion measure d such that no loss in transmission quality is incurred. In other words, under the appropriate channel input cost function and distortion measure, uncoded transmission achieves the same performance as the best source compression followed by the best channel code. In this paper, we determined explicit formulae to select ρ and d. We showed that these formulae are also necessary conditions in the sense that if ρ and d are not chosen according to them, then the overall system performs suboptimally.

The separation principle is limited to ergodic point-to-point communication. Interestingly, single-letter codes perform optimally in certain non-ergodic and multiuser communication scenarios. For example, a simple single-source broadcast situation was shown to have this property.

A question of practical interest that we have not considered in this paper is the following. Suppose that a source distribution p_S and a channel conditional distribution $p_{Y|X}$ are fixed. For any single-letter code (f, g), we can determine ρ and d to make the overall system

optimal, but these distortion measures and cost functions may not be meaningful for the given source/destination pair and for the physical constraints of the channel, respectively. Can f and g be cleverly chosen in such a way that the ρ and d from our formulae *are* physically meaningful? Moreover, if codes of block length M are permitted, how closely can some desired ρ and d be approximated?

Acknowledgments

We greatly appreciated the various suggestions by Emre Telatar (EPFL); in particular, the basic idea of Section 5.2 came up during a discussion of the authors with Emre. Initial discussions with Kannan Ramchandran (UC Berkeley) are also acknowledged.

A $R(\Delta) = C(\Gamma)$ Does Not Imply Optimality

In Section 3.2, it was shown that most cases of interest satisfy Condition *(ii)* of Lemma 1. Are there examples which do not? This section presents such an example: a source/channel/code triplet that satisfies $R(\Delta) = C(\Gamma)$ and yet does not represent an optimal communication system.

Example 5 (noisy typewriter channel). Let all involved alphabets be $\mathcal{S} = \mathcal{X} = \mathcal{Y} = \hat{\mathcal{S}} = \{0, 1, \ldots, L-1\}$, where L is an even integer. The channel conditional pmf is the noisy typewriter channel as in [8, p. 185], that is, $p_{Y|X}(k|k) = 1/2$ and $p_{Y|X}((k+1) \mod L|k) = 1/2$, for all k. The unconstrained capacity of this channel is found to be $C_0 = \log_2 \frac{L}{2}$. Let the encoder and decoder be the identity function. For the source pmf, define $p_{odd}(s)$ to be the uniform pmf over the odd inputs, and $p_{even}(s)$ the uniform pmf over the even inputs. Let the source pmf be a convex combination of these two, i.e. $p_\lambda(s) = \lambda p_{odd}(s) + (1-\lambda)p_{even}(s)$, where $0 \leq \lambda \leq 1$. Notice that $p_\lambda(s)$ achieves capacity on the unconstrained noisy typewriter channel for any λ.

Define the following distortion measure:

$$d(s, \hat{s}) \quad = \quad \begin{cases} 0, & \hat{s} = s \text{ or } \hat{s} = (s+1) \bmod L \\ 1, & \text{otherwise.} \end{cases} \tag{44}$$

Certainly, $\Delta = Ed(S, \hat{S}) = 0$. Moreover, we find that for any λ,

$$R(\Delta = 0) \quad = \quad \log_2 \frac{L}{2} \tag{45}$$

Let the input cost function be

$$\rho(x) \quad = \quad \begin{cases} 1, & x \text{ even}, \\ 0, & x \text{ odd}. \end{cases} \tag{46}$$

Suppose now that the source has $\lambda = 1/2$. Is the overall communication system optimal in that case? For $\lambda = 1/2$, we compute $\Gamma = \frac{L}{2}$, and hence

$$C(\Gamma) \quad = \quad C_0 = \log_2 \frac{L}{2}. \tag{47}$$

Evidently, the condition $R(\Delta) = C(\Gamma)$ is satisfied. Unfortunately, however, this is *not* an optimal communication system. Consider for example the source with parameter $\lambda = 1$. We compute $\Gamma' = 0 < \Gamma$, but clearly, $C(\Gamma) = C(\Gamma')$. Hence, the second condition of Lemma 1 is violated: It is indeed possible in this case to lower Γ without changing $C(\Gamma)$. Practically, this means that for the source with parameter $\lambda = 1/2$, there exists a *coded* communication system that achieves the same distortion but requires lower cost.

