Channel Coding in the Presence of Side Information: Subject Review

Channel Coding in the Presence of Side Information: Subject Review

Guy Keshet

Department of Electrical Engineering
Technion – Israel Institute of Technology
Haifa 32000, Israel

guyk@tx.technion.ac.il

Yossef Steinberg

Department of Electrical Engineering
Technion – Israel Institute of Technology
Haifa 32000, Israel

ysteinbe@ee.technion.ac.il

Neri Merhav

Department of Electrical Engineering
Technion – Israel Institute of Technology
Haifa 32000, Israel

merhav@ee.technion.ac.il

now

Contents

Abstract

In this paper we review concepts and methods of communication systems equipped with side information. We focus on the channel coding problem, where side information is available to the transmitter in either a causal or non-causal manner, and we also consider the source coding problem with side information at the receiver.

We first summarize the main results for channels with causal/non-causal side information and the associated capacity formulas. Next, we consider specific channel models, such as Costa's dirty-paper model, the AWGN channel model with fading and the modulo additive noise channel. Further, we provide applications to the models considered here, in particular, we present the watermarking problem and the Gaussian MIMO broadcast channel. We also consider algorithms for the calculation of the channel's capacity, and practical coding schemes for the communication systems explored in this paper. Finally, we study several related information-theoretic problems and present both the Wyner-Ziv and the Slepian-Wolf problems. The source coding problems and the channel coding problems, are presented in a unified version and the duality between the problems is presented. We also present extensions for the MAC and broadcast channel models, to the case where they are controlled by a state process, and consider several hybrid models, e.g., joint source-channel coding for the Wyner-Ziv source and the Gel'fand-Pinsker channel, and the achievable tradeoff between message and state information rates.

1

Introduction

This paper gives an overview of results pertaining to the capacity of a channel whose conditional output probability distribution depends on a state process, and where the channel state information (CSI) signal (also referred to as "side information") is available at the transmitter (CSIT) or at the receiver (CSIR) or at both ends. These channels have been widely studied over the years and they can serve for modeling in a wide range of problems, depending on some assumptions regarding the channel state and on the availability and quality (clean or noisy) of the side information at the transmitter and/or the receiver. For CSI available at the transmitter, we will distinguish between channels where the CSIT is causal and channels where it is non-causal. In the causal case, the transmission, at every time instant, depends only on the past and present CSI, whereas in the non-causal case, the transmitter knows in advance the realization of the entire state sequence from the beginning to the end of the block. The causal CSIT channel model was introduced in 1958 by Shannon [105], who also found its capacity. The non-causal CSIT channel model was introduced in 1974 by Kusnetsov and Tysbakov [75], and its capacity was found in 1980 by Gel'fand and Pinsker [56]. Regarding CSIR, we will not distinguish between causal and non-causal CSI at the receiver, because the receiver, in this case, waits until the end of the block anyway, before decoding. Channels with CSIR are studied in [101], [13], [63].

A crucial element in the solutions of the information theory problems presented in this work, including the Slepian-Wolf problem [107], the Wyner-Ziv problem [135], and the Gel'fand-Pinsker problem [56], is the "binning" concept. A binning scheme divides a set of codewords into subsets or "bins", such that the codewords in each bin are as far apart as possible. These bins are constructed at random and each bin is assigned to a different message index. When given with a message to be transmitted in a channel coding scenario, we use only the codewords from the bin with the same index as our message, and we choose a codeword which is the closest to (or jointly typical with) the side information vector. We present this scheme in more details later on.

One interesting example which can be modeled as a channel with non-causal transmitter CSI, is a computer memory with defective cells [75], [76], [74], [122], [123], [63], [62]. In this example, the process of storing to this memory suffers from random errors caused by noise. A computer memory may also have some cells which are defected, e.g., a memory cell whose stored value seems to be fixed regardless of the input, e.g., the cell is "stuck at one" or "stuck at zero". The location and values ("defect information") of the defective cells may be found by storing the all-one bit pattern (or the all-zero bit pattern) in the memory, reading the contents of the memory and comparing it with the stored pattern. This process may be repeated several times in order to exclude the effect of random errors. If this process is not repeated, we can refer to the "defect information" received from this process as a noisy version of the "defect information". By knowing in advance the "defect information", we can design codes that are more efficient than the usual error correcting codes. In this example, the "defect information" plays the role of the channel states and transmitter's CSI.

Another important example for a channel model with transmitter CSI, is the power constrained Gaussian additive noise channel model with additive interference which is known non-causally to the transmitter as side information and is statistically independent of the noise.

In this example, the channel output is given by

$$Y_i = X_i + S_i + Z_i, \quad i = 1, 2, ..., N, \tag{1.1}$$

where $S^N = (S_1, ..., S_N)$ is the i.i.d. side information sequence (the interference) distributed as $S^N \sim \mathcal{N}(0, QI)$ (I being the identity matrix), Q is the side information variance, and $Z^N = (Z_1, ..., Z_N)$ is the i.i.d. noise sequence distributed as $Z^N \sim \mathcal{N}(0, BI)$, B being the noise variance. Based on the message to be sent and on the interference samples $S_1, ..., S_N$, the encoder sends a codeword $X^N = (X_1, ...X_N)$ which must satisfy the power constraint

$$\frac{1}{N}\sum_{n=1}^{N} E(X_n^2) \leq \Gamma, \tag{1.2}$$

where $\Gamma > 0$ is a given constant. This channel with non-causal CSIT, is known as Costa's channel [34], which has recently received much attention, as it has been proven useful for modeling in various communication problems, among them precoding for intersymbol interference (ISI) channel, digital watermarking, and various broadcasting schemes. Again, this problem can be divided to the causal and non-causal CSIT scenarios. When the CSIT is non-causal, this problem is known as "writing on dirty paper" (WDP), or the "dirty-paper" problem. When the CSIT is causal, the problem is known as "writing on dirty tape" (WDT), or the "dirty-tape" problem [15]. The dirty-tape setting is often used to describe cases where we restrict the encoder to be causal in order to reduce the implementation complexity compared to dirty-paper implementation.

Costa showed, for the dirty-paper setting, that the capacity of this channel is the same as if the interference was not present or, equivalently, if it was also known at the decoder and could be subtracted off, i.e., the capacity is given by

$$C = \frac{1}{2}\log(1 + \frac{\Gamma}{B}). \tag{1.3}$$

This surprising result is another reason why this problem has received so much attention.

Yet another interesting example for a channel model with transmitter CSI, is digital watermarking [20]-[24], [31], [38], [44], [45], [46], [79],

[80], [83], [87], [89], [108], [109]. Digital watermarking is the process of embedding a message within a host signal to form a composite (watermarked) signal. The embedding must not cause a noticeable distortion relative to the host signal. On the other hand, the embedding should be robust to attacks on the watermarked signal. In some applications, these attacks are the result of standard signal processing operations. In other cases they are malicious. The digital watermarking problem can be modeled as a channel whose conditional output probability depends on a state process, and where the transmitter has channel state information. Here the input power constraint is replaced by a constraint on the distortion between the channel input and the host signal. This expresses the requirement that the embedding does not cause a noticeable distortion to the host signal. The dirty-paper problem may be used to model watermarking in a Gaussian environment. We can model watermarking as a communication system in which the transmitter, which must satisfy a distortion constraint with regard to the host signal, sends a watermark message through a noisy channel with the host signal playing the role of state. The state (the host signal), is available to the transmitter, and therefore it can be used by the transmitter as side information, just like in the dirty-paper problem.

The purpose of this paper is to give an overview of the subject of coding for channels with side information, both from the theoretical point of view (capacity, fundamental limits, duality with source coding), and the practical point of view, including coding and decoding techniques, structures of classes of codes, application aspects, etc.

The outline of this work is as follows. In Section 2, we specify the notation conventions that will be used and we formalize the model of a channel with CSIT. Section 3 describes the main theoretical results pertaining to the capacity of channels with side information. In Section 4, we describe some specific channel models for various problems. In Section 5, we present several applications for the models which were presented in the preceding sections. Section 6 presents related problems which are linked to the problem of coding for a channel with side information. These problems include source coding dual, which is the Wyner-Ziv problem, the Gaussian vector broadcast channel and multiuser channels. In Section 7, we describe algorithms for computation of the capacity of channels that were presented in the previous sections, as well as several coding techniques for these channels. Section 8 concludes the paper.

2

Notation and Model Formulation

Throughout this paper, scalar random variables will be denoted by capital letters, their sample values will be denoted by the respective lower case letters, and their alphabets will be denoted by the respective calligraphic letters. A similar convention will apply to random vectors and their sample values, which will be denoted with the same symbols superscripted by the dimension. Thus, for example, W^k will denote a random k-vector $(W_1, ..., W_k)$, and $w^k = (w_1, ..., w_k)$ will be a specific vector value in $\mathcal{W}^k$, the k-th Cartesian power of $\mathcal{W}$. The notations w_i^j and W_i^j, where i and j are integers and $i \leq j$, will designate segments $(w_i, \ldots, w_j)$ and $(W_i, \ldots, W_j)$, respectively, where for $i = 1$, the subscript will be omitted (as above). For $i > j$, w_i^j (or W_i^j) will be understood as the null string. Sequences without specifying indices are denoted by $\{\cdot\}$. The cardinality of a set $\mathcal{W}$ will be denoted by $|\mathcal{W}|$. The empty set will be denoted by $\emptyset$. $\mathbb{R}$ stands for the set of real numbers, $\mathbb{C}$ for the set of complex numbers and $\mathbb{R}_+$ for the set of positive reals.

Sources and channels will be denoted generically by the letter P subscripted by the name of the RV and its conditioning, if applicable, e.g., $P_U(u)$ is the probability function of U at the point $U = u$, $P_{Z|S}(z|s)$ is the conditional probability of $Z = z$ given $S = s$, and so on.

Whenever clear from the context, these subscripts will be omitted. A sequence of finite-dimensional distributions will be denoted by bold letters, for example, $\mathbf{\Lambda} = \{P_{W^N}(w^N)\}_{N=1}^{\infty}$. $U(a,b)$ stands for the uniform distribution on the interval (a,b). Information theoretic quantities like entropies, divergences, and mutual informations will be denoted following the usual conventions of the information theory literature, e.g., $H(U^N)$, $I(Z^n;W^k)$, $D(P_{Y|XS}\|P_Y)$, and so on.

We consider the channel depicted in Fig.2.1, whose input, state, and output, at time index n, are $X_n \in \mathcal{X}$, $S_n \in \mathcal{S}$ and $Y_n \in \mathcal{Y}$, respectively, where $\mathcal{X}, \mathcal{Y}, \mathcal{S}$ are the corresponding alphabets. Unless otherwise specified, we assume throughout that the channel and state are memoryless, i.e.,

$$P_{Y^N|X^N,S^N}(y^N|x^N,s^N) = \prod_{n=1}^{N} P_{Y|X,S}(y_n|x_n,s_n). \tag{2.1}$$

$$P_{S^N}(s^N) = \prod_{n=1}^{N} P_S(s_n). \tag{2.2}$$

Many variants of this channel model have been studied. In some of these variants, the state sequence is generated in a different manner. The first example is the compound channel, in which the channel state is fixed during the course of transmission. Another example is the arbitrary varying channel (AVC), in which, the states vary arbitrarily from

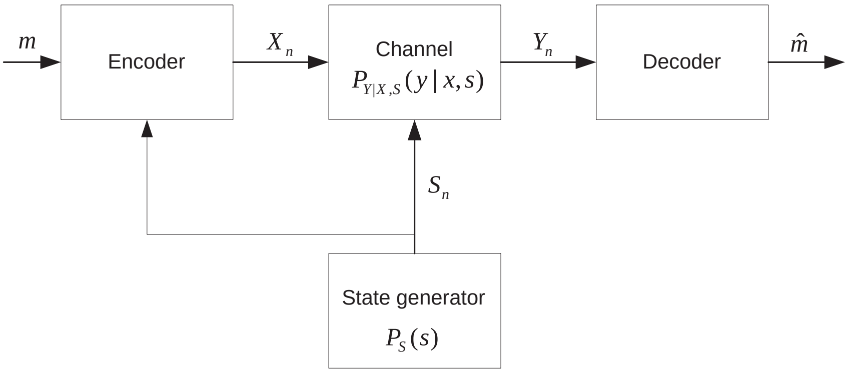

Fig. 2.1 CSIT channel model.

symbol to symbol during the transmission (the states are not generated stochastically, but in an arbitrary deterministic manner). Usually, in AVC channel models, the state sequence is unknown to the transmitter [77]. Ahlswede [1] analyzed the case where this sequence is known to the transmitter. Winshtok and Steinberg [134] considered the problem of joint source channel coding for the arbitrary varying source transmitted over an AVC where the state sequence is known to the transmitter. This problem combines the model considered by Ahlswede for an AVC and the problem of an arbitrary varying Wyner-Ziv source. Other examples include channels in which the state sequence is a more general stochastic process (e.g., the Gilbert-Elliot channel) or some hybrid situations like a compound Gilbert-Elliot channel. More details on many of these channel models can be found in [77].

