To Code, or Not to Code: Lossy Source–Channel Communication Revisited

Michael Gastpar, *Associate Member, IEEE*, Bixio Rimoldi, *Fellow, IEEE*, and Martin Vetterli, *Fellow, IEEE*

***Abstract*—What makes a source–channel communication system optimal? It is shown that in order to achieve an optimal cost–distortion tradeoff, the source and the channel have to be matched in a probabilistic sense. The match (or lack of it) involves the source distribution, the distortion measure, the channel conditional distribution, and the channel input cost function. Closed-form necessary and sufficient expressions relating the above entities are given. This generalizes both the separation-based approach as well as the two well-known examples of optimal uncoded communication.**

The condition of probabilistic matching is extended to certain nonergodic and multiuser scenarios. This leads to a result on optimal single-source broadcast communication.

***Index Terms*—Joint source–channel coding, separation theorem, single-letter codes, single-source broadcast, uncoded transmission.**

I. 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 [3, Theorem 21], and is discussed, e.g., in [4]–[7]. The result is of surprisingly wide validity in point-to-point communication (see, e.g., [8]). Consequently, the separation idea has split the research community into two camps, those who examine source compression and those who investigate channel coding. It is well known that the combination of the results from the two communities leads to an optimal communication system in terms of the cost–distortion tradeoff, but it is generally highly complex, and it disregards delay.

In order to determine whether a lossy source–channel communication system is optimal, it suffices to measure the average cost and the average distortion, and to verify that this cost–distortion pair lies on the optimal cost–distortion tradeoff curve. But the average cost depends only on the marginal distribution $p(x)$ at the channel input, and the average distortion depends only on the joint (marginal) distribution $p(s, \hat{s})$ of the source and the destination symbols. The inevitable conclusion is that achieving optimality is a matter of achieving the correct marginals $p(x)$ and $p(s, \hat{s})$. The system designed according to the separation principle achieves the right marginals by means of (asymptotically) long codewords, but the use of long codes is by no means a requirement.

For instance, it is well known that an optimal communication system results when a binary uniform source is plugged into a binary-symmetric channel without any coding at all, provided that the distortion is measured in terms of the Hamming distance (see, e.g., [9, Sec. 11.8], or [5, p. 117]). In another well-known example of such behavior, a Gaussian source is transmitted across the additive white Gaussian noise (AWGN) channel [10].

The reason why we do not need coding in these two examples is that the source and the channel together produce the right marginals. We say that they are *probabilistically matched*. If, instead, we follow the separation principle, the capacity-approaching channel code destroys this favorable condition by attempting to create a deterministic channel (with high probability). Note that this also implies that the mapping from the source output sequence to the reconstruction sequence at the destination becomes *deterministic* (with high probability). In contrast to this, the source-to-destination mapping in the binary and Gaussian examples quoted above is *random*. This is discussed in more detail in [11].

The goal of this paper is to provide a basis for point-to-point source–channel communication systems that are optimal, including the uncoded ones and those designed according to the separation principle. We derive a set of necessary and sufficient conditions for any discrete memoryless point-to-point communication system to be optimal. These conditions show that by considering a broader class of source-to-destination mappings (rather than only the deterministic ones), there is an arbitrarily large number of source–channel pairs for which the complexity and delay can be reduced to the absolute minimum, yet the optimal cost–distortion tradeoff is achieved.

The paper is organized as follows. In Section II, a brief review of Shannon's conditions for the optimality of a source–channel communication system is given.

In Section III, these conditions are specialized to the case of single-letter codes, i.e., codes where the encoder maps every source output symbol separately onto a channel input symbol, and the decoder maps every channel output symbol separately onto a source reconstruction symbol. In a nutshell, suppose that a discrete memoryless source specified by the random variable S with distribution $p(s)$ is encoded (symbol by symbol) into

Manuscript received May 23, 2001; revised January 6, 2003. The material in this paper was presented in part at the IEEE International Symposium on Information Theory, Sorrento, Italy, June 2000 [1], and at the IEEE International Symposium on Information Theory, Washington DC, June 2001 [2].

M. Gastpar is with the Department of Electrical Engineering and Computer Science, University of California, Berkeley, Berkeley, CA 94720-1770 USA (e-mail: gastpar@eecs.berkeley.edu).

B. Rimoldi is with the Institute of Communication Systems, Swiss Federal Institute of Technology (EPFL), CH-1015 Lausanne, Switzerland (e-mail: bixio.rimoldi@epfl.ch).

M. Vetterli is with the Institute of Communication Systems, Swiss Federal Institute of Technology (EPFL), CH-1015 Lausanne, Switzerland, and the Department of Electrical Engineering and Computer Science, University of California, Berkeley, Berkeley, CA 94720-1770 USA (e-mail: martin.vetterli@epfl.ch).

Communicated by P. Narayan, Associate Editor for Shannon Theory.

Digital Object Identifier 10.1109/TIT.2003.810631

$X = f(S)$. The symbol X is transmitted across a discrete memoryless noisy channel specified by a conditional distribution $p_{Y|X}$. The channel output Y is decoded to yield the estimate of the source $\hat{S} = g(Y)$. We show that this is an optimal communication system if and only if the channel input cost function $\rho(x)$ and the distortion measure $d(s, \hat{s})$ relate to the source and channel distributions according to (up to shifts and scaling)

$$\rho(x) = D\left(p_{Y|X}(\cdot|x)||p_Y(\cdot)\right) \tag{1}$$

$$d(s, \hat{s}) = -\log_2 p(s|\hat{s}) \tag{2}$$

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 6. Equations (1) and (2) suggest our perspective of *probabilistic matching*. In order to achieve an optimal cost–distortion tradeoff, the source and the channel have to be matched in a probabilistic sense. The match (or lack of it) involves the source distribution, the distortion function, the channel conditional distribution, and the channel input cost function. The two conditions above were known as necessary and sufficient conditions for a channel input distribution and for a test channel to achieve the maximum and the minimum, respectively, in the computation of the capacity–cost function and the rate-distortion function, respectively [6]. However, the more important point is new: the above conditions characterize an optimal system in a way that is more fundamental than the well-known condition that the rate-distortion function be equal to the capacity–cost function.