As a last remark, let us point out that the fact that the distortion and the cost Γ' are zero is *not* crucial for this example.

B Proof Of Theorem 10

Proof of Theorem 10. ($\Leftarrow$.) If there is a single-letter code with optimal performance, then trivially there is also a code of length M with optimal performance.
($\Rightarrow$.) Under the stated assumptions, the existence of a code of length M with optimal performance implies the existence of a single-letter code with optimal performance for the same source and channel. To prove this, we consider single-letter codes for the length-M extension source and channel.

Notation: Let $\underline{s} = (s_1, \ldots, s_M)$ be the vector of M consecutive source symbols, and define $\hat{\underline{s}}$ accordingly. By assumption, all alphabets are of the same size. Without (further) loss of generality, we use the generic alphabet $\{1, 2, \ldots, K\}$. The length-M extension source is $p(\underline{s}) = \prod_{m=1}^{M} p_S(s_m)$ with $d^{(M)}(\underline{s}, \hat{\underline{s}}) = \sum_{m=1}^{M} d(s_m, \hat{s}_m)$. For some of the considerations below, it will be more convenient to map $\underline{s}$ into an *extension alphabet* of size K^M according to $\mathbf{s} = \sum_{i=1}^{M} K^i s_i$. Both representations will be used interchangeably. Similarly, the extension channel is $p(\underline{y}|\underline{x}) = \prod_{m=1}^{M} p_{Y|X}(y_m|x_m)$ with $\rho^{(M)}(\underline{x}) = \sum_{m=1}^{M} \rho(x_m)$. In the proof, it will also be handy to use matrix notation. We will use $P_{Y|X}$ for the matrix of channel transition probabilities, where y indexes the rows and x the columns. Note that in the extension alphabet, $P_{\mathbf{Y}|\mathbf{X}} = P_{Y|X} \otimes \ldots \otimes P_{Y|X}$ (M terms), where $\otimes$ denotes the Kronecker product (tensor product).

Outline: The single-letter code for the length-M extension will be denoted $(f^{(M)}, g^{(M)})$. Obviously, this is an M-letter code for the original source and channel. We will now apply the theory developed in this paper to the extension source and channel, and their single-letter code $(f^{(M)}, g^{(M)})$. Plugging $p_{\mathbf{S}}$, $p_{\mathbf{Y}|\mathbf{X}}$ and the code $(f^{(M)}, g^{(M)})$ into Formulae

(3) and (11) of Theorems 3 and 4, we obtain the $\rho^{(M)}$ and $d^{(M)}$ that are necessary and sufficient for optimal performance.[5] However, by assumption, they have to be averaging (or *single-letter*) measures, that is, $\rho^{(M)}(\underline{x}) = \sum_{m=1}^{M} \rho(x_m)$ for some cost function $\rho(\cdot)$, and $d^{(M)}(\underline{s}, \underline{\hat{s}}) = \sum_{m=1}^{M} d(s_m, \hat{s}_m)$ for some distortion measure $d(\cdot, \cdot)$. This excludes many of the possible M-letter codes $(f^{(M)}, g^{(M)})$.

From Theorem 4, the distortion measure has to be chosen as

$$d^{(M)}(\underline{s}, \underline{\hat{s}}) \quad = \quad -\log_2 p(\underline{s}|\underline{\hat{s}}). \tag{48}$$

Clearly, for this to split additively into equal functions each of which depends only on one of the pairs $(s_i, \hat{s}_i)$, it is necessary that $p(\underline{s}|\underline{\hat{s}}) = p_{S|\hat{S}}(s_1|\hat{s}_1) \cdot \ldots \cdot p_{S|\hat{S}}(s_M|\hat{s}_M)$. In terms of transition probability matrices, this can be expressed as

$$P_{\mathbf{S}|\hat{\mathbf{S}}} \quad = \quad P_{S|\hat{S}} \otimes \ldots \otimes P_{S|\hat{S}}. \tag{49}$$

By symmetry, the second key insight follows from the fact that the cost function has to split additively. However, the derivation is somewhat more technical. Therefore, we state the result in the shape of the following lemma, to be proved below:

Lemma A. If $\rho^{(M)}$ is averaging and $P_{Y|X}$ invertible, then X and Y are iid.