The channel input may be subjected to a transmission-cost constraint

$$E\left\{\sum_{i=1}^{N}\phi(X_i)\right\} \leq N\Gamma, \tag{2.3}$$

where ϕ is a given function from $\mathcal{X}$ to $\mathbb{R}^+$ and $\Gamma \geq 0$ is a prescribed constant. In this problem, the transmitter sends a random message M to the receiver in N uses of the channel. The random message M is uniformly distributed in $\mathcal{M} = \{1, ..., 2^{NR}\}$, R being the code rate. The transmitter is also provided with causal or non-causal CSI.

An $(N, 2^{NR})$ code consists of the following:

1. A set of N encoding functions $f_n : \mathcal{M} \times \mathcal{S}^n \longrightarrow \mathcal{X}$, for $n = 1, ..., N$, such that $x_n = f_n(m, s^n)$ and equation (2.3) is satisfied, where m ranges over $\mathcal{M}$, and s^n is the vector of the channel states realizations up to time n. For the non-causal case, a mapping $f : \mathcal{M} \times \mathcal{S}^N \longrightarrow \mathcal{X}^N$, such that $x^N = f(m, s^N)$, or equivalently, $x_n = f_n(m, s^N)$, $n = 1, ..., N$, and again equation (2.3) is satisfied.

2. A decoding function $g_n : \mathcal{Y}^N \longrightarrow \mathcal{M}$, such that the decoded message is $\hat{m} = g(y^N)$.

The average probability of error is given by

$$P_e = \frac{1}{2^{NR}} \sum_{m=1}^{2^{NR}} \sum_{\{y^N : g(y^N) \neq m\}} \sum_{s^N} P_{S^N}(s^N) P_{Y^N|X^N,S^N}(y^N|x^N, s^N), \tag{2.4}$$

where x^N depends on the inputs to the encoder in a manner that depends on whether it is the causal or non-causal case, as described above.

We define an achievable rate and capacity in the following way:

Definition 1. A rate R is said to be achievable if for every $\epsilon > 0$, there exists an integer $N_0 = N_0(\epsilon)$ such that for all $N > N_0$, there exists an $(N, 2^{NR})$ code with probability of error $P_e \leq \epsilon$.

Definition 2. The capacity is the supremum of all achievable rates.

3

Fundamental Results

3.1 Causal Side Information

The causal side information case was first investigated by Shannon [105] in 1958. Shannon showed that the capacity of the channel, depicted in Fig. 2.1, is equal to the capacity of an ordinary discrete memoryless channel (DMC), with the same output alphabet and an input alphabet of size $|\mathcal{X}|^{|\mathcal{S}|}$. The input letters of the new channel consist of all mappings from $\mathcal{S}$ to $\mathcal{X}$. Any coded communication system for the equivalent channel, without side information, can be translated into a coded communication system for the original channel with the same probability of error. The equivalent channel and its relation to the original channel are depicted in Fig. 3.1.

The equivalent channel input variable is denoted by T, and is called "strategy". Every input letter t of this channel is a particular function from the state alphabet $\mathcal{S}$ to the input alphabet $\mathcal{X}$ of the original channel (i.e., $t : \mathcal{S} \to \mathcal{X}$). The alphabet of T, which will be denoted by $\mathcal{T}$, consists of all $|\mathcal{X}|^{|\mathcal{S}|}$ distinct functions from the state set to the input alphabet. To achieve capacity, it is enough to use at most $|\mathcal{Y}|$ of them, which is the number of output letters (see [36] Section 8.3). The

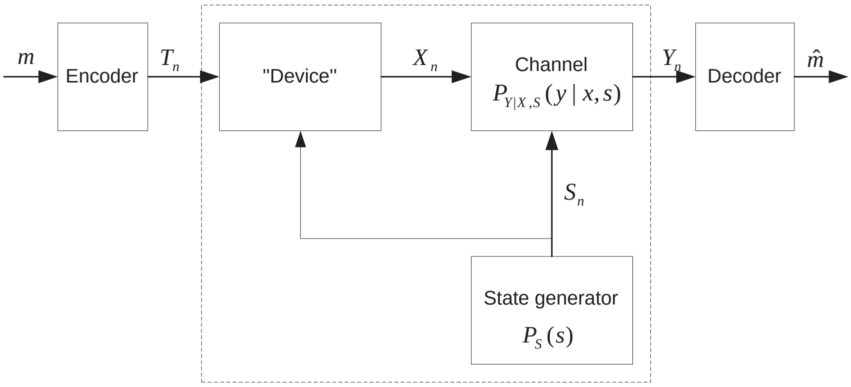

Fig. 3.1 Shannon's Equivalent channel.

"device" shown in Fig. 3.1 takes T_n and S_n as the inputs and produces $X_n = T_n(S_n)$ for a function T_n and state S_n. The equivalent channel is characterized by a set of conditional output probability functions $\{P_{Y|T}(y|t) : t \in \mathcal{T}, y \in \mathcal{Y}\}$ such that

$$P_{Y^N|T^N}(y^N|t^N) = \prod_{n=1}^{N} P_{Y|T}(y_n|t_n), \tag{3.1}$$

where

$$P_{Y|T}(y|t) = \sum_{s\in\mathcal{S}} P_S(s)P_{Y|X,S}(y|t(s),s). \tag{3.2}$$

Shannon used strategies in which the input to the channel depends only on the current state of the channel and not on previous states. He showed that using only this type of strategies, the capacity of the equivalent channel is equal to the capacity of the original channel. The capacity of this channel is therefore the capacity of the equivalent DMC

$$C = \max_{P_T(\cdot)} I(T;Y), \tag{3.3}$$

where $P_T(\cdot)$ is a probability distribution of T, which is independent of the state S and the maximization is taken over all joint distributions satisfying

$$P_{T,S,X,Y}(t,s,x,y) = P_S(s)P_T(t)\delta(x,t(s))P_{Y|X,S}(y|x,s), \tag{3.4}$$

where $\delta(x,t(s)) = 1$ for $x = t(s)$, and $\delta(x,t(s)) = 0$ otherwise.

In order to prove the equivalence of the two channels, we first note that from (3.2), it can easily be seen that each code for the equivalent channel can be translated to a code for the original channel with the same probability of error. The translation of codes consists merely of using the input $x_n = t_n(s_n)$ for a function t_n and state s_n. Next, we prove that the rate of the original channel cannot be larger than the capacity of the equivalent channel, thus establishing the converse part of (3.3). Letting $T_n = (M, S^{n-1})$, and C being given by (3.3), we can write:

$$\begin{aligned} NR - H(M|Y^N) &= H(M) - H(M|Y^N) \\ &= I(M;Y^N) \\ &= \sum_{n=1}^{N} I(M;Y_n|Y^{n-1}) \\ &\leq \sum_{n=1}^{N} I(M,Y^{n-1};Y_n) \\ &\leq \sum_{n=1}^{N} I(M,S^{n-1};Y_n) \\ &= \sum_{n=1}^{N} I(T_n;Y_n) \\ &\leq NC, \end{aligned} \tag{3.5}$$

where the second inequality follows from the data processing inequality using the Markov chain $(M,Y^{n-1}) \longrightarrow (M,S^{n-1}) \longrightarrow Y_n$ and the last inequality follows since T_n is independent of S_n. If the message is recovered reliably, then $H(M|Y^N)$ must be small (Fano's inequality). Thus, the rate R cannot be larger than C.

The capacity in equation (3.3) is expressed in terms of strategies. This might pose some conceptual and practical problems for code construction, especially for large $|\mathcal{S}|$. Large alphabets make the implementation of the encoder and decoder more difficult.

The encoder, in this problem, is able to produce the channel input X_n, when given with $T_n = (M, S^{n-1})$ and S_n, as one should expect in a causal CSIT model.

Shannon's capacity formula can be extended [6] to the case where the alphabets $\mathcal{X}, \mathcal{Y}, \mathcal{S}$ are the real line and where the transmitter is subjected to an average power constraint

$$E\{\phi(X)\} \leq \Gamma, \tag{3.6}$$

and the capacity is given by

$$C = \sup_{P_T(\cdot): E\{\phi(T(S))\} \leq \Gamma} I(T;Y), \tag{3.7}$$

where the expectation is relative to the product distribution $P_{S,T}(s,t) = P_S(s)P_T(t)$, $t \in \mathcal{T}$, $s \in \mathcal{S}$. In this case, the supremum in (3.7) is over all the distributions $P_T(\cdot)$ of the continuous functions $t : \mathcal{S} \longrightarrow \mathcal{X}$, where the alphabets $\mathcal{X}, \mathcal{S}$ are the real line.

Shannon's results have been used to compute the capacity of a discrete modulo-additive noise channel [50] with causal side information at the transmitter, which represents a specific channel whose capacity was determined using (3.3). These results were also used to bound the capacity for the dirty-tape problem [48], which is the causal counterpart of the dirty-paper problem. These problems will be discussed in Section 4.

Shannon's model was extended by Salehi [101] to the case where the transmitter has access to one noisy version of the state information, W_n, and the receiver has access to another version, V_n. We will see later that this model is not really more general than Shannon's model. This communication system is depicted in Fig. 3.2.

Denote by $\mathcal{W}$ and $\mathcal{V}$ the alphabets of W_n and V_n, respectively, and let $\mathcal{T}$ be the set of all $|\mathcal{X}|^{|\mathcal{W}|}$ possible functions from $\mathcal{W}$ to $\mathcal{X}$. The variables T_n, S_n, W_n, V_n and Y_n, $n = 1, ..., N$, take values according to the joint distribution

$$P_{T,S,W,V,Y}(t,s,w,v,y) = P_T(t)P_{S,W,V}(s,w,v)P_{Y|X,S}(y|x = t(w), s), \tag{3.8}$$

for some arbitrary distribution $P_T(t)$, where T is independent of W and V. The capacity of the channel is given by

$$C = \max I(T;Y|V), \tag{3.9}$$

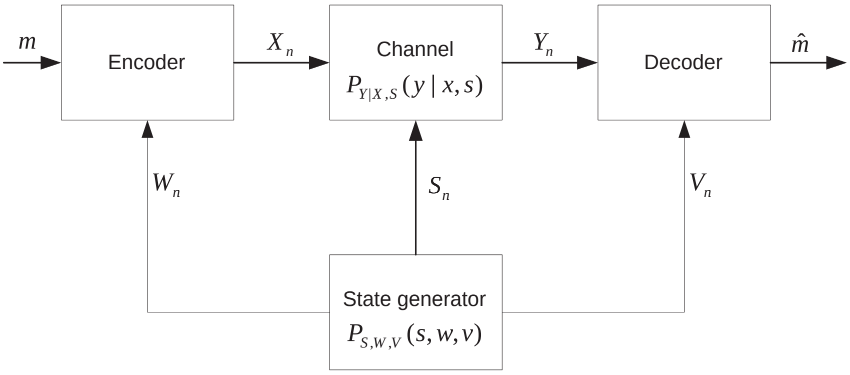

Fig. 3.2 Transmitter and receiver side information channel model.

where the maximum is taken over all joint probability distributions as in (3.8).

Salehi proved his capacity formula (3.9) directly in [101]. Caire and Shamai [13] showed that this result follows from Shannon's result for the capacity of a channel with perfect causal transmitter side information by a simple argument: They considered the receiver side information $\{V_n\}$ as an additional channel output in an equivalent channel. For this channel, the conditional probability distribution of the output (Y, V) is given by

$$\begin{aligned} P'_{Y,V|X,W}(y,v|x,w) &= \sum_s P_{Y,V|X,W,S}(y,v|x,w,s)P_{S|X,W}(s|x,w) \\ &= \sum_s P_{Y|X,S}(y|x,s)P_{S,W,V}(s,w,v)/P_W(w), \end{aligned} \tag{3.10}$$

and $P_W(w) = \sum_{s,v} P_{S,V,W}(s,v,w)$.

This equivalent channel is clearly of the type studied by Shannon, with causal transmitter side information and no receiver side information. The capacity of this channel is therefore $C = \max_{P_T(t)} I(T;Y,V)$. But since V is independent of T we have $I(T;V) = 0$, so that (3.9) follows immediately. It follows then that, as argued earlier, Salehi's model is not really more general than Shannon's original model.