In Section IV, we argue that the results obtained in Section III extend to arbitrary encoders, mapping k source symbols onto m channel input symbols, and to arbitrary decoders, mapping m channel output symbols to k source reconstruction symbols, at least as long as all alphabets are assumed discrete. What code length is required in order to achieve the optimum match? For some source/channel pairs, a single-letter code is sufficient, as shown in Section III. Consider now the source–channel pairs for which there is *no* single-letter code that achieves this. We establish for a subset of those source–channel pairs that there is no code of finite block length that achieves the optimal match either.

The significance of joint source–channel codes extends beyond memoryless point-to-point systems. This is discussed in Section V. 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 nonergodic channels and multiuser communication, where the separation-based approach is generally suboptimal (see Example 4 in Section V).

Finally, the application of information theory to biology, in particular to neural communication, has received increasing attention. The work of Berger [12] is pioneering in this area; one way of applying the concept of probabilistic matching to neural communication is illustrated in Example 2 in Section III-D.

II. Optimal Source–Channel Communication Systems

The key elements of the problem studied in this paper are the discrete 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 probability mass function (pmf) of the random variable S is denoted by $p_S(s)$. 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 memoryless source (p_S, d) is specified by a pmf $p_S(s)$ on a discrete alphabet $\mathcal{S}$ and a nonnegative function $d(s, \hat{s})$: $\mathcal{S} \times \hat{\mathcal{S}} \to \mathbb{R}^+$ called the distortion measure. This implicitly specifies a discrete alphabet $\hat{\mathcal{S}}$ in which the source is reconstructed. The rate-distortion function (see e.g., [13]) of the source (p_S, d) is defined as

$$R(D) = \min_{p_{\hat{S}|S}:\, Ed(S,\hat{S}) \le D} I(S; \hat{S}) \tag{3}$$

where $\hat{S}$ is a random variable over the alphabet $\hat{\mathcal{S}}$.

Definition 2 (Channel): A discrete memoryless channel $(p_{Y|X}, \rho)$ is specified by a conditional pmf $p_{Y|X}(y|x)$, where $X \in \mathcal{X}$ and $Y \in \mathcal{Y}$, and where $\mathcal{X}$ and $\mathcal{Y}$ are discrete alphabets, and a nonnegative function $\rho(x)$: $\mathcal{X} \to \mathbb{R}^+$ called the channel input cost function. The capacity–cost function (see, e.g., [5]) of the channel $(p_{Y|X}, \rho)$ is defined as

$$C(P) = \max_{p_X:\, E\rho(X) \le P} I(X; Y). \tag{4}$$

This function is also called capacity-constraint function in [6].

In order to decide on the optimality of a communication system, 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,

$$C_0 = \max_{p_X} I(X; Y). \tag{5}$$

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 $C_0 = \lim_{P \to \infty} C(P)$.

In this paper, we study communication by means of a source–channel code of rate κ, defined as follows.

Definition 4 (Source–Channel Code of Rate κ): A source–channel code (F, G) of rate κ is specified by an encoding function F: $\mathcal{S}^k \to \mathcal{X}^m$ and a decoding function G: $\mathcal{Y}^m \to \hat{\mathcal{S}}^k$, such that $k/m = \kappa$.

Remark 1: κ is part of the problem statement, *not* a parameter in the optimization: κ is the number of source symbols that have to be transmitted per channel use.

Suppose the source (p_S, d) is transmitted across the channel $(p_{Y|X}, \rho)$ using the source–channel code (F, G). This is illustrated in Fig. 1. The average input cost used on the channel is found to be

$$\Gamma \stackrel{def}{=} \frac{1}{m} \sum_{i=1}^{m} E\rho(X_i) \tag{6}$$

and the average distortion achieved by the code (F, G) is

$$\Delta \stackrel{def}{=} \frac{1}{k}\sum_{i=1}^{k} Ed(S_i, \hat{S}_i). \tag{7}$$

We will sometimes refer to (Γ, Δ) as the *cost–distortion* pair, or the operating point. The main goal of this paper is to determine necessary and sufficient conditions such that the communication system of Fig. 1 is an optimal communication system, 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 source–channel code (F, G) of rate κ is optimal if both[1]

i) the distortion Δ incurred using (F, G) is the minimum distortion that can be achieved at input cost Γ with the best possible source–channel code of rate κ (regardless of complexity), and

ii) the cost Γ incurred using (F, G) is the minimum cost needed to achieve distortion Δ with the best possible source–channel code of rate κ (regardless of complexity).

Following Definition 5, in order to establish the optimality of a given communication system, the issue is to optimize over all possible source–channel codes of a fixed rate κ, regardless of complexity and delay. This problem can be solved by the aid of the separation theorem [3, Theorem 21]. For the purpose of this paper, we formulate it as follows.

Lemma 1: For a source (p_S, d) and a channel $(p_{Y|X}, \rho)$, transmission using a source-channel code (F, G) of rate κ is optimal if and only if 1) or 2) is satisfied:

1) i) $\kappa R(\Delta) = C(\Gamma)$, and
   ii) neither can Δ be lowered without changing $R(\Delta)$ nor can Γ be lowered without changing $C(\Gamma)$;

2) $$\Delta = \min_{D}\{D\colon R(D) = \max_{D'} R(D')\}$$
and
$$\Gamma = \min_{P}\{P\colon C(P) = \min_{P'} C(P')\}.$$

Remark 2: The crucial condition of case 1) is $\kappa R(\Delta) = C(\Gamma)$. For condition ii), note that 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 III-B.

Remark 3: Case 2) is degenerate in the sense that there is no tradeoff between cost and distortion; there is only one optimal operating point. This is illustrated in Fig. 2 by the dashed lines. In that example, the only optimal operating point is $\Delta = D_{\min}$ and $\Gamma = P_{\min}$.

Outline of the Proof: This lemma is essentially [3, Theorem 21]. For an outline of the proof, note that by combining the proofs of the rate-distortion and the capacity theorem with the data processing inequality, it is easy to show that any source–channel communication system of rate κ must satisfy

[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.