The third insight is that under the additional assumptions on the alphabet sizes and $p(s)$, the encoder and decoder have to be bijective. It is given by the following lemma (to be proved below):

Lemma B. If all alphabets are of the same cardinality, $p(s) > 0$ for all s, $P_{S|\hat{S}}$ and $P_{Y|X}$ are invertible and $d^{(M)}$ is averaging, then encoder $f^{(M)}$ and decoder $g^{(M)}$ are bijections.

To complete the proof, consider first the encoder. Suppose that for fixed distribution of S and X, there exists indeed a bijective encoder $f^{(M)}$ that maps $\mathbf{S}$ to $\mathbf{X}$. Equivalently, this means that there exists a permutation matrix $F^{(M)}$ such that $\underline{p}_{\mathbf{X}} = F^{(M)} \underline{p}_{\mathbf{S}}$, where $\underline{p}_{\mathbf{X}}$ is a vector containing the probabilities $p_{\mathbf{X}}(\mathbf{x})$, and $\underline{p}_{\mathbf{S}}$ the corresponding for the random variable $\mathbf{S}$. By Lemma A, X is iid, hence we can write

$$\underline{p}_X \otimes \ldots \otimes \underline{p}_X \quad = \quad F^{(M)}(\underline{p}_S \otimes \ldots \otimes \underline{p}_S). \tag{50}$$

But this can only be true if there exists also a permutation matrix F such that

$$\underline{p}_X \quad = \quad F \underline{p}_S. \tag{51}$$

[5] Suppose there exists a code $(f^{(M)}, g^{(M)})$ such that $I(\underline{X}; \underline{Y}) = MC_0$. When all alphabets are of the same cardinality and $p(s) > 0$ for all s, it is a simple matter to prove that there exists also a single-letter code that achieves $I(X; Y) = C_0$. For this reason, the interesting case is when $I(\underline{X}; \underline{Y}) < MC_0$, in which case the formula for ρ is indeed a necessary condition. A similar comment applies to the case $I(S; \hat{S}) = 0$.

In other words, there exists also a single-letter encoder f that maps S to X.

This argument can be applied to the matrix $P_{\mathbf{S}|\hat{\mathbf{S}}}$ to conclude that the decoder can also be implemented by a single-letter mapping. First, recall that $P_{\mathbf{S}|\hat{\mathbf{S}}} = P_{\mathbf{S}|\mathbf{X}} P_{\mathbf{X}|\mathbf{Y}} P_{\mathbf{Y}|\hat{\mathbf{S}}}$. On the right hand side, $P_{\mathbf{X}|\mathbf{Y}}$ can be written as an M-fold Kronecker product because the channel is memoryless and X and Y are iid. Moreover, we have just shown that the encoder is a permutation matrix, and that it can be written also as an M-fold Kronecker product. Using Eqn. (49), we find

$$\begin{aligned} P_{S|\hat{S}} \otimes \ldots \otimes P_{S|\hat{S}} &= (P_{S|X} \otimes \ldots \otimes P_{S|X})(P_{X|Y} \otimes \ldots \otimes P_{X|Y}) P_{\mathbf{Y}|\hat{\mathbf{S}}} \\ &= (A \otimes \ldots \otimes A) P_{\mathbf{Y}|\hat{\mathbf{S}}} \end{aligned} \tag{52}$$

for some matrix A. But if there does indeed exist a permutation matrix $P_{\mathbf{Y}|\hat{\mathbf{S}}}$ that satisfies the above equation, then there exists also a permutation matrix $P_{Y|\hat{S}}$ that satisfies $P_{S|\hat{S}} = A P_{Y|\hat{S}}$, which implies the existence of a single-letter decoder. □

Remark. Let us explain at this point why the additional assumptions in Theorem 10 are necessary: To ensure that the encoder and the decoder are bijective maps. If this is not ensured, then the step from Eqn. (50) to Eqn. (51) seems to become surprisingly tricky.