The capacity of this channel is expressed in terms of strategies. We mentioned earlier that this might pose some conceptual and practical problems for code construction, especially for large $|\mathcal{W}|$, that will require very large codebooks. Caire and Shamai [13] found a special case in which the capacity can be expressed without using strategies. In this case, the CSIT is a deterministic function of the CSIR. Let $W_n = g(V_n)$, where $g(\cdot)$ is a deterministic function from $\mathcal{V}$ to $\mathcal{W}$. Then, the channel capacity is given by

$$C = \sum_{w \in \mathcal{W}} p(w) \max_{p(x|w)} I(X;Y|V,W=w). \tag{3.11}$$

This capacity can be achieved by a multiplexed multiple codebook scheme. For each value of $w \in \mathcal{W}$, a codebook of length $(p(w) - \delta)N$, where δ is a small positive number, and rate slightly less than $I(X;Y|V,w)$ is generated i.i.d. according to the probability distribution $p(x|w)$. For the message m, a set of $|\mathcal{W}|$ codewords is selected, one for each codebook. At time index n, if $W_n = w$, the transmitter sends the first not yet transmitted symbol of the w-th codeword. Then the codewords are multiplexed according to the transmitter side information sequence W^N. If $g(\cdot)$ is deterministic, the receiver can demultiplex the received sequence before decoding since it can perfectly recover W^N from V^N. After demultiplexing, the $|\mathcal{W}|$ codewords are independently decoded. The case where W_n is a deterministic function of V_n, can describe a scenario in which the transmitter's side information is obtained via an error-free low rate causal feedback channel from the receiver to the transmitter. Heegard and El Gamal [63] studied a related problem, in which the CSIT and CSIR are subject to rate constraints R_e and R_d, respectively. We will examine this problem is Subsection 3.2.

A special case of Caire and Shamai's result, is the case where the side information available to the encoder is equal to the side information available to the decoder, i.e., $W_n = V_n$. We will then denote this as Z_n, i.e., $W_n \triangleq V_n = Z_n$. In this special case, $\mathcal{T}$ denotes the set of all possible functions from $\mathcal{Z}$ to $\mathcal{X}$. The capacity in this case is given by

$$C = \max I(X;Y|Z), \tag{3.12}$$

where the maximum is taken over all joint probability distributions of the form

$$P_{S,Z,X,Y}(s,z,x,y) = P_{S,Z}(s,z)P_{X|Z}(x|z)P_{Y|X,S}(y|x,s). \tag{3.13}$$

The case where the same side information is available to both the encoder and decoder has been used in [6] to model *private* information embedding which has a wide variety of applications, such as digital watermarking, data hiding and steganography. Information embedding can be viewed as a problem of channel coding with side information. We refer to the information embedding case where both the encoder and decoder have the same side information signal as *private* information embedding, and to the case where only the encoder has side information as *public* information embedding.

We mentioned, in Section 2, that we assume that the states of the channel are generated by a memoryless source of a given distribution (2.2). In other models, the state sequence is not memoryless. We will elaborate on this in Section 6.

3.2 Non-Causal Side Information

In this section, we consider the case of non-causal side information. This model was first investigated by Kuznetsov and Tsybakov [75] in 1974, who considered the problem of coding for a computer memory with defective cells. Here, the positions of the defective cells serve as the channel state information, and they are known non-causally to the encoder.

Kuznetsov and Tsybakov presented coding techniques for this channel, but they have not determined the capacity. The capacity was found in 1980 by Gel'fand and Pinsker [56]. As in Section 2, the channel is stationary and memoryless and the output probability is given by (2.1). The capacity of this channel is given by

$$C = \max_{P_{U,X|S}} [I(U;Y) - I(U;S)], \tag{3.14}$$

where U is an auxiliary random variable with cardinality $|\mathcal{U}| \leq |\mathcal{S}||\mathcal{X}| + 1$, and the maximization is subjected to the constraint that $U \longrightarrow (X,S) \longrightarrow Y$ forms a Markov chain. An auxiliary RV was originally used by Wyner and Ziv [135] for the dual problem of source coding (rate-distortion) with decoder side information under a distortion constraint, and it was at that time a relatively new technique. We will elaborate more on the Wyner-Ziv problem and its duality to the Gel'fand-Pinsker problem in Section 6.

Denoting $R_{GP}(P_{U,X|S}) = I(U;Y) - I(U;S)$, we now present a few properties [56] of this function.

Claim 1. [56]
i) For fixed $P_{U|S}(u|s)$, $R_{GP}(P_{U,X|S})$ is a convex functional of the distribution $P_{X|U,S}(\cdot|u,s)$. ii) For fixed $P_{X|U,S}$, $R_{GP}(P_{U,X|S})$ is a concave functional of the distribution $P_{U|S}(\cdot|s)$.

By Claim 1, the optimum $P_{X|U,S}(\cdot|u,s)$ (in the sense of (3.14)) puts all its mass on a single x for every u, s, and so, there exists a deterministic mapping $f : \mathcal{U} \times \mathcal{S} \longrightarrow \mathcal{X}$ such that $P_{X|U,S}(x|u,s) = 1$ if and only if $x = f(u,s)$. This remains true [6], even when the transmitter is subjected to an average power constraint (2.3).

We will now outline the proof of the direct part of this capacity formula. To this end, we will need to use a random binning technique, which is also being used in the dual problem of source coding with side information [36]. For each message m, generate 2^{NR_0} codewords (forming a bin) $\{u^N(m,1),...,u^N(m,2^{NR_0})\}$ [1] i.i.d. according to the distribution $P_U(u)$. Given the message m and the state sequence s^N, the encoder seeks a codeword in bin m that is jointly typical with s^N, say $u^N(m,j)$. If multiple such codewords in bin m exist, choose the one with the smallest j. If no such j exists, then declare an encoding error. The encoder then creates the input to the channel as $x_i = f(u_i(m,j), s_i), i = 1,...,N$. The decoder finds an $\hat{m}$ and a $\hat{j}$ such that $u^N(\hat{m},\hat{j})$ is jointly typical with the channel output sequence y^N, the decoded message in this case will be $\hat{m}$. If multiple or no such $\hat{m}$ and $\hat{j}$ exist, then declare a decoding error. The probability of encoding failure goes to zero as long as $R_0 > I(U;S)$ and the probability of decoding failure goes to zero as long as $R + R_0 < I(U;Y)$. Thus, the overall probability of error goes to zero as long as $R < I(U;Y) - I(U;S)$.

[1] $u^N(m,j)$ represents a codeword indexed by $m \in \mathcal{M}$, which is the bin index and $j \in \{1,...,2^{NR_0}\}$, which is an index within the bin.

The above encoding procedure which finds a jointly typical codeword $u^N(m,j)$ with s^N in bin m, defines a source coding process [36], thus the collection of codewords in each bin plays the role of a source code. The above decoding procedure, finds a codeword from the entire collection of codewords which is jointly typical with y^N, thus the collection of all the codeword in all the bins plays the role of a channel code. We will make use of these observations later on.

3.3 Relations Between the Causal and Non-Causal Cases

We can rewrite (3.14) and present it in terms of strategies. Equation (3.14) is maximized over $P_{U|S}(u|s)P_{X|U,S}(x|u,s)$ and by Claim 1, as discussed in Subsection 3.2, the optimal $P_{X|U,S}(x|u,s)$ may take values of 0 or 1 only, and thus, X is a deterministic function of U and S. Therefore, we can extend the input alphabet to the set of all functions $t : \mathcal{S} \longrightarrow \mathcal{X}$ in order to eliminate $P_{X|U,S}(x|u,s)$ from the problem. This is very important for the numeric calculation of channel capacity. We shall see in Subsection 7.1 that the new form of the capacity formula will make the numeric algorithms for the computation of the channel capacity more efficient.

The capacity for the non-causal case, in terms of strategies, is given by

$$C = \max_{P_{T|S}(t|s)} [I(T;Y) - I(T;S)], \tag{3.15}$$

where the maximization is taken over all joint distributions satisfying

$$P_{T,S,X,Y}(t,s,x,y) = P_S(s)P_{T|S}(t|s)\delta(x,t(s))P_{Y|X,S}(y|x,s). \tag{3.16}$$

Alternatively, we can rewrite the capacity formula for the causal case (3.3), and present it in terms of auxiliary random variable:

$$C = \max_{P_U(\cdot), f:\mathcal{U}\times\mathcal{S}\longrightarrow\mathcal{X}} I(U;Y), \tag{3.17}$$

where U is an auxiliary random variable and the joint distributions of the random variables S, U, X, Y is given by

$$P_{S,U,X,Y}(s,u,x,y) = P_S(s)P_U(u)\delta(x,f(u,s))P_{Y|X,S}(y|x,s). \tag{3.18}$$

We can think of this auxiliary random variable as indexing a set of functions from $\mathcal{S}$ to $\mathcal{X}$. There are $|\mathcal{X}|^{|\mathcal{S}|}$ possible functions, but we only need to use at most $|\mathcal{S}||\mathcal{X}| + 1$ to achieve capacity in the non-causal case, and $|\mathcal{Y}|$ in the causal case.

The causal side information capacity formula, discussed in Subsection 3.1, can be obtained as a special case of the non-causal one if we restrict U to be independent of S. Let us look at the proof of the direct part in the non-causal case. The encoder, in the causal case, is only given the state sequence until time index i, $i = 1, ..., N$, i.e., S^i. Therefore, the entire state sequence S^N and the codeword U^N will be jointly typical if the distribution of the auxiliary random variable U is independent of S. For the converse part, it was shown in [56] that

$$NR - N\epsilon \leq \sum_{i=1}^{N} I(U_i; Y_i) - I(U_i; S_i), \tag{3.19}$$

where U_i is defined as $U_i = (m, Y^{i-1}, S_{i+1}^N)$ and $(m, Y^{i-1}, S_{i+1}^N) \longrightarrow (X_i, S_i) \longrightarrow Y_i$ forms a Markov chain. Therefore, when the side information is given in a causal manner, U_i is independent of S_i ($(m, Y^{i-1}) \longrightarrow (m, S^{i-1}) \longrightarrow Y_i$ forms a Markov chain), resulting in Shannon's capacity (3.17).

Another important observation is the following. The last step in Gel'fand and Pinsker's proof of the converse part is

$$\sum_{i=1}^{N} I(U_i; Y_i) - I(U_i; S_i) \leq N \max_i [I(U_i; Y) - I(U_i; S)], \tag{3.20}$$

which proves the existence of a RV U, s.t $R \leq I(U; Y) - I(U; S)$. If we have a power constraint, we cannot use (3.20). The index i which maximizes $I(U_i; Y) - I(U_i; S)$, does not have to fulfill the power constraint $E[\phi(X_i)] \leq \Gamma$. Therefore, we have to take a different approach to prove the converse. We define a random variable J uniformly distributed in $\{1, 2, ..., N\}$, $U_J \triangleq (m, Y^{J-1}, S_{J+1}^N)$, and $U \triangleq (U_J, J)$. Now, the proof of the converse part will be complete with these steps:

$$\begin{aligned} R-\epsilon &\le \frac{1}{N}\sum_{i=1}^{N} I(U_i;Y_i)-I(U_i;S_i) \\ &= I(U_J;Y_J|J)-I(U_J;S_J|J) \\ &= I(U;Y_J|J)-I(U;S_J|J) \\ &= I(U;Y_J|J)-I(U;S) \\ &\le I(U;Y)-I(U;S), \end{aligned} \tag{3.21}$$

where the last equality follows from the fact that S is independent of J, and the last inequality since conditioning reduces entropy.

We can also link Gel'fand and Pinsker's result to the case where both the transmitter and receiver have the same side information. This case is a special case of equation (3.12), in which both the transmitter and receiver have the same noisy version of the side information- $W = V = Z$. In our case, $Z = S$, and the capacity is equal to

$$C = \max_{P_{X|S}(x|s)} I(X;Y|S). \tag{3.22}$$

In general, the capacity in this case is bigger than when we have only perfect non-causal CSIT and no CSIR (Gel'fand and Pinsker's setting), but this capacity can be equal to Gel'fand and Pinsker's capacity formula (3.14), if and only if [95] the distribution achieving the maximum in (3.22) (say $P^*_{X,S,Y}$) can be represented in the form of having the auxiliary variable $U : P_{U,X,S,Y}(u,x,s,y)$ such that:

- $\sum_u P_{U,X,S,Y}(u,x,s,y) = P^*_{X,S,Y}(x,s,y)$,
- the channel input x can be represented as $x = f^*(u,s)$, where $f^*(\cdot,\cdot)$ is the encoding mapping which maximizes (3.14), and
- the following two Markov chains are satisfied: $U \longrightarrow (X,S) \longrightarrow Y$ and $S \longrightarrow Y \longrightarrow U$.

This is needed because the objective function in (3.22), $I(X;Y|S) = I(U;Y|S)$ from the first Markov chain and the objective function in (3.14), $I(Y;U) - I(U;S) = I(U;Y|S)$ from the second Markov chain. The second Markov chain can be interpreted as follows: all the dependency between the side information and the auxiliary variable U is captured through the channel output Y. So there is no loss in performance in terms of getting information about U, even if we do not have CSIR. These conditions link (3.14) and (3.22). They present another way to determine if (3.14) equals (3.22). The obvious method to determine if they are equal, is to separately calculate the capacity of (3.14) and (3.22). In the dirty-paper problem, as we shall see in Subsection 4.1, these conditions are met and the capacity with CSIT is equal to the capacity with CSIT and CSIR.

3.4 Modifications and Extensions

We can also extend (3.15) to the case where the alphabets $\mathcal{X}, \mathcal{Y}, \mathcal{S}$ are the real line (see [6],[48]) and where the transmitter is subject to an average power constraint to yield

$$C = \sup_{P_{T|S}(t|s):E\{\phi(T(S))\}\leq\Gamma} [I(T;Y) - I(T;S)], \tag{3.23}$$

where $t \in \mathcal{T}$ is the set of all possible mappings $t : \mathcal{S} \longrightarrow \mathcal{X}$.