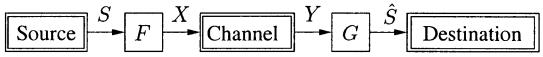

Fig. 1. The source–channel communication system.

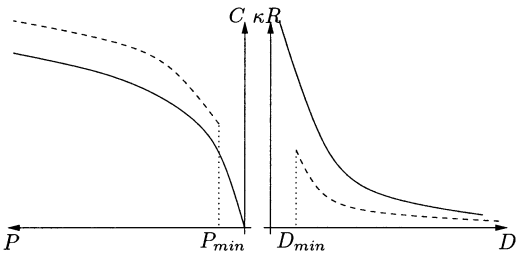

Fig. 2. The bold lines represent a source–channel pair for which $\kappa R(\Delta) = C(\Gamma)$ is feasible, and hence case 1) of Lemma 1 applies; the dashed lines represent a source–channel pair for which it is not feasible, i.e., case 2) of Lemma 1 applies.

$\kappa R(\Delta) \le C(\Gamma)$. For case 1), suppose that for the source–channel pair $\kappa R(\Delta) = C(\Gamma)$ is feasible. Then, if a communication strategy does not satisfy this, it cannot be optimal: there is a better system (namely, one that does satisfy $\kappa R(\Delta) = C(\Gamma)$). However, $\kappa R(\Delta) = C(\Gamma)$ is still not sufficient as it may occur that Δ can be reduced without increasing $R(\Delta)$. This is prevented by condition ii). The same comment applies to $C(\Gamma)$. Hence, i) and ii) together are necessary and sufficient conditions for optimality. For case 2), $\kappa R(\Delta) = C(\Gamma)$ is not feasible, i.e., $\kappa R(\Delta) < C(\Gamma)$. The optimality is clear from the dashed lines in Fig. 2.

III. 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 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. In line with Definition 4, we define the following.

Definition 6 (Single-Letter Source–Channel Code): A single-letter source–channel code (f, g) is specified by an encoding function $f(\cdot)\colon \mathcal{S} \to \mathcal{X}$ and a decoding function $g(\cdot)\colon \mathcal{Y} \to \hat{\mathcal{S}}$.

Note that for single-letter codes, $\kappa = 1$. Lemma 1 contains two conditions[2] that together are necessary and sufficient to establish the optimality of any communication system, including those that use single-letter codes. These conditions will now be examined in detail. In Section III-A, we elaborate on the first condition, i.e., $R(\Delta) = C(\Gamma)$. The second condition is somewhat subtler; it will be discussed in Section III-B. In Section III-C, the results are combined to yield a general criterion for the optimality of single-letter codes.

A. Condition i) of Lemma 1

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

[2]Here, we only derive detailed results for Case 1) (i.e., the nondegenerate case) of Lemma 1.

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 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 $I(S;\hat{S}) = I(X;Y)$, i.e., they are "information lossless."

Proof: For any source–channel communication system that employs a single-letter code

$$R(\Delta) = \min_{q_{\hat{S}|S}:\, Ed(S,\hat{S})\le\Delta} I(S;\hat{S}) \overset{(a)}{\le} I(S;\hat{S}) \overset{(b)}{\le} I(X;Y) \overset{(c)}{\le} \max_{q_X:\, E\rho(X)\le\Gamma} I(X;Y) = C(\Gamma) \quad (8)$$

where (b) is the data processing inequality. Equality holds in (a) if and only if $p_{\hat{S}|S}$ achieves the rate-distortion function of the source, 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 last 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 lemma 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. Lemma 3 gives an explicit formula to select the input cost function ρ for given channel conditional and input distributions. By analogy, Lemma 4 gives a similar condition for the distortion measure.

The results of the following Lemmas 3 and 4 are not new. They have already appeared, in effect, in [6]. Specifically, Lemma 3 appears as Problem 2 (without explicit proof) in [6, p. 147]; it can also be seen as an extension of [4, Theorem 4.5.1] to the case of constrained channel inputs. Lemma 4 appears as Problem 3 (without explicit proof) in [6, p. 147].

Lemma 3: For fixed source distribution p_S, a 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

$$\rho(x) \begin{cases} = c_1 D(p_{Y|X}(\cdot|x)\|p_Y(\cdot)) + \rho_0, & \text{if } p(x) > 0 \\ \ge c_1 D(p_{Y|X}(\cdot|x)\|p_Y(\cdot)) + \rho_0, & \text{otherwise} \end{cases} \quad (9)$$

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 $\rho(x)$.

Proof: The proof is given in Appendix I.

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.

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

Lemma 4: For fixed source distribution p_S, channel conditional distribution $p_{Y|X}$, and a 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

$$d(s, \hat{s}) = -c_2 \log_2 p(s|\hat{s}) + d_0(s) \quad (10)$$

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 $d(s, \hat{s})$.

Proof: The proof is given in Appendix I.

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.

Apparently, there is a slight asymmetry between Lemmas 3 and 4: In the former, when $p(x) = 0$, $\rho(x)$ satisfies a less strin-

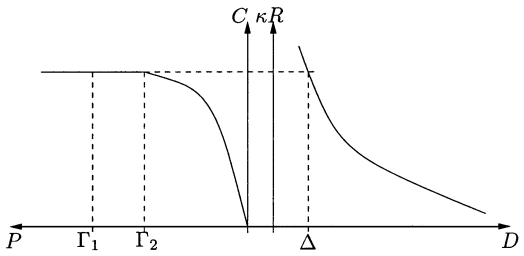

Fig. 3.When $R(\Delta) = C(\Gamma)$ is not sufficient to guarantee optimality.

gent condition. In the latter, a similar behavior occurs: when $p(s, \hat{s}) = 0$, the condition can be relaxed to

$$d(s, \hat{s}) \geq -c_2 \log_2 p(s|\hat{s}) + d_0(s).$$

However, since the right-hand side is infinity in that case, requiring equality is equivalent.

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 subsection.