Proof of Lemma A. From Theorem 3, the cost function $\rho(\underline{x})$ has to be chosen as

$$\rho(\underline{x}) = D\left(p_{\underline{Y}|\underline{X}}(\cdot|\underline{x}) || p_{\underline{Y}}(\cdot)\right) = \sum_{\underline{y}} p(\underline{y}|\underline{x}) \log_2 \frac{p(\underline{y}|\underline{x})}{p(\underline{y})}. \tag{53}$$

By definition, the cost function of a memoryless channel has to split additively into M equal functions, each depending only on one of the x_i. It is now shown that this implies that $p(y_1, \ldots, y_M) = p_Y(y_1) \cdot \ldots \cdot p_Y(y_M)$. For the case $M = 2$,

$$\rho(x_1, x_2) = H(Y|X = x_1) + H(Y|X = x_2) - \sum_{y_1, y_2} p(y_1|x_1) p(y_2|x_2) \log_2 p(y_1, y_2). \tag{54}$$

The last double sum has to split additively into two parts, one depending only on x_1, the other only on x_2. As a first step, we now show that this implies that Y_1 and Y_2 are independent random variables. Equivalently, we show that the matrix $P_{Y_1 Y_2}$ containing the joint pmf of Y_1 and Y_2 has rank at most 1.

To see why this holds, let us introduce the following shorthand: $z_i^j = p(y = j|x_i)$, where $1 \leq i \leq K$ and $1 \leq j \leq K$. Moreover, in this paragraph, we use $p(\cdot, \cdot)$ in place of $p_{Y_1 Y_2}(\cdot, \cdot)$ to make the formulae more readable. With this, we can rewrite the double sum on the RHS

of Eqn. (54) as

$$\begin{aligned} & z_1 z_2 \log_2 p(1,1) + z_1 z_2^2 \log_2 p(1,2) + \ldots + z_1 z_2^K \log_2 p(1,K) \\ + \quad & z_1^2 z_2 \log_2 p(2,1) + z_1^2 z_2^2 \log_2 p(2,2) + \ldots + z_1^2 z_2^K \log_2 p(1,K) \\ + \quad & \vdots \\ + \quad & z_1^K z_2 \log_2 p(K,1) + z_1^K z_2^2 \log_2 p(K,2) + \ldots + z_1^K z_2^K \log_2 p(K,K), \end{aligned} \tag{55}$$

with the constraint

$$z_i + z_i^2 + \ldots + z_i^K \quad = \quad 1, \text{ for all } i. \tag{56}$$

To split the sum additively into terms that depend only on one of the x_i (or, equivalently, of the z_i), it is necessary that the coefficients of all terms that involve more than one of the variables z_i are zero. Substitute for instance $z_1^2 = 1 - z_1 - z_2^3 - \ldots - z_2^K$ and $z_2^2 = 1 - z_2 - z_2^3 - \ldots - z_2^K$. Then it is quickly verified that the coefficient of $z_1 z_2$ is

$$\log_2 p(1,1) + \log_2 p(2,2) - \log_2 p(1,2) - \log_2 p(2,1). \tag{57}$$

But this is precisely the determinant of a 2×2 submatrix of $P_{Y_1 Y_2}$. In a similar fashion, we find that the determinants of all 2×2 submatrices of $P_{Y_1 Y_2}$ have to be zero. But this implies that $\operatorname{rank} P_{Y_1 Y_2} \leq 1$ (a well-known fact for which we did not find a reference, but which has a short proof; therefore it is given below as Lemma 12), which implies that Y_1 and Y_2 must be independent random variables.

For $M > 2$, define two sets of indices, $\mathcal{I}$ and $\mathcal{J}$, such that $\mathcal{I} \cap \mathcal{J} = \emptyset$. Let $Y^{(I)} = \{Y_i : i \in \mathcal{I}\}$ and $Y^{(J)} = \{Y_j : j \in \mathcal{J}\}$. But since Y are discrete random variables, $Y^{(I)}$ and $Y^{(J)}$ can be interpreted as two discrete random variables over larger alphabets. Denote the joint pmf matrix of $Y^{(I)}$ and $Y^{(J)}$ by P_{IJ}. For this matrix, it can again be shown that all 2×2 submatrices have zero determinant, and from Lemma 12, that P_{IJ} has rank one. Hence, the joint distribution matrix is $P_{IJ} = p_{Y^{(I)}} p'_{Y^{(J)}}$. Since this holds for any two index sets, it follows that the Y_i are independent random variables.