Gel'fand and Pinsker's results were used by Costa [34] to analyze the additive interference channel, which is known also as the dirty-paper channel, as mentioned earlier. In the Introduction, the importance of this result and its relevance to analyzing many communication problems was discussed. This result was also used to analyze digital watermarking and the capacity of a computer memory with defects. Costa's work and other related results will be presented in Section 4.

Cover and Chiang [25],[37] extended Gel'fand and Pinsker's model to the case where both the transmitter and receiver observe different non-perfect CSI sequences, which are correlated to the state sequence. This extension can be used to model a wider range of problems than the previous model. For example, it can model a high-definition television (HDTV) system, where the noisy analog version of the TV signal is the side information to the decoder. In this example, the side information is not really part of the channel, because the analog transmission serves as the side information and there is no state process in the channel.

The model which they used assumes a CSI signal, W_n, available to the transmitter, a correlated CSI signal, V_n, available to the receiver, and a memoryless channel with transition probability

$P_{Y|X,W,V}(y|x,w,v)$. We assume that the CSI signals (W_n, V_n) are generated by a memoryless source

$$P_{W^N,V^N}(w^N, v^N) = \prod_{n=1}^{N} P_{W,V}(w_n, v_n). \tag{3.24}$$

The output Y^N has the conditional distribution

$$P_{Y^N|X^N,W^N,V^N}(y^N|x^N, w^N, v^N) = \prod_{n=1}^{N} P_{Y|X,W,V}(y_n|x_n, w_n, v_n). \tag{3.25}$$

For this model, we define a block code of length N as a sequence of N encoding functions $f_n : \mathcal{M} \times \mathcal{W}^N \longrightarrow \mathcal{X}$, $n = 1, ..., N$, such that $x_n = f_n(m, w^N)$, where m ranges over the set of possible source messages $\mathcal{M}$. The decoding function is $g_n : \mathcal{Y}^N \times \mathcal{V}^N \longrightarrow \mathcal{M}$, such that the decoded message is $\hat{m} = g(y^N, v^N)$. The capacity in this case will be

$$C = \max_{P_{U,X|W}(u,x|w)} [I(U; V, Y) - I(U; W)]. \tag{3.26}$$

Cover and Chiang's result follows directly from Gel'fand and Pinsker's result. Therefore, Cover and Chiang's model is not really a generalization of Gel'fand and Pinsker's original model, just like Salehi's [101] model which is not really a generalization of Shannon's original model.

Another interesting aspect of the Gel'fand-Pinsker channel model presented in Section 2, is that similarly to a DMC with feedback [36], adding a feedback to the Gel'fand-Pinsker channel model or to the Shannon channel model does not increase the capacity [131]. As in the classical DMC, feedback can help in simplifying the coding scheme in the Gel'fand-Pinsker channel, so there is no need for a complicated binning scheme.

We can look back at the last few steps in Gel'fand and Pinsker's converse proof [56], and see that it is general enough as to include the feedback. We have defined $U_i = (m, Y^{i-1}, S_{i+1}^N)$ where $U_i \longrightarrow (X_i, S_i) \longrightarrow Y_i$ forms a Markov chain. $U_i \longrightarrow (X_i, S_i) \longrightarrow Y_i$ is a Markov chain even if we include a feedback in the channel. This is due to the fact that $(m, S^N, Y^{i-1}) \longrightarrow (X_i, S_i) \longrightarrow Y_i$ is a Markov chain and therefore, since U_i is a deterministic function of (m, S^N, Y^{i-1}), $U_i \longrightarrow (X_i, S_i) \longrightarrow Y_i$ is a Markov chain.

We have mentioned in Subsection 3.3 that Shannon's causal counterpart of the problem is a special case of the Gel'fand-Pinsker non-causal problem. If we add a feedback to Shannon's model, we arrive at the same conclusion, i.e., feedback does not increase capacity in the causal model. This is due to the fact that the independence between U_i and S_i is maintained in the presence of feedback (see Subsection 3.3).

3.5 Rate-Limited Side Information

While we have assumed thus far that the CSI is delivered to the transmitter and/or receiver, "free-of-charge", a more appropriate scenario would describe these deliveries as being carried out across capacity-limited channels, in particular, capacities that are not necessarily higher than the entropy of the state process. Heegard and El-Gamal [63] assumed that the CSIT and the CSIR are subjected to rate constraints R_e and R_d, respectively. Therefore, the CSI signals carry only partial information about the underlying state process. Their model is depicted in Fig. 3.3.

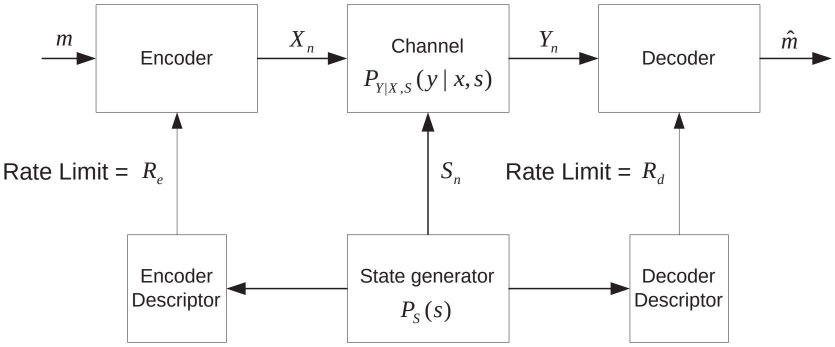

Fig. 3.3 A channel with two-sided non-causal state information with rate constraint model.

The corresponding definition of the coding system is as follows:

Definition 3. An $(N, R, R_e, R_d, \epsilon)$ code consists of the following four functions:
$J_e : \mathcal{S}^N \longrightarrow \{1, ..., 2^{NR_e}\}$

$J_d : \mathcal{S}^N \longrightarrow \{1, ..., 2^{NR_d}\}$
$f : \{1, ..., 2^{NR}\} \times \{1, ..., 2^{NR_e}\} \longrightarrow \mathcal{X}^N$
$g : \mathcal{Y}^N \times \{1, ..., 2^{NR_d}\} \longrightarrow \{1, ..., 2^{NR}\}$
$P_e \le \epsilon$, where P_e is defined in Section 2.

The mapping J_e generates a description of the state vector S^N to the encoder f, at rate R_e. Similarly, J_d generates a description of S^N to the decoder g, at rate R_d. The mappings J_e and J_d, are implemented by the encoder descriptor (ED) and decoder descriptor (DD) blocks, respectively.

Definition 4. For fixed R_e and R_d, a rate triplet (R, R_e, R_d) is achievable if for $\epsilon > 0$ there exists an $(N, R, R_e, R_d, \epsilon)$ code for all sufficiently large N.

The capacity is defined as the supremum over all R, such that (R, R_e, R_d) is achievable.

The complete characterization of the capacity region is still unknown. An inner bound to the capacity region was provided by Heegard and El-Gamal, and is presented in Theorem 1.

Theorem 1. [63] Fix $(\mathcal{S}, P_S(s), \mathcal{X}, P_{Y|X,S}(y|x,s), \mathcal{Y})$ and alphabets $\mathcal{U}, \mathcal{S}_0, \mathcal{S}_e$ and $\mathcal{S}_d$. All rate triples (R, R_e, R_d) in the convex hull of the set

$$\left\{\begin{array}{l} (R, R_e, R_d) | R_e > I(S_0, S_e; S) \\ R_d > I(S_0, S_d; S) - I(S_0, S_d; Y) \\ R_d > I(S_d; S|S_0) - I(S_d; Y|S_0) \\ R_e + R_d > I(S_0, S_e, S_d; S) - I(S_0, S_d; Y) + I(S_e; S_d|S_0) \\ R_e + R_d > I(S_e, S_d; S|S_0) - I(S_d; Y|S_0) + I(S_e; S_d|S_0) \\ R < I(U; Y, S_d|S_0) - I(U; S_e|S_0) \end{array}\right\}$$

for some probability mass function

$$\begin{aligned} &P_{S,S_0,S_e,S_d,U,X}(s, s_0, s_e, s_d, u, x) = \\ &\qquad P_S(s) P_{S_0,S_e,S_d|S}(s_0, s_e, s_d|s) P_{U,X|S_0,S_e}(u, x|s_0, s_e) \end{aligned} \tag{3.27}$$

are achievable.

The variables S_0, S_e, S_d denote common side information, encoder side information and decoder side information, respectively. U is an auxiliary random variables that plays the role of the coding RV of Gel'fand

and Pinsker. The first condition, is associated with the rate-distortion function for the channel from S to (S_0, S_e). The second condition, is associated with the Wyner-Ziv rate-distortion function [135] for the channel from S to (S_0, S_d) with the channel output Y serving as a decoder side information. The third condition, is associated with the Wyner-Ziv rate-distortion function for the channel from S to S_d with Y serving as a decoder side information, and S_0 as common side information. The fourth and fifth conditions, are combinations of the first condition with the second and third conditions, respectively. We also add to both of these conditions, the information between S_e and S_d, i.e., $I(S_e; S_d|S_0)$. This is due to the fact that if we want to look at the sum of the rates $R_e + R_d$ we have to take into account the influence of S_e on S_d, because now we do not have two separate problems of encoding S_e, and S_d. The last condition is similar to Gel'fand and Pinsker's capacity formula, conditioned on S_0, which is the common part of side information sent to the transmitter and receiver.

Although Theorem 1 gives only an inner bound on the achievable region, Heegard and El-Gamal found several special cases for which it is the exact achievable region.

(a) $R_e = R_d = 0$ (no description of defects), the capacity is

$$C = \max_{P_X(x)} I(X; Y). \tag{3.28}$$

(b) $R_e > H(S), R_d > H(S|Y)$ (complete description of defects at encoder and decoder), the capacity is equal to (3.22).

(c) $R_e > H(S), R_d = 0$ (complete description of defects at encoder and no description at decoder, i.e., the Gel'fand-Pinsker model), the capacity is equal to (3.14).

(d) $R_e > 0, R_d = H(S|Y)$ (complete description of defects at decoder).

In case (d), Heegard and El-Gamal suggested an expression which is too optimistic. This case was later corrected by Rosenzweig, Steinberg and Shamai [99], which gave a different expression for the capacity, as we shall see later.

The achievability of these capacity formulas follows from Theorem 1 by identifying the auxiliary random variables as follows:

(a) $S_0 = S_e = S_d = \emptyset, U = X$.

(b) $S_0 = S, S_e = S_d = \emptyset, U = X$.

(c) $S_0 = S_d = \emptyset, S_e = S$.
(d) $S_e = \emptyset, (S_0, S_d) = S, U = X$.
In [99], it was mentioned that substituting the variables $(S_0, S_d) = S$ in case (d), forces the auxiliary variable S_0 to be a deterministic function of S. Instead, Rosenzweig, Steinberg and Shamai suggested the assignment $S_d = S$. In this case, the capacity is given by

$$C = \max_{P_{X|S_0}(x|s_0)} \max_{P_{S_0|S}(s_0|s)} I(X;Y|S,S_0), \tag{3.29}$$

where the second maximum is taken over the distribution $P_{S_0|S}(s_0|s)$ satisfying

$$R_e \leq I(S_0;S). \tag{3.30}$$

Partial CSIT has received considerable attention in the context of multi-input-multi-output (MIMO) systems used over the block fading channel. The presence of CSIT has been shown to yield significant performance gains in various aspects [10],[67],[100]. Many studies consider a MIMO system where the transmitter is given only a quantized finite rate CSI via an error-free feedback channel. Rosenzweig, Steinberg and Shamai [99], studied these types of channels. They extended Heegard and El-Gamal's model for a channel with partial CSIT and perfect CSIR (case (d)), to the case where the CSIT comprises two parts, both subjected to a rate constraint. This model was later further extended by Cemal and Steinberg to the case where the transmitter receives multiple partial CSI signals (see [17] and [18]). Cemal and Steinberg [19] extended the model of a single-user channel with partial CSIT and perfect CSIR to multiple access channel (MAC), where the encoders have access to partial rate-limited CSIT and the decoder have access to perfect CSIR. Multiple access channels will be discussed in Section 6.6.1.