B. 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, $R(\Delta) = C(\Gamma)$, was studied and developed in detail in Section III-A; 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 Fig. 3. 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 in regions where either $C(\cdot)$ or $R(\cdot)$ are horizontal, i.e., when they have reached their asymptotic values $\lim_{P\to\infty} C(P)$ and $\lim_{D\to\infty} R(D)$. This only happens when the mutual information $I(X;Y)$ is C_0 or zero. 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 [14].

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

Proposition 5: 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 we have the following.

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 $\Gamma_{\max} = \min\{P\colon C(P) = C_0\}$. 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 $C(\Gamma) = C_0$.

Part ii): The proof goes along the same lines. To see that condition a) is sufficient, define $\Delta_{\max} = \min\{D\colon R(D) = 0\}$. For every $\Delta < \Delta_{\max}$, the value $R(\Delta)$ uniquely specifies Δ. This follows from the fact that $R(\cdot)$ is convex and nonincreasing. 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 4: In the most general case of Proposition 5, 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 III-A.

C. 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 Lemma 3 ensures that the channel input distribution achieves capacity. Similarly, the choice of the distortion measure according to Lemma 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 $R(\Delta) = C(\Gamma)$. On the other hand, Lemma 1 requires 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 Lemmas 3 and 4. Rather, it was discussed in Section III-B and led to Proposition 5. It is now a simple matter to combine the insight gained in the latter proposition with the statements from Lemmas 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 6: Consider a source (p_S, d) and a channel $(p_{Y|X}, \rho)$ for which $R(\Delta) = C(\Gamma)$ is feasible.[3] For the transmission using a 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 $d(s, \hat{s})$ satisfy Lemmas 3 and 4, respectively.
- ii) If $0 < I(S;\hat{S}) = I(X;Y) = C_0$, the system is optimal if and only if $d(s, \hat{s})$ satisfies Lemma 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 $\rho(x)$ satisfies Lemma 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 $\tilde{p}_{\hat{S}|S}$.

[3]This condition rules out the source–channel pairs of Case 2) of Lemma 1.

Proof:

Part o). From the Data Processing Theorem (e.g., [7, Theorem 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 5). 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 $I(X;Y) < C_0$, Lemmas 3 and 4 establish that the first two are satisfied *if and only if* ρ and d are chosen according to (9) and (10), 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 5). 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 $I(X;Y) = C_0$, Lemmas 3 and 4 establish that the first two are satisfied *if and only if* d is chosen according to (10).

Part iii). If $0 = I(S;\hat{S})$, the system is 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 5). 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 $I(X;Y) < C_0$ but $I(S;\hat{S}) = 0$, Lemmas 3 and 4 establish that the first two are satisfied *if and only if* ρ is chosen according to (9).

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. □

D. Illustrations of Theorem 6

To illustrate this theorem, pick any probability measures for the source and the channel, and determine the cost function and distortion measure according to Lemmas 3 and 4, respectively. For the well-known example of a binary uniform source across a binary-symmetric channel, this is done as follows.

Example 1 (Binary): Let the source be binary and uniform with Hamming distortion measure, and let the channel be binary and symmetric (with transition probability $\epsilon < 1/2$) 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., [5] and [6].

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-dis-

tortion function for the binary source is $R(D) = 1 - H_b(D)$ (see, e.g., [7]). 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 5, 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 6. Trivially, $I(X;Y) = I(S;\hat{S})$, and we find

$$p(s|\hat{s}) = \frac{p_{Y|X}(\hat{s}|s)p_S(s)}{p_Y(\hat{s})} = p_{Y|X}(\hat{s}|s) = \begin{cases} 1-\epsilon, & \text{if } \hat{s} = s \\ \epsilon, & \text{otherwise.} \end{cases} \tag{11}$$

Taking

$$d_0(s) = \frac{\log_2(1-\epsilon)}{\log_2 \frac{1-\epsilon}{\epsilon}} \quad \text{and} \quad c_2 = \frac{1}{\log_2 \frac{1-\epsilon}{\epsilon}}$$

in Lemma 4 reveals that one of the distortion measures that satisfy the requirement in Theorem 6 is indeed the Hamming distance.

As shown in this example, Theorem 6 can be applied directly by fixing the probability measures and the single-letter code, and determining the cost function and distortion measures according to the formulas. But the conditions of Theorem 6 are also useful if, e.g., the channel conditional probability distribution and the cost function are specified, and the source probability distribution and the distortion measure have to be determined accordingly, as illustrated by the following example [11].

Example 2: In this example, the alphabets are binary sequences of length n, denoted by bold symbols $\boldsymbol{x}$. Let the channel conditional distribution be any permutation of the length-n sequence, i.e.,

$$p(\boldsymbol{y}|\boldsymbol{x}) = \begin{cases} \binom{n}{w(\boldsymbol{x})}^{-1}, & \text{if } w(\boldsymbol{x}) = w(\boldsymbol{y}) \\ 0, & \text{otherwise} \end{cases} \tag{12}$$

where $w(\boldsymbol{x})$ denotes the Hamming weight (number of 1's) in the sequence $\boldsymbol{x}$. Moreover, let the cost function be

$$\rho(x) = a_1 w(\boldsymbol{x}) + a_0. \tag{13}$$

This can be seen as a simple model of neural communication [11]. By Lemma 3, the capacity-achieving input distribution satisfies

$$a_1 w(\boldsymbol{x}) + a_0 = c_1 D\left(p_{\boldsymbol{Y}|\boldsymbol{X}}(\cdot|\boldsymbol{x}) \| p_{\boldsymbol{Y}}(\boldsymbol{y})\right). \tag{14}$$

In [11], this condition is used to determine the capacity-achieving input distribution. The probability that $w(\boldsymbol{x}) = k$ is found to be

$$q(k) = \frac{b^n}{b^{n+1}-1}(b-1)b^{-k} \tag{15}$$

for $k \in \{0, 1, \ldots, n\}$. In [11], the distortion measure according to Lemma 4 is also determined.

Example 2 considered a simple model of neural communication. This also illustrates the point that in certain applications, the source and the channel *can* be selected in a favorable fashion: for the case of neural communication, evolution had the opportunity to do so.