Up to now, we have established that $Y_1, \ldots, Y_M$ have to be independent random variables, thus we can write

$$\begin{aligned} \rho(x_1, \ldots, x_M) \quad = \quad & H(Y|X = x_1) - \sum_{y_1} p(y_1|x_1) \log_2 p(y_1) \\ & + \ldots + H(Y|X = x_M) - \sum_{y_2} p(y_2|x_M) \log_2 p(y_2), \end{aligned} \tag{58}$$

which indeed splits additively into M functions, each of which depends only on one of the x_i. Moreover, it has to split into *equal* functions. That is, whenever $x_i = x_j$, we must have

that

$$\sum_{y_i} p(y_i|x_i) \log_2 p(y_i) \quad = \quad \sum_{y_j} p(y_j|x_j) \log_2 p(y_j), \tag{59}$$

which can be rewritten (by letting $x = x_i = x_j$)

$$\sum_{y} p(y|x)(\log_2 p_{Y_i}(y) - \log_2 p_{Y_j}(y)) \quad = \quad 0. \tag{60}$$

This must hold for every choice of x. In other words, the vector $\{\log_2 p_{Y_i}(y) - \log_2 p_{Y_j}(y)\}_y$ must be orthogonal to all of the K vectors $\{p(y|x)\}_y$. Hence, if those K vectors span the entire K-dimensional space, then $p_{Y_i} = p_{Y_j}$, and thus Y_i and Y_j are identically distributed random variables. Thus, if the channel transition probability matrix $P_{Y|X}$ admits a right inverse, then the channel outputs $Y_1, \ldots, Y_M$ must be iid random variables.

Under certain circumstances, the fact that $Y_1, \ldots, Y_M$ are iid implies that $X_1, \ldots, X_M$ are also iid. A sufficient (but not necessary) condition for this is that the channel transition probability matrix $P_{Y|X}$ admit a left inverse. For codes of length $M = 2$, this can be shown as follows. Construct matrices $P_{Y_1Y_2} = \{p(y_1, y_2)\}_{y_1,y_2}$ and $P_{X_1X_2} = \{p(x_1, x_2)\}_{x_1,x_2}$. Then, we can write

$$P_{Y_1Y_2} \quad = \quad P_{Y|X} P_{X_1X_2} P_{Y|X}^T. \tag{61}$$

Denote the left inverse of $P_{Y|X}$ by $P_{Y|X}^L$. Then,

$$P_{Y|X}^L P_{Y_1Y_2} P_{Y|X}^{LT} \quad = \quad P_{X_1X_2}. \tag{62}$$

However, since Y_1 and Y_2 are iid, $P_{Y_1Y_2} = pp^T$ for some vector p, and thus

$$\operatorname{rank} P_{X_1X_2} \quad = \quad \operatorname{rank}(P_{Y|X}^L P_{Y_1Y_2} P_{Y|X}^{LT}) \leq \operatorname{rank} P_{Y_1Y_2} \quad = \quad 1. \tag{63}$$

Since moreover, $P_{X_1X_2} = P_{X_1X_2}^T$, there must exist a vector q such that $P_{X_1X_2} = qq^T$.

To extend this argument to $M > 2$, we use again the sets $\mathcal{I}$ and $\mathcal{J}$ as defined above. The joint distribution of $Y^{(I)}$ and $Y^{(J)}$ can thus be written in matrix form as $P_{IJ} = p_{Y^{(I)}} p'_{Y^{(J)}}$. This is a rectangular matrix of dimension $K^{|\mathcal{I}|} \times K^{|\mathcal{J}|}$. By construction, it has only one non-zero singular value. The transition probability matrices are Kronecker products of multiple copies of $P_{Y|X}$ and are therefore also left invertible. This implies (by analogy to the argument for $M = 2$) that the joint pmf matrix of $X^{(I)}$ and $X^{(J)}$ has also only one non-zero singular value, which means that it must be the outer product of two vectors, hence $X^{(I)}$ and $X^{(J)}$ are independent. But since this holds for arbitrary sets $\mathcal{I}$ and $\mathcal{J}$, we have that $X_1, \ldots, X_M$ must be independent. The fact that they are also identically distributed can then be derived by considering X_i and X_j for all $i \neq j$, and using the same argument as in the case $M = 2$. □