Cemal and Steinberg [18] have also extended the model for a channel with partial CSIT and perfect CSIR, to the case where the partial CSI is conveyed through a noisy channel (genie channel) to the encoder. They considered an extended communication model where an information source generates i.i.d random variables $W_n \in \mathcal{W}$, drawn under distribution $P_W(\cdot)$, independent of the state source S_n, is to be transmitted through the channel, where the transmission is subjected to a distortion constraint between the source signal W and its reconstruction $\hat{W}$. This communication model is depicted in Fig. 3.4. The Gel'fand-Pinsker channel $P_{Y|X,S}(\cdot|\cdot,\cdot)$ operates at a rate of ρ_c^{-1} channel uses per information symbol and the genie channel $P_{S_e|K}(\cdot|\cdot)$ operates at a rate of ρ_s channel uses per state source (or at ρ_s/ρ_c channel uses per information symbol). The joint source-channel state encoder generates a block code of length $N_s = \rho_s N$, $k^{N_s} \in \mathcal{K}^{N_s}$, from a state source block of length N, $s^N \in \mathcal{S}^N$. The block K^{N_s} is subjected to a transmission cost constraint such that $E[\phi_s(K^{N_s})] \leq N_s\Gamma_s$ where $\phi_s(K^{N_s}) = \sum_{n=1}^{N_s} \phi_s(k_n)$ for some cost function $\phi_s : \mathcal{K} \longrightarrow \mathbb{R}_+$. The joint source-channel encoder, receives the partial CSIT $S_e^{N_s}$ which is the output of the genie channel, generates the input to the Gel'fand-Pinsker channel $x^N \in \mathcal{X}^N$ based on $S_e^{N_s}$ and an information source block of length $N_c = \rho_c N$, $W^{N_c} \in \mathcal{W}^{N_c}$. The input block x^N is subjected to a transmission cost constraint (2.3). The decoder has access to perfect CSIR s^N and produces an estimate of the source $\hat{w}^{N_c} \in \hat{\mathcal{W}}^{N_c}$ from the channel output y^N and the CSIR s^N. The quality of the estimation is measured by the expectation of an additive distortion measure: $d\left(w^{N_c}, \hat{w}^{N_c}\right) = \sum_{n=1}^{N_c} d\left(w_n, \hat{w}_n\right)$ where $d : \mathcal{W} \times \hat{\mathcal{W}} \longrightarrow \mathbb{R}_+$ is a single-letter distortion measure.

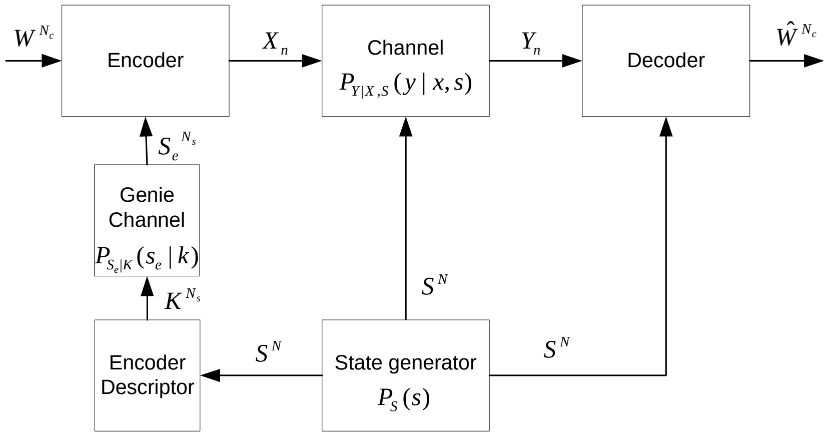

Fig. 3.4 Joint state-channel source-channel coding with partial state information at the transmitter.

Definition 5. An $(N, N_c, N_s, D, \Gamma, \Gamma_s)$ joint state-source-channel (SSC) code for the information source $P_W(\cdot)$, state source $P_S(\cdot)$, Gel'fand-Pinsker channel $P_{Y|X,S}(\cdot|\cdot,\cdot)$ and genie channel $P_{S_e|K}(\cdot|\cdot)$, consists of three mappings:

$$\begin{aligned} f_s &: S^N \longrightarrow K^{N_s} \\ f &: W^{N_c} \times S_e^{N_s} \longrightarrow X^N \\ g &: Y^N \times S^N \longrightarrow \hat{W}^{N_c} \end{aligned} \tag{3.31}$$

such that $E[d\left(W^{N_c}, g(Y^N, S^N)\right)] \leq N_c D$, $E[\phi(f(W^{N_c}, S_e^{N_s}))] \leq N\Gamma$ and $E[\phi_s(f_s(S^N))] \leq N_s \Gamma_s$.

The distortion cost triple (D, Γ, Γ_s) is achievable with bandwidth expansion factors ρ_c and ρ_s if for any $\epsilon > 0$ and sufficiently large N, there exists an $(N, \rho_c N, \rho_s N, D + \epsilon, \Gamma, \Gamma_s)$ joint SSC code. Given bandwidth expansion factors ρ_c and ρ_s, the distortion-cost region is the closure of the set of achievable distortion-cost triples (D, Γ, Γ_s) and is denoted by $\mathcal{D}$.

Cemal and Steinberg [18] have characterized $\mathcal{D}$.

Theorem 2. [18] For any information source $P_W(\cdot)$, state source $P_S(\cdot)$, Gel'fand-Pinsker channel $P_{Y|X,S}(\cdot|\cdot,\cdot)$ and genie channel $P_{S_e|K}(\cdot|\cdot)$ with bandwidth expansion factors ρ_c and ρ_s, $\mathcal{D}$ is the set of all distortion-cost triples (D, Γ, Γ_s) for which there exist a random variable S_0 taking values in $\mathcal{S}_0$ such that the following conditions are satisfied simultaneously: (i) The Markov relations $S \longrightarrow S_0 \longrightarrow X$, $S_0 \longrightarrow (X, S) \longrightarrow Y$ hold. (ii) The auxiliary alphabet $\mathcal{S}_0$ satisfies $|\mathcal{S}_0| \leq |S| + 1$. (iii) The distortion-cost triple (D, Γ, Γ_s) satisfies

$$\begin{aligned} \rho_c R_W(D) &\leq I(X; Y|S, S_0) \\ I(S, S_0) &\leq \rho_s C_g(\Gamma_s) \\ E[\phi(X)] &\leq \Gamma \end{aligned} \tag{3.32}$$

where $R_W(D) \triangleq \min_{P_{\hat{W}|W}(\hat{w}|w): E[d(w,\hat{w})] \leq D} I(W; \hat{W})$ is the rate-distortion function and $C_g(\Gamma_s) \triangleq \max_{P_K(k): E[\phi_s(k)] \leq \Gamma_s} I(K; S_e)$ is the capacity-cost function.

If we compare the achievable rate region given in (3.29), (3.30) to the distortion-cost region characterized in Theorem 2, we see that the partial description of R_e in (3.30) is replaced by $\rho_s C_g(\Gamma_s)$ and the Gel'fand-Pinsker channel capacity C in (3.29) is replaced by $\rho_c R_W(D)$. This connection implies a separation principle in our extended scenario. A separation principle holds for state coding independent of the genie channel (but depending on the Gel'fand-Pinsker channel), and a separation principle holds for the main source coding, independent of the Gel'fand-Pinsker channel statistics.

In [112], [111] a dual problem was investigated, where the transmitter have access to full non-causal CSIT and the receiver have access to rate-limited CSIR. The capacity of this model is given by

$$C = \max\left[I(U;Y|S_d) - I(U;S|S_d)\right] \tag{3.33}$$

where the maximization is over all $P_{X,S_d,U|S}(\cdot,\cdot,\cdot|\cdot)$ satisfing

$$R_d \geq I(S;S_d) - I(Y;S_d). \tag{3.34}$$

Coding of side information intended for the channel decoder is a Wyner-Ziv like problem, since the channel output depends statistically on the state, thus serving as side information in the decoding of the encoded state. Steinberg [113] suggested as an application for this model the problem of reversible information embedding with compressed host at the decoder. Information embedding, and reversible information embedding is discussed in Section 5.2.

4

Specific Channel Models

In this section, we describe two specific models of channels with side information. We use the results presented in Section 3 to find the capacities of these channels and we present coding techniques. The specific channels are the "dirty paper" channel, the AWGN channel with fading and the modulo additive noise channel.

4.1 The Dirty Paper Channel

In this subsection, we consider the problem of a power constrained additive noise channel, where part of the noise is known at the transmitter as side information. This part of the noise may be the result of an additive interference. This model is suitable to describe, for example, a scenario in which a user, which is located close to the main transmitter, is interfering, but cooperating by revealing his messages in advance. It is also suitable to describe a scenario in which one transmitting antenna, is interfering another antenna, serving the same user in a MIMO scenario.

The channel output, in this model, is given by

$$Y^N = X^N + S^N + Z^N, \tag{4.1}$$

where S^N is the interference sequence which is known to the encoder, and Z^N is the statistically independent unknown additive noise sequence. The encoder satisfies the power constraint (1.2).

This problem was investigated by Costa [34] and is known as the "dirty-paper" problem, and the channel model is known as Costa's channel. In his work, Costa assumed that the interference and noise are sequences of i.i.d. random variables distributed according to $S \sim \mathcal{N}(0, Q)$ and $Z \sim \mathcal{N}(0, B)$, respectively. Costa showed that the capacity is given by

$$C = \frac{1}{2} \log \left(1 + \frac{\Gamma}{B}\right), \tag{4.2}$$

which is the same as if S^N was absent altogether.

Let X be a $\mathcal{N}(0, \Gamma)$ random variable independent of S and let $U = X + \alpha S$. Costa directly computed $I(U;Y) - I(U;S)$ for this joint distribution and then optimized the result over α. He found that the optimal value of α is $\alpha^* = \frac{\Gamma}{\Gamma + B}$. We observe that $\alpha^*(X + Z)$ is the MMSE estimator of X given $X + Z$, and therefore, $X - \alpha^*(X + Z)$ is independent of $X + Z$. Furthermore, $X - \alpha^*(X + Z)$ is independent of $Y = X + S + Z$ since they are jointly Gaussian and uncorrelated, or more generally, since $X - \alpha^*(X + Z) \longrightarrow X + Z \longrightarrow X + S + Z$ form a Markov chain. This more general independence quality allows us to drop Costa's assumption that S^N is Gaussian. Therefore, S^N can have any (power limited) ergodic distribution [31], [32]. Costa's result was also extended to any known deterministic sequence S^N with a common source of randomness available to both the transmitter and receiver [48].

Costa's result still holds, with a more general and sufficient condition on the noise Z^N: if X^N has the capacity-achieving distribution for the additive noise channel $Y^N = X^N + Z^N$ and there exist a linear function $\alpha(\cdot)$ such that $X^N - \alpha\left(X^N + Z^N\right)$ is independent of $X^N + Z^N$. This condition is met if Z^N is a colored Gaussian process since the capacity achieving distribution is also Gaussian. In [141] this result was further extended to the case where the noise and interference are not necessarily stationary or ergodic (see [141] for more details).

The dirty-paper approach can also be extended to the vector case [140]. This extension is given in the following Lemma.

Lemma 1. Given a fixed power constraint, a Gaussian vector channel with side information $y^N = x^N + s^N + z^N$, where z^N and s^N are independent Gaussian random vectors, and s^N is known non-causally at the transmitter but not at the receiver, has the same capacity as if s^N did not exist, i.e.,

$$\begin{aligned} C &= \max_{P_{U^N,X^N|S^N}(u^N,x^N|s^N)} \left\{ I(U^N;Y^N) - I(U^N;S^N) \right\} \\ &= \max_{P_{X^N|S^N}(x^N|s^N)} I(X^N;Y^N|S^N). \end{aligned} \tag{4.3}$$

Further, the capacity-achieving x^N is statistically independent of s^N.

Costa presented a transmission strategy based on the binning technique. He did not give a practical coding scheme for this channel. Several practical coding schemes for this problem will be presented in Section 7.

In the binary modulo-2 additive noise channel case, Barron, Chen and Wornell [6], showed that if the known interference S is a binary symmetric source, the known noise Z is an independent Bernoulli-p source, and the channel input satisfies an input Hamming constraint $\frac{1}{N} E\omega_H(X^N) \leq \gamma$, where $\omega_H(\cdot)$ denotes the Hamming weight, then the capacity with side information at the transmitter is given by

$$C = u.c.e \left\{ H(\gamma) - H(p), (0,0) \right\}, \quad 0 \leq \gamma \leq 0.5 \tag{4.4}$$

where $u.c.e\{\cdot\}$ denotes upper convex envelope as a function of γ. The proof of (4.4) is dual to the proof given by Wyner and Ziv for the binary symmetric source coding problem with side information.

A closed-form formula for the capacity of the dirty-tape channel is still unavailable. In [48] Erez, Shamai and Zamir used Shannon's general capacity formula (3.7) for the causal case to find the capacity of this channel for the worst case interference in the sense of the statistics of S, which is the asymptotic case of strong interference. In their calculations, they have also used a common random variable, available to both the transmitter and receiver (common randomness). This common variable does not increase the capacity [77].