Beyond such a direct application, Theorem 6 is also useful in certain proofs. Example 1 suggests the question of the *uniqueness* of the solution. Suppose that all involved alphabets are binary, the distortion measure is Hamming, and the channel input cost function a constant. Then, is Example 1 the unique instance of optimal uncoded transmission? Using Theorem 6, one can establish the following lemma.

Lemma 7 (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.

Proof: The proof is given in Appendix II.

If the alphabets are not binary, the following similar result can be established.

Lemma 8 (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: The proof is given in Appendix II.

There is a nice intuition going along with the last result. Suppose that the channel is symmetric [7, 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.

The claims made in this paper are for discrete alphabets only. However, the proofs of the sufficiency of Lemmas 3 and 4 given in Appendix I can be extended to continuous alphabets with appropriate technical assumptions. For example, suppose that the source distribution p_S is Gaussian of variance P, and the channel is an AWGN channel with noise variance σ^2. Suppose that uncoded transmission is used, and the decoder is $\hat{S} = P/(P+\sigma^2)Y$. Then, Lemmas 3 and 4 give

$$\rho(x) = c_1 x^2 + \rho_0 \tag{16}$$

$$d(s, \hat{s}) = c_2(s - \hat{s})^2 + d_0(s). \tag{17}$$

In words, if the cost on the channel is power, and the distortion the mean-square error, then uncoded transmission is optimal, confirming the well-known example reported in [10].

IV. Optimal Source–Channel Communication Systems, Revisited

In Section III, we developed results for single-letter codes. However, it is clear that any source–channel code can be seen as a single-letter code in appropriately extended alphabets, at least as long as all alphabets are assumed to be discrete (as we have done throughout the present paper). Hence, the results of Section III can be applied directly to arbitrary source–channel codes. In other words, we have developed a criterion to establish the optimality of any source–channel communication system, and that criterion is no less general than the separation theorem, Lemma 1.

More precisely, suppose that a source–channel code (F, G) is used, with F: $\mathcal{S}^k \to \mathcal{X}^m$ and G: $\mathcal{Y}^m \to \hat{\mathcal{S}}^k$. This situation can be addressed by merging k source symbols to yield a new source, denoted by S^k with alphabet $\mathcal{S}^k$. The distribution of the new source is

$$p(s^k) = \prod_{j=1}^{k} p(s_j). \tag{18}$$

Similarly, m channel symbols are merged to yield a new channel with conditional distribution

$$p(y^m|x^m) = \prod_{j=1}^{m} p(y_j|x_j). \tag{19}$$

Consequently, the cost function and the distortion measure are also defined in the new, extended alphabets, and are denoted $\rho^{(m)}(x^m)$ and $d^{(k)}(s^k, \hat{s}^k)$. In the new alphabets, the source–channel code (F, G) is a single-letter code. Hence, Theorem 6 can be used to obtain the following statement:

Corollary 9: For a source $(p_{S^k}, d^{(k)})$, and a channel $(p_{Y^m|X^m}, \rho^{(m)})$, suppose that $R(\Delta) = C(\Gamma)$ is feasible.[4] Consider the transmission using a single-letter source–channel code (F, G) with F: $\mathcal{S}^k \to \mathcal{X}^m$ and G: $\mathcal{Y}^m \to \hat{\mathcal{S}}^k$, and suppose that

$$0 < I(S^k; \hat{S}^k) = I(X^m; Y^m) < C_0.$$

This is optimal if and only if

$$\rho^{(m)}(x^m) \begin{cases} = c_1 D(p_{Y^m|X^m}(\cdot|x^m)||p_Y(\cdot)) + \rho_0, \\ \qquad\qquad \text{if } p(x^m) > 0 \\ \geq c_1 D(p_{Y|X}(\cdot|x)||p_Y(\cdot)) + \rho_0, \\ \qquad\qquad \text{otherwise} \end{cases} \tag{20}$$

$$d^{(k)}(s^k, \hat{s}^k) = -c_2 \log_2 p(s^k|\hat{s}^k) + d_0(s^k). \tag{21}$$

Proof: This corollary is Theorem 6, Part i), applied to suitably extended alphabets. □

Corollary 9 makes the concept of *probabilistic matching* precise. For given source and channel statistics, any source–channel code is optimal with respect to an appropriately chosen cost function and distortion measure. The goal of the code design can be understood as the determination of the code that achieves the closest match with the *desired* cost function and distortion measure. Note that the cost function $\rho^{(m)}(x^m)$ and distortion measure $d^{(k)}(s^k, \hat{s}^k)$ need not generally decompose in an additive fashion in terms of the original alphabet.

It is clear that longer codes generally permit to better match the source and the channel. Corollary 9 can, therefore, also be interpreted as follows. Suppose a certain finite complexity is available to implement a source–channel communication system. Following Lemma 1, one would design (suboptimal) source and channel codes independently. The advantage of additional complexity appears as a lower error probability on the channel and a smaller size of the quantization cells for the source. In contrast to this, Corollary 9 suggests a very different perspective: additional coding complexity (in the shape of longer codes) is used to better match $\rho^{(m)}$ and $d^{(k)}$ to the desired cost and distortion measures.

While longer codes permit to *better* match the source (p_S, d) to the channel $(p_{Y|X}, \rho)$, we would also like to know what code length is necessary to obtain the *optimal* match. More precisely, attention shall still be restricted to discrete memoryless sources and channels as defined in Definitions 1 and 2, but the code is now an arbitrary source–channel code of (finite) length M. For simplicity, we consider only codes of rate $\kappa = 1$. Corollary 9 gives 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. We can prove this conjecture under some additional assumptions.

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: The proof of this theorem is given in [14]. □

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 such as 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.

V. Extensions to Nonergodic and Multiuser Communication Systems

Optimal transmission systems designed according to the separation principle may be quite sensitive to parameter mismatch.

[4]This condition rules out the source–channel pairs of Case 2) of Lemma 1.

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.