Proof of Lemma B. Consider the matrix $P_{\mathbf{S}|\hat{\mathbf{S}}}$. It may be expressed as $P_{\mathbf{S}|\hat{\mathbf{S}}} = P_{\mathbf{S}|\mathbf{X}} P_{\mathbf{X}|\mathbf{Y}} P_{\mathbf{Y}|\hat{\mathbf{S}}}$. The distortion measure has to be averaging, which, by Eqn. (49), implies that $P_{\mathbf{S}|\hat{\mathbf{S}}} = P_{S|\hat{S}} \otimes \ldots \otimes P_{S|\hat{S}}$ (M terms). By assumption, $P_{S|\hat{S}}$ is nonsingular. This is true if and only if $P_{\mathbf{S}|\hat{\mathbf{S}}}$ is also nonsingular. Hence, $P_{\mathbf{S}|\mathbf{X}}$ and $P_{\mathbf{Y}|\hat{\mathbf{S}}}$ must be full-rank matrices.

Moreover, using the requirement that $I(\mathbf{X};\mathbf{Y}) = I(\mathbf{S};\hat{\mathbf{S}})$, we now infer that $P_{\mathbf{S}|\mathbf{X}}$ and $P_{\mathbf{Y}|\hat{\mathbf{S}}}$ have to be permutation matrices. Consider the mutual information

$$\begin{aligned} I(\mathbf{S},\mathbf{X};\mathbf{Y}) &= I(\mathbf{S};\mathbf{Y}) + I(\mathbf{X};\mathbf{Y}|\mathbf{S}) \\ &= I(\mathbf{X};\mathbf{Y}) + I(\mathbf{S};\mathbf{Y}|\mathbf{X}), \end{aligned} \tag{64}$$

where $I(\mathbf{S};\mathbf{Y}|\mathbf{X}) = 0$ since $\mathbf{S} \to \mathbf{X} \to \mathbf{Y}$ is a Markov chain, and hence $I(\mathbf{X};\mathbf{Y}) = I(\mathbf{S};\mathbf{Y}) + I(\mathbf{X};\mathbf{Y}|\mathbf{S})$. To satisfy $I(\mathbf{X};\mathbf{Y}) = I(\mathbf{S};\mathbf{Y})$, it is therefore necessary that $I(\mathbf{X};\mathbf{Y}|\mathbf{S}) = H(\mathbf{X}|\mathbf{S}) - H(\mathbf{X}|\mathbf{Y},\mathbf{S}) = 0$. This is true if and only if $\mathbf{X}$ and $\mathbf{Y}$ are independent given $\mathbf{S}$. Hence consider the joint distribution matrix $P_{\mathbf{Y},\mathbf{X}|\mathbf{S}=\mathbf{s}}$. Denoting by $P_{\mathbf{X}|\mathbf{S}=\mathbf{s}}$ a diagonal matrix with entries $p(\mathbf{x}|\mathbf{s})$ along the diagonal, we can write

$$P_{\mathbf{Y},\mathbf{X}|\mathbf{S}=\mathbf{s}} = P_{\mathbf{Y}|\mathbf{X}} P_{\mathbf{X}|\mathbf{S}=\mathbf{s}}. \tag{65}$$