The worst interference capacity of Shannon's causal channel without randomness is given by

$$C = \inf_{P_S(\cdot)} \sup_{P_T(\cdot):E\{(T(S)^2)\} \leq \Gamma} I(T;Y). \tag{4.5}$$

Let U be a RV uniformly distributed over $[-L/2, L/2)$, where L is a positive number. Define

$$h_{min} = \inf_{t \in \mathcal{T}} h(t(U) + U + Z), \tag{4.6}$$

where $h(\cdot)$ denotes differential entropy and Z is the channel noise, and where $t(\cdot)$ is a strategy belonging to the set of strategies $\mathcal{T}(\Gamma)$, defined over the interval $[-L/2, L/2)$, and satisfying the power constraint (1.2), i.e.,

$$\mathcal{T}(\Gamma) = \left\{ t : E[t(U)]^2 \leq \Gamma \right\}. \tag{4.7}$$

Define

$$\widetilde{C}_L(\Gamma) = \log L - h_{min}, \tag{4.8}$$

and

$$\widetilde{C}(\Gamma) = \lim_{L \to \infty} \widetilde{C}_L(\Gamma). \tag{4.9}$$

The worst interference capacity $C(\Gamma)$ of the dirty-tape problem is given by the upper convex envelope of $\widetilde{C}(\Gamma)$.

Erez, Shamai and Zamir used a coding scheme which is similar to the one that was proposed by Costa for this problem. Costa's encoding scheme is, in general, suboptimal for general SNR's, but is asymptotically optimal at high SNR's. The asymptotic (high SNR) rate loss with respect to the no interference case, is equal to the shaping gain, $\frac{1}{2} \log \frac{2\pi e}{12} \approx 0.254$ bits per channel use [48].

We will present coding schemes for this problem in Subsection 7.2.

4.2 The AWGN Channel With Fading

The additive white Gaussian noise (AWGN) channel with fading can be modeled as a state dependent channel with CSIT and/or CSIR using the models presented in Section 3. A large number of works have been devoted to assess the capacity of this channel. These works differ in their assumptions on the availability of CSI at both the transmitter and receiver.

In this subsection, we consider a single-user channel model with flat-fading[1]. We assume that the channel output is matched-filtered to the pulse shape and subsequently sampled. These samples are modeled as

$$Y_n = S_n X_n + Z_n, \tag{4.10}$$

where the channel input and output, at time n, are represented by $X_n \in \mathbb{C}, Y_n \in \mathbb{C}$ respectively. The complex, circularly symmetric i.i.d. Gaussian noise samples are represented by Z_n, where $E(|Z_n|^2) = \sigma^2$. S_n represents the complex circularly symmetric fading samples with a power $(R_n = |S_n|^2)$ and power distribution designated by $P_R(\cdot)$. The phase of the fading samples $\arg(S_n)$ ($\arg(S_n)$ stands for the argument of S_n) is distributed uniformly in $[-\pi, \pi)$ and is assumed independent of R. We further assume that $E(R) = E(|S|^2) = 1$. The channel input, in this model, is subjected to an average-power constraint

$$E(|X_n|^2) \leq \Gamma. \tag{4.11}$$

In the following subsections, we present different models for the AWGN channel with fading. In each model, we will use different assumptions on the availability of side information at the transmitter and/or the receiver.

4.2.1 Perfect CSIR and no CSIT

This case has been treated in [60],[11]. It applies, for example, to the case of a flat-fading channel where the receiver is informed by a third party of the fading realizations s^N and use them as it's CSIR. We assume that $\{R_n\}$ is a stationary ergodic process. In this case, we can easily derive the capacity formula by

$$C = E\left[\log(1 + \frac{r\Gamma}{\sigma^2})\right] = \int_0^\infty P_R(r) \log\left(1 + \frac{r\Gamma}{\sigma^2}\right) dr, \tag{4.12}$$

where $P_R(\cdot)$ is the distribution function of R. This channel is equivalent to a memoryless channel and the CSIR is interpreted as an additional

[1] A passband pulse amplitude modulated signal is said to experience flat-fading if it is transmitted over a fading channel of a delay spread that is negligible compared to the symbol duration.

channel output [13]. Therefore, ordinary channel coding will be sufficient to achieve capacity in this case (see Section 3.1).

4.2.2 Perfect CSIT and CSIR

This model applies, for example, to the case of time-division duplex (TDD) based systems, where reciprocity facilitates channel measuring. The receiver, as in Section 4.2.1, is informed by a third party of the fading realizations, s^N, and use them as it's CSIR. Here, we assume that the channel state information is known to both the transmitter and receiver in a causal manner. The capacity formula, in this case, is a special case of (3.11) and is given by

$$C = E\left[\sup \ \log(1 + \frac{\Psi(r)r}{\sigma^2})\right], \tag{4.13}$$

where the supremum is over all nonnegative power assignments $\Psi(r)$ satisfying

$$E\left[\Psi(r)\right] \le \Gamma. \tag{4.14}$$

The optimal power assignment $\Psi_{opt}(r)$, given in [60], satisfies

$$\frac{\Psi_{opt}(r)}{\Gamma} = \begin{cases} \frac{1}{r_0} - \frac{1}{r}, & r \ge r_0 \\ 0, & r < r_0 \end{cases} \tag{4.15}$$

where the constant r_0 is determined by the average power constraint and the specific distribution of the fading power $P_R(r)$. This optimal power control policy can be interpreted as a time-water-pouring solution [60], that is, above a threshold r_0 the stronger the channel gain, the larger is the instantaneous transmitted power.

The capacity, in this case, can be achieved by a variable-rate, variable-power communication technique as was suggested in [60]. However, in [13], Caire and Shamai showed that it is possible to achieve capacity using an ordinary channel coding scheme and using, at the input of the channel, an amplifier, whose gain is $\sqrt{\Psi_{opt}(r)/\Gamma}$, controlled by the observed fading power r. This amplifier is interpreted as part of the channel, the instantaneous power gains are revealed to the receiver, and we can use (4.12) to calculate the capacity in this case by replacing r with the effective power gains $\Psi_{opt}(r)/\Gamma$.

4.2.3 Perfect CSIT and no CSIR

This model is less interesting than the previous ones, because it rarely reflects practical situations. This model may be applied to the example given in Section 4.2.2 without CSIR. and the channel output is a complex scalar Y with probability distribution given by

$$P_{Y|T,R}(y|t(r),r) = \int_0^\infty P_R(r)\frac{1}{2\pi\sigma^2}\left\{\exp -\frac{1}{2\sigma^2}|y-\sqrt{r}t(r)|^2\right\}dr, \tag{4.16}$$

with the input $t(r), r \geq 0$ subjected to the average power constraint $E\left[T^2(R)\right] \leq \Gamma$.

We note here that we have assumed that $s = \sqrt{r}$. This assumption is justified by the reason that if the transmitter has full access to the fading coefficients, it can fully neutralize their phase by rotation at no additional power cost.

A general formula of the capacity, has not been found. A simple upper bound is the case were the CSI is also available to the receiver. Lower bounds were derived by using suboptimal strategies. One of them is the truncated channel inversion [60]. This strategy uses a standard long codebook and an amplitude amplifier at the transmitter with the power gain function

$$\frac{\Psi_{opt}(r)}{\Gamma} = \begin{cases} \frac{\alpha}{r}, & r \geq r_0 \\ 0, & r < r_0 \end{cases}, \tag{4.17}$$

where

$$\alpha^{-1} = \int_{r_0}^\infty r^{-1}P_R(r)dr. \tag{4.18}$$

The receiver, in this case, receives an unfaded Gaussian channel $Y_n = X_n + Z_n$ with probability

$$P_G = \int_{r_0}^\infty P_R(r)dr, \tag{4.19}$$

and a pure noise channel $Y_n = Z_n$ with the complementary probability $1 - P_G$.

A lower bound is obtained by setting: $P_G = 1$, i.e., the channel is unfaded and there is no power constraint. In this case, the lower bound is given by

$$C \geq \log\left(1 + \frac{SNR}{E[1/r]}\right). \tag{4.20}$$

As a simple upper bound, we can take (4.13), which is the case where the CSI is also available to the receiver.

In [61], tighter upper bounds were given both for the causal and the non-causal cases. These upper bounds were derived directly from Gel'fand and Pinsker's capacity formula (3.14) for the non-causal case, and from Shannon's capacity formula (3.3) for the causal case (see [61]). Therefore, they are superior to the trivial upper bound which assumes that the side information is also available to the receiver.

4.3 Modulo Additive Noise Channels

The modulo additive noise channel was studied in [50],[51],[52]. This channel belongs to the class of channels with states and side information at the encoder which were presented in Section 3.

Very few explicit solutions exist for the capacity of channels with side information at the encoder. This is due to the computational complexity of the solution. The modulo additive noise channel is one of the few cases for which the capacity was found. This model may be used, for example, in the problem of writing to a computer memory with defective cells.

The modulo additive noise channel problem is divided into two categories, according to whether the side information is available to the encoder in a causal or non-causal manner. We will present both.

A simple coding scheme that achieves the capacity of this channel was presented in [50]. This coding structure achieves also the random-coding error exponent, and therefore is optimal for some range of rates below capacity [51],[52].

Let $\mathcal{X} = \mathcal{Y} = \mathcal{Z} = \{0, ..., |\mathcal{X}| - 1\}$. A symmetric, or modulo additive noise channel can be described by

$$Y = X + Z, \tag{4.21}$$

where Z is conditionally independent of X given the state S, and addition or substraction is performed modulo $|\mathcal{X}|$. The side information

S may be thought of as a version of the channel noise Z. No power constraint is imposed on the transmitter.

We begin with the causal version of the problem. By the additivity of the channel, we have $P_{Y|X,S}(y|x,s) = P_{Z|S}(y-x|s)$. Therefore, the transition probability of the equivalent channel to the modulo additive noise channel is given by

$$P_{Y|T}(y|t) = \sum_s P_S(s) P_{Z|S}(y - t(s)|s) = P_r(Z + t(s) = y). \quad (4.22)$$

The capacity of the memoryless modulo additive noise channel with causal side information is give by [50]

$$C = \log |\mathcal{X}| - H_{min} \quad (4.23)$$

where

$$H_{min} \triangleq \min_{t \in \mathcal{T}} H(Z - t(S)), \quad (4.24)$$

where $H(\cdot)$ is the entropy. It can be easily seen that by (4.22) we get $H_{min} = \min_{t\in\mathcal{T}} H(Y|T=t) = \min_{t\in\mathcal{T}} H(Z+t(S))$, this is because the minimization of the entropy of $Z+t(S)$ and of $Z-t(S)$ is the same.

We start with the converse part of the proof, i.e., $C \leq \log|\mathcal{X}| - H_{min}$. Since

$$H(Y) \leq \log |\mathcal{X}| \quad \textit{and} \quad H(Y|T) \geq \min_{t\in\mathcal{T}} H(Y|T=t) = H_{min}, \quad (4.25)$$

we have $I(T;Y) = H(Y) - H(Y|T) \leq \log|\mathcal{X}| - H_{min}$ for any distribution on $\mathcal{T}$, and the converse follows from (3.3).

We next show the direct part, i.e., $C \geq \log|\mathcal{X}| - H_{min}$. Let t^* denote a strategy for which $H(Y|T=t^*) = H_{min}$. Define the following class of strategies

$$\mathcal{T}^* = \{t_j\} \textit{ where } t_j(s) = t^*(s) + j, \quad j = 1 \cdots |\mathcal{X}|. \quad (4.26)$$

From (4.22), we see that $P_{Y|T}(y|t_j) = P_r(Z + t_j(S) = y) = Pr(Z + t^*(S) = y - j)$ i.e., $P_{Y|T}(y|t_j)$ is the transition probability $P_{Y|T}(y|t^*)$ shifted modulo $|\mathcal{X}|$ by j. Therefore, $H(Y|t_j) = H(Y|t^*) = H_{min}$ for all j. Furthermore, choosing T to be uniformly distributed within $\mathcal{T}^*$ induces uniform distribution on $\mathcal{Y}$. Thus, for such T we have equality in both inequalities (4.25), and the direct part follows.

We have given the full proof for the capacity of this channel because it helps to understand the optimal transmission scheme which is called "the instantaneous-predictor encoder" that is depicted in Fig. 4.1. By restricting the input alphabet to the set $\mathcal{T}^*$, the resulting channel from $\mathcal{T}^*$ to $\mathcal{Y}$ can be viewed as an additive noise channel, whose alphabet is $\mathcal{X}$, and whose noise is distributed as $\widetilde{Z} \triangleq Z+t^*(S) \triangleq Z-t_{min}(S)$, where $t_{min}(\cdot) \triangleq arg\min_{t:\mathcal{S}\to\mathcal{X}} H(Z-t(S))$. The transmission scheme, depicted in Fig. 4.1, has a simple modular structure, consisting of an ordinary (i.e., state independent) Shannon code for a symmetric channel with rate bellow (4.23), followed by a shift by $t_{min}(s)$. This structure leads to the following interpretation. Since the receiver does not know S, it "sees" a channel with effective noise $\widetilde{Z} = Z - t_{min}(S)$. Shifting Z by $t_{min}(S)$ thus makes the effective noise the least harmful for the receiver. The function $t_{min}(\cdot)$ minimizes the entropy of $Z-t_{min}(S)$. We can view this function as the prediction of Z from S with minimum error entropy criterion.