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

Example 3 (Example 1 With Fading): Let the source be the binary uniform source as in Example 1. The channel is slightly different from Example 1: the transition probability ϵ varies during transmission. Take the encoder and the decoder to be identity mappings (i.e., uncoded transmission). From Example 1, it is clear that this code performs optimally irrespective of the value of ϵ.

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

Definition 7 (Universality): The source–channel code (F, G) is called universal for the source (p_S, d) and the class of channels given by

$$\mathcal{W} = \{(p_{Y|X}^{(0)}, \rho^{(0)}), (p_{Y|X}^{(1)}, \rho^{(1)}), \ldots\}$$

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 6 to the present scenario. For example, for single-letter codes, Theorem 6, Part i), provides the following corollary.

Corollary 11: Consider a source (p_S, d) and a class of channels $\mathcal{W}$ such that for every channel in $\mathcal{W}$, $R(\Delta_i) = C_i(\Gamma_i)$ is feasible.[5] Suppose that for the single-letter code (f, g), it is true that

$$0 < I(S; \hat{S}^{(i)}) = I(X; Y^{(i)}) < C_0^{(i)}$$

for all i. The single-letter code (f, g) is universal if and only if for all i

$$\rho^{(i)}(x) \begin{cases} = c_1^{(i)} D(p_{Y|X}^{(i)}(\cdot|x) \| p_Y(\cdot)) + \rho_0^{(i)}, & \\ \qquad\qquad \text{if } p(x) > 0 & \\ \geq c_1^{(i)} D(p_{Y|X}^{(i)}(\cdot|x) \| p_Y(\cdot)) + \rho_0^{(i)}, & \\ \qquad\qquad \text{otherwise} & \end{cases} \tag{22}$$

$$d(s, \hat{s}) = -c_2^{(i)} \log_2 p^{(i)}(s|\hat{s}) + d_0^{(i)}(s), \tag{23}$$

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 6. □

[5]This condition rules out the source–channel pairs of Case 2) of Lemma 1.

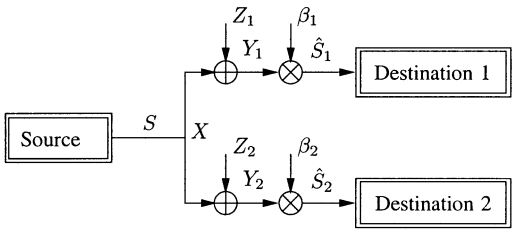

Fig. 4. Single-source Gaussian broadcast using uncoded transmission.

By analogy, one can again include all the special cases of Theorem 6. This is left to the reader. The main reason for studying this particular property of 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 a universal source–channel code achieves the performance of the best source compression followed by the best channel code. However, the significance of universal source–channel codes extends beyond the validity of the separation theorem. Two scenarios under which source–channel codes outperform any code designed according to the separation paradigm are mentioned and illustrated explicitly in the sequel.

Implication 1 (Nonergodic Channels): Let the source–channel 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 source–channel code (F, G) achieves optimal performance, regardless of which particular channel is selected.

Implication 2 (Single-Source Broadcast): Let the source–channel 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 source–channel code (F, G) achieves optimal performance on each channel individually.

Example 4 (Single-Source Gaussian Broadcast): Let the source be independent and identically distributed (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 without loss of generality (w.l.o.g.) $\sigma_1^2 < \sigma_2^2$. This is illustrated in Fig. 4. It is well known 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 [15], [16]. 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

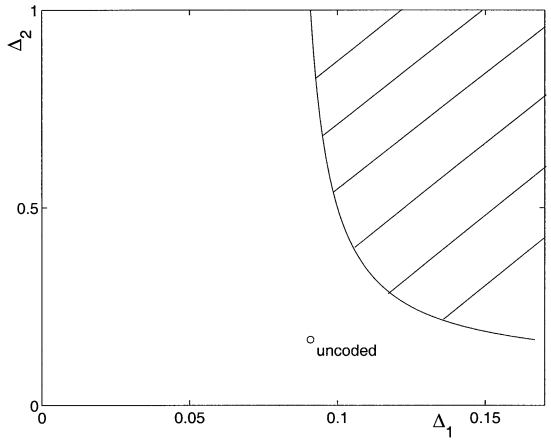

Fig. 5. The distortion achievable by uncoded transmission (circle) versus the distortion region achievable by a transmission scheme based on the separation principle for Example 4. Parameters are $P = 1$, $\sigma_1^2 = 0.1$, and $\sigma_2^2 = 0.2$.

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 [13]. 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 [7]. 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. 5 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.

VI. Conclusion and Extensions

To code, or not to code: that is the question. Undoubtedly, "not to code" is very appealing when it leads to an optimal cost–distortion tradeoff, since it involves the smallest possible delay and complexity. Optimality is a matter of matching up six quantities, namely, the source (p_S, d), the channel $(p_{Y|X}, \rho)$, and the encoder–decoder pair (F, G). Various approaches can lead to such a match.

From a traditional point of view, one can think of the source and the channel as being fixed. Then, one has to design the encoder and the decoder in such a way that they match up the source and the channel probabilistically. Although we do not have a specific design procedure, one expects that matching up probabilities can be a much simpler task than designing good source and channel codes.

If the engineer gets to design a complete system (like nature did in the case of neural communication), then conceivably one can design the source and the channel in such a way that they are already matched, or that they can be matched with a low-complexity encoder and decoder. This is illustrated by Example 2.

Furthermore, the separation principle is limited to ergodic point-to-point communication. Interestingly, very simple source–channel codes perform optimally in certain nonergodic and multiuser communication scenarios. For example, a simple single-source broadcast situation was shown to have this property. It was shown for this example that the separation-based approach leads to a strictly suboptimal performance. Therefore, another promising extension of the results of this paper, and in particular of the condition of probabilistic matching, is to source–channel networks. We have studied one extension into this direction in [14], [17], [18]. As another step into this direction, the approach developed in this paper has been extended to scenarios with side information by Pradhan, Chou, and Ramchandran in [19], and by Merhav and Shamai in [20].