For $\mathbf{X}$ and $\mathbf{Y}$ to be independent given $\mathbf{S}$, the matrix $P_{\mathbf{Y},\mathbf{X}|\mathbf{S}=\mathbf{s}}$ has to have rank 1 for all $\mathbf{s}$. However, since by assumption, any set of two columns of $P_{\mathbf{Y}|\mathbf{X}}$ are linearly independent, the diagonal matrix $P_{\mathbf{X}|\mathbf{S}=\mathbf{s}}$ can have at most one non-zero entry, hence $p(\mathbf{x}|\mathbf{s}) = 1$ for exactly one of the $\mathbf{x}$. Hence, the matrix $P_{\mathbf{X}|\mathbf{S}}$ has only ones and zeros as entries. Moreover, $p(\mathbf{s}) > \mathbf{0}$ for all $\mathbf{s}$ and $P_{\mathbf{S}|\mathbf{X}}$ is invertible, which implies that $P_{\mathbf{X}|\mathbf{S}}$ is also invertible. Hence $P_{\mathbf{X}|\mathbf{S}}$ is a permutation matrix (and so is $P_{\mathbf{S}|\mathbf{X}}$).

By analogy, consider

$$\begin{aligned} I(\mathbf{S};\mathbf{Y},\hat{\mathbf{S}}) &= I(\mathbf{S};\mathbf{Y}) + I(\mathbf{S};\hat{\mathbf{S}}|\mathbf{Y}) \\ &= I(\mathbf{S};\hat{\mathbf{S}}) + I(\mathbf{S};\mathbf{Y}|\hat{\mathbf{S}}). \end{aligned} \tag{66}$$

To satisfy $I(\mathbf{S};\mathbf{Y}) = I(\mathbf{S};\hat{\mathbf{S}})$, we need that $I(\mathbf{S};\mathbf{Y}|\hat{\mathbf{S}}) = H(\mathbf{Y}|\hat{\mathbf{S}}) - H(\mathbf{Y}|\hat{\mathbf{S}},\mathbf{S}) = 0$. This is true if and only if $\mathbf{S}$ and $\mathbf{Y}$ are independent given $\hat{\mathbf{S}}$. By analogy to the first half of the proof, this implies that $P_{\mathbf{Y}|\hat{\mathbf{S}}}$ can have only zero or one as entries. Since moreover, it is invertible, it follows that $P_{\mathbf{Y}|\hat{\mathbf{S}}}$ is a permutation matrix.

To conclude the proof, note that permutation matrices represent bijective mappings. □

Lemma 12. *For any matrix $A \in M(n \times m)$, $\mathrm{rank}(A) \leq 1$ if and only if*

$$A_{ij}A_{kl} - A_{kj}A_{il} = 0, \tag{67}$$

for all $1 \leq i,k \leq n$ and $1 \leq j,l \leq m$.

Proof. $rank(A) \leq 1 \Leftrightarrow A = xy^T$ for some vectors x and y. But then, the forward part is immediate.

For the reverse, we show that any two rows of A are dependent. Pick row i and row k, and form the $2 \times m$ submatrix A'. The rows of A' are independent if and only if we can find two columns j and l that are linearly independent. This happens if and only if the 2×2 submatrix A'' containing only columns j and l of A' has full rank. However, by assumption, this matrix has determinant zero. Hence any two rows of A are dependent, and the rank cannot be larger than 1. □

References

[1] C. E. Shannon, "A mathematical theory of communication," *Bell Sys. Tech. Journal*, vol. 27, pp. 379–423, 623–656, 1948.

[2] S. Vembu, S. Verdú, and Y. Steinberg, "The source-channel separation theorem revisited," *IEEE Transactions on Information Theory*, vol. 41, pp. 44–54, January 1995.

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

[4] R. J. McEliece, *The Theory of Information and Coding.* Encyclopedia of mathematics and its applications, Reading MA: Addison-Wesley, 1977.

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

[6] S. Shamai, S. Verdú, and R. Zamir, "Systematic lossy source/channel coding," *IEEE Transactions on Information Theory*, vol. 44, pp. 564–579, March 1998.

[7] D. G. Luenberger, *Optimization by Vector Space Methods.* New York: John Wiley, 1969.

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

[9] W. H. R. Equitz and T. M. Cover, "Successive refinement of information," *IEEE Transactions on Information Theory*, vol. 37, pp. 460–473, March 1991.

[10] B. Rimoldi, "Successive refinement of information: Characterization of the achievable rates," *IEEE Trans. Inform. Theory*, vol. IT–40, pp. 253–259, Jan. 1994.