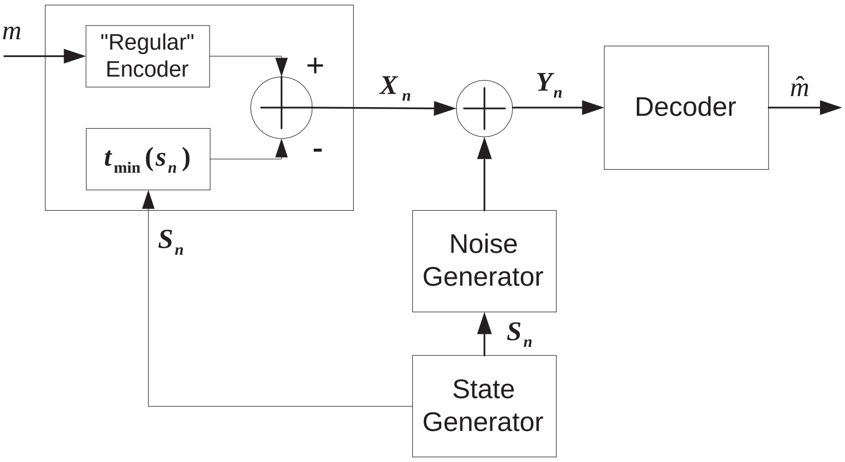

Fig. 4.1 Instantaneous prediction encoding scheme.

This result was also expanded to the case of non-causal side information, and to the case where the state process has memory. We use the following definition of the state weight independent prediction (SWIP).

Definition 6. [50] Fix the set of conditional noise distributions $\{P_{Z|S}(z|s), z \in \mathcal{X}, s \in \mathcal{S}\}$. If the same function $t = t_{min}(\cdot)$ minimizes the entropy of $Z - t(S)$ for every state set of weights $\{P_S(s), s \in \mathcal{S}\}$, then we say that the noise satisfies the state weight independent prediction (SWIP) property, or in short, "the noise is SWIP".

We also define

$$H^N_{min} \triangleq \frac{1}{N} \min_{t^N : \mathcal{S}^N \longrightarrow \mathcal{X}^N} H(Z^N - t^N(S^N)). \tag{4.27}$$

The capacity for the SWIP noise with memory both for the causal and non-causal cases is given in Theorem 3.

Theorem 3. [50] For SWIP noise and stationary state process, the instantaneous shift function

$t^N(s^N) = (t_{min}(s_1), \cdots, t_{min}(s_N))$ where t_{min} achieves H^1_{min} in (4.27), is optimal for both causal and non-causal side information at the encoder. Thus

$$C = \log |\mathcal{X}| - \lim_{N\to\infty} \frac{1}{N} H(Z_1 - t_{min}(S_1), \cdots, Z_N - t_{min}(S_N)) \tag{4.28}$$

i.e., the optimum (causal or non-causal) encoder reduces to the instantaneous prediction encoder.

The limit in (4.28) exists since $Z_i - t_{min}(S_i)$, $i = 1, 2...$, is a stationary process.

The capacity for the memoryless SWIP additive noise channel is given by

$$C = \log |\mathcal{X}| - H(Z - t_{min}(S)). \tag{4.29}$$

Therefore, the capacity in the causal case is equal to the capacity of the non-causal case, if the noise satisfies the SWIP property.

5

Applications

Next, we present several applications that will be described by the models presented in Section 3, and Section 4.

We distinguish between applications and specific channel models in which the side information is inherently a part of the channel (e.g., the AWGN channel with fading), and cases where the side information is "artificial" or "man-made", e.g., a channel with two cooperating transmitters where the side information to the first transmitter is the message transmitted by the second transmitter.

5.1 The Gaussian Vector Broadcast Channel

A broadcast channel model describes a communication scenario in which a single transmitter sends independent information to multiple non-cooperating receivers. A Gaussian vector broadcast channel or MIMO broadcast channel model describes a communication scenario in which a single transmitter with multiple antennas sends independent information to multiple non-cooperating receivers with multiple antennas through a Gaussian channel.

This model is used to describe the downlink of a wireless system, where the base station is equipped with an antenna array, for example, a CDMA downlink [128]. It is also being used to describe a DSL line [58] with coordinated transmission and uncoordinated receivers.

We use the following notation in this subsection. Upper case letters are used to denote scalar random variables as before, or matrices, e.g., H, where the context should make the distinction clear. Bold face letters are used to denote vectors, e.g., **x,y**, or vector random variables, e.g., **X,Y**. For matrices, the superscript † denotes the conjugate transpose (Hermit) operation, the superscript t denotes the transpose operation and $|\cdot|$ denotes the determinant operation.

The broadcast channel was first introduced by Cover [35], who also proposed an achievable coding strategy based on superposition. Superposition coding has been shown to be optimal for a class of degraded broadcast channels[1].

Consider a standard scalar two-user Gaussian broadcast channel, defined by

$$\begin{aligned} Y_1 &= h_1 X + Z_1 \\ Y_2 &= h_2 X + Z_2 \end{aligned} \tag{5.1}$$

where $Z_i \in \mathbb{C}$ is the i.i.d. noise which is a complex circularly symmetric Gaussian RV with variance $E[||Z_i||^2] = 1$, i.e., $Z_i \sim \mathcal{N}_{\mathbb{C}}(0,1)$, and where h_1, h_2, $|h_1|^2 \geq |h_2|^2$, are the channel gains for the two users known to both the transmitter and receivers, and where $X \in \mathbb{C}$ and $Y_i \in \mathbb{C}$ are the input and outputs of the channel.

This channel can be regarded as a degraded broadcast channel for which the capacity region is well established [36].

The capacity region of this channel can be attained by two different coding techniques. The first is the superposition technique, and the second one is the dirty-paper technique [140]. We will now describe how the dirty-paper technique is used for this channel.

The transmitter sends $X = X_1 + X_2$, where X_1 and X_2 are scalar independent Gaussian signals with powers $\alpha\Gamma$ and $(1-\alpha)\Gamma$ for $\alpha \in [0,1]$,

[1] A two-user broadcast channel is called physically degraded if $P_{\boldsymbol{Y}_1,\boldsymbol{Y}_2|\boldsymbol{X}}(\boldsymbol{y}_1,\boldsymbol{y}_2|\boldsymbol{x}) = P_{\boldsymbol{Y}_1|\boldsymbol{X}}(\boldsymbol{y}_1|\boldsymbol{x})P_{\boldsymbol{Y}_2|\boldsymbol{Y}_1}(\boldsymbol{y}_2|\boldsymbol{y}_1)$, and stochastically degraded if its conditional marginal distributions are the same as that of a physically degraded broadcast channel.

respectively. The message intended for Y_1 is transmitted through X_1, and the message intended for Y_2 is transmitted through X_2. Once a codeword $x^N(2)$ for user 2 is generated, the transmitter non-causally knows the interference signal $h_1 x^N(2)$ that this codeword will cause to user 1. Therefore, by using Costa's results an achievable rate for user 1 will be $\log(1+\alpha|h_1|^2\Gamma)$. User 2 (the weaker user) decodes $x^N(2)$ by treating $x^N(1)$ as an additional background noise. Hence, an achievable region for this channel is given by

$$\bigcup_{\alpha\in[0,1]} \left\{ (R_1,R_2) : \begin{array}{l} R_1 \leq \log(1+\alpha|h_1|^2\Gamma), \\ R_2 \leq \log\left(1+\frac{(1-\alpha)|h_2|^2\Gamma}{1+\alpha|h_2|^2\Gamma}\right) \end{array} \right\}. \tag{5.2}$$

This achievable region is also the capacity region for this channel. The converse which proves this, was given by Bergmans [9] using the entropy power inequality (EPI). Note that in applying the dirty-paper coding (DPC) technique in the broadcast channel, we have not used the degraded structure of the channel. This fact hints that DPC technique can be useful in non degraded settings, where successive cancelation at the receivers is not necessarily possible, due to the lack of order between the receivers.

Next, we consider the Gaussian vector broadcast (MIMO broadcast) channel depicted in Fig. 5.1, with K users (receivers), user i equipped with r_i antennas ($i=1,2,...,K$), and a transmitter with t antennas, defined by

$$\begin{bmatrix} \boldsymbol{Y}_1 \\ \cdot \\ \cdot \\ \cdot \\ \boldsymbol{Y}_K \end{bmatrix} = \begin{bmatrix} H_1 \\ \cdot \\ \cdot \\ \cdot \\ H_K \end{bmatrix} \boldsymbol{X} + \begin{bmatrix} \boldsymbol{Z}_1 \\ \cdot \\ \cdot \\ \cdot \\ \boldsymbol{Z}_K \end{bmatrix} \tag{5.3}$$

where $\boldsymbol{X} \in \mathbb{C}^t$, $\boldsymbol{Y}_i \in \mathbb{C}^{r_i}$, and $\boldsymbol{Z}_i \in \mathbb{C}^{r_i}$ where $\boldsymbol{Z}_i \sim \mathcal{N}(0,\Sigma_{z_i z_i})$ are the input, the outputs and the noise vectors (at any given channel use), and $H_i \in \mathbb{C}^{r_i \times t}$ are the channel transfer matrices from the transmitter to the i-th receiver. We assume that the channel matrices H_i are constant and are perfectly known to the transmitter and to all the receivers. The transmitter, in this setting, sends independent information $m_i \in \{1,...,2^{NR_i}\}$, $i=1,...,K$, to each receiver. Let $\boldsymbol{X} = \boldsymbol{X}_1 + ... + \boldsymbol{X}_K$,

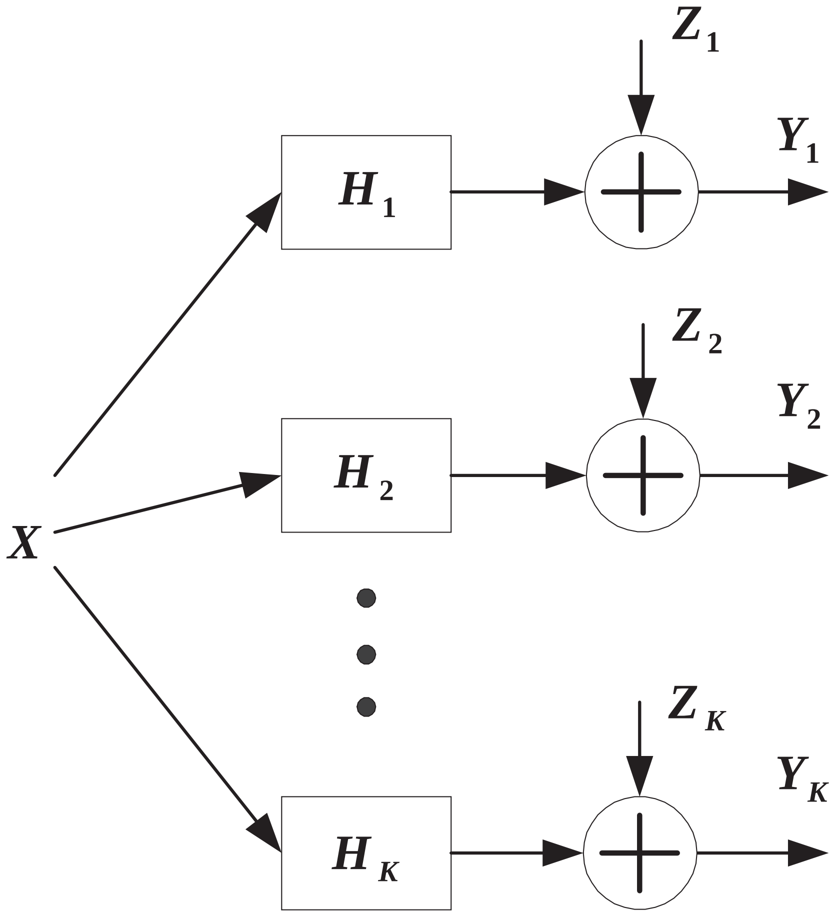

Fig. 5.1 The K-user Gaussian vector broadcast channel.

where $\boldsymbol{X}_i$, $i = 1, ..., K$, are independent Gaussian vectors whose covariance matrices are $\Sigma_{x_i x_i}$. The message intended for $\boldsymbol{Y}_i$ is transmitted through $\boldsymbol{X}_i$, for $i = 1, 2..., K$. The input signal satisfies an input constraint $tr(E[\boldsymbol{X}\boldsymbol{X}^{\dagger}]) \leq \Gamma$.

When a broadcast channel has a vector input and vector outputs, it is no longer necessarily degraded. Superposition coding and successive decoding does not apply to a general non-degraded broadcast channel. However, the dirty-paper approach is still applicable to a general non-degraded broadcast channel. Caire and Shamai [14], [16] were the first to consider using the dirty-paper approach to find the achievable rate region for a two-user non-degraded broadcast channel, with t antennas for the transmitter, and one antenna per user. Caire and Shamai have also found the sum capacity for this two-user case.