Appendix I
Proofs of Lemmas 3 and 4

Proof of Lemma 3

This lemma appears as Problem 2 in [6, p. 147], and its proof is a consequence of [6, Theorem 3.4]. In the following, we prove the sufficiency of the formula for $\rho(x)$ using a slightly different approach. Our proof extends to continuous alphabets under the appropriate technical conditions.

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

$$I'_{p_X}(x) = D(p_{Y|X}(\cdot|x)||p_Y) \tag{24}$$

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, $\langle f, g\rangle = \sum_x f(x)g(x)$. With this notation, we may write

$$D(p_{Y|X}(\cdot|x)||p_Y) = \left\langle p_{Y|X}, \log_2 \frac{p_{Y|X}}{p_Y} \right\rangle_y$$

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 $\tilde{p}_X$

$$I_{\tilde{p}_X}(X;Y) - I_{p_X}(X;Y) \le \langle \tilde{p}_X - p_X, I'_{p_X}\rangle.$$

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

$$\begin{aligned}
&\langle \tilde{p}_X, I'_{p_X}\rangle - I_{\tilde{p}_X}(X;Y) \\
&\quad = \langle \tilde{p}_X, I'_{p_X}\rangle - \langle \tilde{p}_X, I'_{\tilde{p}_X}\rangle = \langle \tilde{p}_X, I'_{p_X} - I'_{\tilde{p}_X}\rangle \\
&\quad = \langle \tilde{p}_X, D(p_{Y|X}||p_Y) - D(p_{Y|X}||\tilde{p}_Y)\rangle \\
&\quad = \left\langle \tilde{p}_X, \left\langle p_{Y|X}, \log_2 \frac{\tilde{p}_Y}{p_Y} \right\rangle_y \right\rangle_x \\
&\quad \overset{(a)}{=} \left\langle \langle \tilde{p}_X, p_{Y|X}\rangle_x, \log_2 \frac{\tilde{p}_Y}{p_Y} \right\rangle_y \\
&\quad = \left\langle \tilde{p}_Y, \log_2 \frac{\tilde{p}_Y}{p_Y} \right\rangle_y = D(\tilde{p}_Y||p_Y) \ge 0
\end{aligned} \tag{25}$$

where (a) is a change of summation (or integration) order and the inequality follows from the fact that the Kullback–Leibler distance is nonnegative. Lemma 3 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

$$E_{\tilde{p}_X}\rho(X) \le E_{p_X}\rho(X). \tag{26}$$

For any $\lambda \ge 0$

$$\begin{aligned} I_{p_X}(X;Y) - I_{\tilde{p}_X}(X;Y) &\ge \langle p_X - \tilde{p}_X,\, I'_{p_X}\rangle \\ &\ge \langle p_X - \tilde{p}_X,\, I'_{p_X} - \lambda\rho\rangle \end{aligned} \tag{27}$$

where the first inequality is the preceding lemma, and the second follows by assumption on $\tilde{p}_X$. If $\lambda\rho(x) = I'_{p_X}(x) + c$ for all x with $p(x) > 0$, 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. □

Proof of Lemma 4

This lemma appears in [6, Problem 3, p. 147], and its proof is a consequence of [6, Theorem 3.7]. In the following, we prove the sufficiency of the formula for $d(s,\,\hat{s})$ using a slightly different approach. Our proof extends to continuous alphabets under the appropriate technical conditions.

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

$$I'_W(s,\,\hat{s}) = \log_2 \frac{W(\hat{s}|s)}{p_{\hat{S}}(\hat{s})} \tag{28}$$

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 $\sum_s \sum_{\hat{s}} p_S(s)W(\hat{s}|s)I'_W(s,\,\hat{s})$. In the proof, we use the following auxiliary lemma.

Lemma: For any W and $\tilde{W}$

$$I_{\tilde{W}}(S;\hat{S}) - I_W(S;\hat{S}) \ge \langle p_S\tilde{W} - p_S W,\, I'_W\rangle.$$

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

$$\begin{aligned} & I_{\tilde{W}}(S;\hat{S}) - \langle p_S\tilde{W},\, I'_W\rangle \\ &= \left\langle p_S\tilde{W},\, \log_2 \frac{\tilde{W}}{\tilde{p}_{\hat{S}}}\right\rangle - \left\langle p_S\tilde{W},\, \log_2 \frac{W}{p_{\hat{S}}}\right\rangle \\ &= \left\langle p_S\tilde{W},\, \log_2 \frac{\tilde{V}}{p_S}\right\rangle - \left\langle p_S\tilde{W},\, \log_2 \frac{V}{p_S}\right\rangle \\ &= \left\langle p_{\hat{S}}\tilde{V},\, \log_2 \frac{\tilde{V}}{V}\right\rangle = \langle p_{\hat{S}},\, D(\tilde{V}\|V)\rangle \ge 0 \end{aligned} \tag{29}$$

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 Lemma 4.

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

$$E_{p_S\tilde{W}}d(S,\,\hat{S}) \le E_{p_S W}d(S,\,\hat{S}). \tag{30}$$

For any $\lambda > 0$

$$\begin{aligned} I_{\tilde{W}}(S;\hat{S}) - I_W(S;\hat{S}) &\ge \langle p_S\tilde{W} - p_S W,\, I'_W(s,\,\hat{s})\rangle \\ &\ge \langle p_S\tilde{W} - p_S W,\, I'_W + \lambda d\rangle \end{aligned} \tag{31}$$

where the first inequality is the preceding lemma, and the second follows by assumption on $\tilde{W}$. If

$$\lambda d(s,\,\hat{s}) = -I'_W(s,\,\hat{s}) + \tilde{d}_0(s)$$

for all pairs $(s,\,\hat{s})$ with $p(s,\,\hat{s}) > 0$, 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 (10).

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 $E_{\tilde{W}}d(S,\,\hat{S}) \le E_W d(S,\,\hat{S})$, regardless of d. □

Appendix II
Proofs of Lemmas 7 and 8

Proof of Lemma 7

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: $\epsilon = p_{Y|X}(1|0)$, $\delta = p_{Y|X}(0|1)$, $p_X(x=0) = \overline{\pi}$, 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 6 applies. Hence, it is *necessary* that $d(s,\,\hat{s})$ be chosen according to (10); 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 ϵ, δ, $\overline{\pi}$, and π, the latter implies that

$$\pi = \sqrt{(\epsilon(1-\epsilon))/(\delta(1-\delta))}\,\overline{\pi}.$$

Since, moreover, $\pi + \overline{\pi} = 1$, we find

$$\pi = \frac{1}{1 + \sqrt{(\delta(1-\delta))/(\epsilon(1-\epsilon))}}. \tag{32}$$

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:

$$\begin{aligned} \frac{d}{d\pi} I(X;Y) = (\epsilon+\delta-1)\log_2 \frac{1-((1-\pi)(1-\epsilon)+\pi\delta)}{(1-\pi)(1-\epsilon)+\pi\delta} \\ + H_b(\epsilon) - H_b(\delta) = 0. \end{aligned} \tag{33}$$

Plugging in π from above yields

$$2^{\frac{H_b(\delta)-H_b(\epsilon)}{1-\delta-\epsilon}} = \frac{(1-\epsilon)\sqrt{\delta(1-\delta)} + \delta\sqrt{\epsilon(1-\epsilon)}}{\epsilon\sqrt{\delta(1-\delta)} + (1-\delta)\sqrt{\epsilon(1-\epsilon)}}. \tag{34}$$

Clearly, equality holds if $\epsilon = \delta$ (and thus $\overline{\pi} = \pi$), 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- and the left-hand side of (34), respectively. The left-hand side is convex and decreasing in ϵ. For $0 \leq \epsilon \leq 1-\delta$, the right-hand side is also convex and decreasing. Hence, at most two 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. □

Proof of Lemma 8

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 6 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

$$p(s|\hat{s}) = \begin{cases} 2^{-d_0(s)}, & s = \hat{s} \\ 2^{-c_2 - d_0(s)}, & s \neq \hat{s}. \end{cases} \tag{35}$$

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

$$p(s|\hat{s}) = \begin{cases} \alpha, & s = \hat{s} \\ \frac{1-\alpha}{L-1}, & s \neq \hat{s}. \end{cases} \tag{36}$$

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

$$p(x|y) = \begin{cases} \alpha, & y = x \\ \frac{1-\alpha}{L-1}, & y \neq x. \end{cases} \tag{37}$$

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. □

ACKNOWLEDGMENT

The authors thank the reviewers for many useful remarks, and for pointing out that Lemmas 3 and 4 appear (in a different shape) in [6]. We would like to thank Prof. Kannan Ramchandran for initial discussions, and Prof. Emre Telatar for various suggestions; in particular, the basic idea leading to Theorem 10 came up during a discussion of the authors with him. The authors would also like to acknowledge Prof. Toby Berger for suggesting to look into biological communication and for sharing his insights on this topic.

REFERENCES

[1] M. Gastpar, B. Rimoldi, and M. Vetterli, "To code or not to code," in *Proc. IEEE Int. Symp. Information Theory*, Sorrento, Italy, June 2000, p. 236.
[2] ——, "On source/channel codes of finite block length," in *Proc. IEEE Int. Symp. Information Theory*, Washington, DC, June 2001, p. 261.
[3] C. E. Shannon, "A mathematical theory of communication," *Bell Sys. Tech. J.*, vol. 27, pp. 379–423, 623–656, 1948.
[4] R. G. Gallager, *Information Theory and Reliable Communication*. New York: Wiley, 1968.
[5] R. J. McEliece, *The Theory of Information and Coding*, ser. Encyclopedia of Mathematics and its Applications. Reading, MA: Addison-Wesley, 1977.
[6] I. Csiszár and J. Körner, *Information Theory: Coding Theory for Discrete Memoryless Systems*. New York: Academic, 1981.
[7] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.
[8] S. Vembu, S. Verdú, and Y. Steinberg, "The source-channel separation theorem revisited," *IEEE Trans. Inform. Theory*, vol. 41, pp. 44–54, Jan. 1995.
[9] F. Jelinek, *Probabilistic Information Theory*. New York: McGraw-Hill, 1968.
[10] T. J. Goblick, "Theoretical limitations on the transmission of data from analog sources," *IEEE Trans. Inform. Theory*, vol. IT-11, pp. 558–567, Oct. 1965.
[11] B. Rimoldi, "Beyond the separation principle: A broader approach to source-channel coding," in *Proc. 4th Int. ITG Conf. Source and Channel Coding*, Berlin, Germany, Jan. 2002.
[12] T. Berger, "Living information theory (Shannon Lecture)," presented at the IEEE International Symposium on Information Theory, Lausanne, Switzerland, July 2002.
[13] ——, *Rate Distortion Theory: A Mathematical Basis for Data Compression*. Englewood Cliffs, NJ: Prentice-Hall, 1971.
[14] M. Gastpar, "To code or not to code," Ph.D. dissertation, Ecole Polytechnique Fédérale (EPFL), Lausanne, Switzerland, 2002.
[15] W. H. R. Equitz and T. M. Cover, "Successive refinement of information," *IEEE Trans. Inform. Theory*, vol. 37, pp. 460–473, Mar. 1991.
[16] B. Rimoldi, "Successive refinement of information: Characterization of the achievable rates," *IEEE Trans. Inform. Theory*, vol. 40, pp. 253–259, Jan. 1994.
[17] M. Gastpar and M. Vetterli, "On the capacity of wireless networks: The relay case," in *Proc IEEE INFOCOM 2002*, New York, NY, June 2002.
[18] ——, "On the capacity of large Gaussian relay networks," *IEEE Trans. Inform. Theory*, submitted for publication.
[19] S. S. Pradhan, J. Chou, and K. Ramchandran, "Duality between source and channel coding with side information," *IEEE Trans. Inform. Theory*. Univ. Calif., Berkeley, UCB/ERL Tech. Memo.n M01/34, Dec. 2001, submitted for publication.
[20] N. Merhav and S. Shamai (Shitz), "On joint source-channel coding for the Wyner–Ziv source and the Gel'fand–Pinsker channel," *IEEE Trans. Inform. Theory*, submitted for publication.