The dirty-paper approach can be extended to the vector case to be used for the MIMO broadcast channel [140]. This extension was given in Lemma 1, and it may be applied at the transmitter by choosing codewords for different receivers in a similar way as for the two-user scalar broadcast channel case. The transmitter picks a codeword for receiver 1. The transmitter then chooses a codeword for receiver 2 with full (non-causal) knowledge of the codeword intended for receiver 1. Therefore, it can use a dirty-paper coding technique in order to subtract the interference for receiver 2, which is caused by the codeword intended to receiver 1. Similarly, the codeword for receiver 3 is chosen in order to subtract the interference to receiver 3, which is caused by signals intended for receivers 1 and 2. This process continues for all K receivers.

Using Lemma 1, we can get the achievable rate region for the K users [140] case. In addition, there is no "natural" order of the users, since the channel is not degraded. Therefore, we can permute the users arbitrarily, repeat the above procedure and get another achievable region.

Theorem 4. Consider the Gaussian vector broadcast channel $\boldsymbol{y}_i = H_i\boldsymbol{x} + \boldsymbol{z}_i, i = 1, ..., K$, under a power constraint Γ. The following rate region is achievable:

$$\left\{\begin{array}{l} (R_{\pi(1)}, ..., R_{\pi(K)}) : \\ R_{\pi(i)} \leq \log \frac{H_{\pi(i)} \ \sum_{k=i}^{K} \Sigma_{x_{\pi(k)} x_{\pi(k)}} \ H_{\pi(i)}^{\dagger} + \Sigma_{z_{\pi(i)} z_{\pi(i)}}}{H_{\pi(i)} \ \sum_{k=i+1}^{K} \Sigma_{x_{\pi(k)} x_{\pi(k)}} \ H_{\pi(i)}^{\dagger} + \Sigma_{z_{\pi(i)} z_{\pi(i)}}} \end{array}\right\} \tag{5.4}$$

where π is the permutation function $\pi : \{1, ..., K\} \longrightarrow \{1, ..., K\}$, $\Sigma_{z_{\pi(i)} z_{\pi(i)}}$ is the covariance matrix for $\boldsymbol{z}_{\pi(i)}$, and $\Sigma_{x_{\pi(i)} x_{\pi(i)}}$ is a set of positive semi-definite matrices satisfying the constraint: $\sum_{i=1}^{K} tr(\Sigma_{x_i x_i}) \leq \Gamma$.

We can see that the RHS of (5.4), is neither a concave nor a convex function of the covariance matrices. This is the reason that the capacity region for the non-degraded broadcast channel is very difficult to characterize.

Theorem 4 is a generalization of Caire and Shamai's result [16], for the two-user case. They have succeeded to find the sum capacity for the two-user case by finding an optimal set of $(\Sigma_{x_1x_1}, \Sigma_{x_2x_2})$, and proved that the dirty-paper achievable rate region coincides with an outer bound for the rate region.

The capacity region of a Gaussian MIMO broadcast channel was characterized by Weingarten, Steinberg and Shamai in [130]. This capacity region coincides with the dirty-paper rate region given in Theorem 4.

5.2 Watermarking

Several applications in information hiding require a system that modifies an original host signal (covertext) in order to embed some extra information (digital watermark). The embedding must not cause a noticeable distortion relative to the host signal. We use a distortion measure between the host and watermarked signals, and we constrain the embedding distortion to be at most D_1, as we shall see later. The embedding should also be robust to attacks on the watermarked signal (composite signal). In some applications, the attacks on the watermarked signal are the result of standard signal processing operations, in other cases they are malicious.

Watermarking (information embedding) is a model for copyright protection schemes for audio, video and images that are distributed in digital formats. The embedded signal in watermarking, either notifies a recipient of any copyright or licensing restrictions or inhibits unauthorized copying. This watermark could be, for example, a digital "fingerprint" that uniquely identifies the original purchaser of the authorized copy. If illicit copies were made from the authorized copy, all copies would carry this fingerprint, thus identifying the owner of the authorized copy from which all illicit copies were made. The watermark could also enable or disable copying by a duplication device that will check the watermark before making another copy. This watermark could also be checked by a disc player in order to decide whether or not to play the disc. In other applications, the watermark signal is used for authentication check or detection of tampering to the host signal. Watermarking can also be used for a type of covert communication called “steganography”, in this application a secret message is embedded within the host signal.

Other applications of information hiding include monitoring of airplay of advertisements on commercial radio broadcast. These advertisements could be embedded with a digital watermark and the advertisers could count the number of times the watermark occurs during a given broadcast period, as promised by the radio station operators. Another application for which a watermarking model is used could be a backwards-compatible upgrading of an existing communication system [97]. Examples include standard AM/FM radio and analog TV (NTSC/PAL/SECAM). In this application, we would like to simultaneously transmit a digital signal with existing analog commercial radio or TV without interfering with conventional analog reception. The analog signal in this application, serves as the host signal and the digital signal serves as the watermark.

Information embedding can be viewed as a problem of channel coding with side information, where the host signal plays the role of side information and the power constraint is replaced with a distortion constraint. This distortion constraint is the reason why watermarking can be modeled as a channel with artificial states and CSIT. We refer to the information embedding case where both the encoder and decoder have the same side information signal as *private* information embedding, and to the case where only the encoder has side information as *public* information embedding.

In this section, we consider the model depicted in Fig. 5.2. In this model, we assume a host signal source producing random variables $\{S_n\}$ taking values in $\mathcal{S}$ according to the distribution $P_S(s)$, a side information source producing random variables $\{K_n\} \in \mathcal{K}$ distributed as $P_K(k)$, and a message source producing a message M from a message set $\mathcal{M}$. We assume that the side information K^N is available to both the encoder and the decoder, but not to the attacker. This side information is important for two reasons. First, it may provide a source of randomness that is also known to the decoder and enable the use of random-

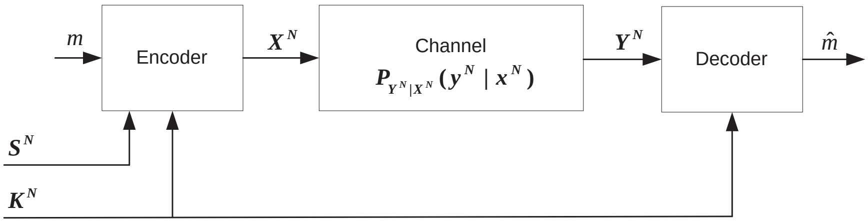

Fig. 5.2 Watermarking model.

ized codes as a means of protection against malicious attacks. Second, it may provide side information about S^N to the decoder. When K^N provides perfect side information, i.e., $K^N = S^N$, the model is suitable to describe a private information embedding scenario, and when K^N is independent of S^N, the model is suitable to describe a public information embedding scenario.

In this model, S^N is not really part of the channel, but it can influence the channel output through a distortion constraint between the host and the watermarked signals. This distortion constraint will be described later. Here, the attacker is represented by the channel $P_{Y^N|X^N}(y^N|x^N)$ which is called the attack channel.

The information hider applies an encoding function f, producing the watermarked signal $x^N = f(s^N, m, k^N)$ that is made publicly available. This watermarked signal is conveyed through an attack channel $P_{Y^N|X^N}(y^N|x^N)$ that produces corrupted data (forgery) Y^N, in an attempt to remove the watermark m from x^N. We assume that the attacker knows the distributions of all random variables in the problem and the actual information hiding code used, but not the side information S^N. The decoder uses Y^N and K^N in order to produce an estimate of the watermark $\hat{m} = g(y^N, k^N)$.

The information hider and the attacker are subjected to distortion constraints between the host and watermarked signals. We define a distortion function for the information hider as a nonnegative function $d_1 : \mathcal{S} \times \mathcal{X} \longrightarrow \mathbb{R}_+$. This constraint replaces the power constraint which was so far being used. The distortion function for the attacker is defined as a nonnegative function $d_2 : \mathcal{X} \times \mathcal{Y} \longrightarrow \mathbb{R}_+$.

The distortion functions $d_1(s,x)$ and $d_2(x,y)$ are extended to distortion functions on N-tuples by $d_1^N(s^N,x^N) = \frac{1}{N}\sum_{n=1}^N d_1(s_n,x_n)$ and $d_2^N(x^N,y^N) = \frac{1}{N}\sum_{n=1}^N d_2(x_n,y_n)$, respectively.

A length-N information-hiding code subject to distortion D_1 is a triple $(\mathcal{M},f,g)$, where:
$\mathcal{M}$ is the message set of cardinality $|\mathcal{M}|$, $f : \mathcal{S}^N \times \mathcal{M} \times \mathcal{K}^N \longrightarrow \mathcal{X}^N$ is the encoder mapping a sequence s^N, a message m, and side information k^N to a sequence $x^N = f(s^N,m,k^N)$. This mapping is subject to the distortion constraint

$$\sum_{s^N\in\mathcal{S}^N}\sum_{k^N\in\mathcal{K}^N}\sum_{m\in\mathcal{M}}\frac{1}{|\mathcal{M}|}P_{S^N,K^N}(s^N,k^N)d_1^N(s^N,f(s^N,m,k^N)) \le D_1, \tag{5.5}$$

and $g : \mathcal{Y}^N \times \mathcal{K}^N \longrightarrow \mathcal{M}$ is the decoder mapping the received sequence y^N and the side information k^N to a decoded message $\hat{m} = g(y^N,k^N)$.

Definition 7. A memoryless attack channel subject to distortion D_2 is a family of conditional output distributions $\{P_{Y|X}(y|x) : x \in \mathcal{X}, y \in \mathcal{Y}\}$ such that $P_{Y^N|X^N}(y^N|x^N) = \prod_{n=1}^N P_{Y|X}(y_n|x_n)$,

$$\sum_{x^N\in\mathcal{X}^N}\sum_{y^N\in\mathcal{Y}^N} d_2^N(x^N,y^N)P_{Y^N|X^N}(y^N|x^N)P_{X^N}(x^N) \le D_2 \tag{5.6}$$

for all $N \ge 1$.

We define the support set of $P_{S,K}(s,k)$,

$$\Omega = \{(s,k) \in \mathcal{S}\times\mathcal{K} : P_{S,K}(s,k) > 0\}. \tag{5.7}$$

Consider an auxiliary random variable U defined over a finite set $\mathcal{U}$ of cardinality $|\mathcal{U}| \le |\mathcal{X}||\Omega| + 1$. When the attack channel $P_{Y|X}(\cdot|\cdot)$ is a fixed known one, the information hiding capacity is given by [89]

$$C = \max I(U;Y|K) - I(U;S|K), \tag{5.8}$$

where the maximum is taken over all distributions $P_{X,U|S,K}(x,u|s,k)$ and functions f satisfying (5.5). In the more general case, information hiding can be thought of as a game between two players, the information hider and the attacker, in cases where the attack channel is not fixed and known. The first player tries to maximize a payoff function (e.g., achievable rate), and the second one tries to minimize it. The information available to each player critically determines the value of the game. In our scenario, we assume that the information hider chooses the encoder f and the attacker is able to learn f and choose the attack channel $P_{Y^N|X^N}(y^N|x^N)$ accordingly. We also assume that the decoder knows the attack channel $P_{Y^N|X^N}(y^N|x^N)$ and chooses g accordingly. These assumptions may be too optimistic. In [31],[108] a more conservative approach for the information hider and the decoder is to assume that they are unable to know $P_{Y^N|X^N}(y^N|x^N)$, but the attacker is able to find out both f and g and design $P_{Y^N|X^N}(y^N|x^N)$ accordingly.

An expression for the information-hiding capacity is derived in terms of optimal covert and attack channels [89].

Definition 8. A memoryless covert channel subject to distortion D_1 is a conditional distribution $P_{X,U|S,K}(x,u|s,k)$ from $\mathcal{S} \times \mathcal{K}$ to $\mathcal{X} \times \mathcal{U}$, such that

$$\sum_{x,s,k,u} d_1(s,x) P_{X,U|S,K}(x,u|s,k) P_{S,K}(s,k) \leq D_1. \tag{5.9}$$

The class $\mathcal{Q}$ is the set of all memoryless covert channels subject to distortion D_1. The class $\mathcal{A}(\mathcal{Q}, D_2)$ is the set of all memoryless attack channels subject to distortion D_2 under covert channels from the class $\mathcal{Q}$.

Additional constraints may be imposed on the attack channel. For this reason, we assume that the attack channel belongs to a subset of $\mathcal{A}(\mathcal{Q}, D_2)$. Assume this subset is of the form $\mathcal{A}(\mathcal{Q}) = \mathcal{A}(\mathcal{Q}, D_2) \bigcap \mathcal{B}$, where $\mathcal{B}$ is some compact set of channels.

Definition 9. The class $\mathcal{A}(f, D_2)$ is the set of all memoryless attack channels that satisfy the distortion constraint (5.6) under the information-hiding code $(\mathcal{M}, f, g)$.

By analogy with $\mathcal{A}(\mathcal{Q})$, we also define $\mathcal{A}(f) = \mathcal{A}(f, D_2) \bigcap \mathcal{B}$.

Theorem 5. Assume that for any $N \geq 1$, the attacker knows f, and the decoder knows both f and the attack channel. A rate R is achievable for distortion D_1 and attacks in the class $\{\mathcal{A}(f)\}$ if and only if $R < C$, where

$$C = \max_{P_{X,U|S,K}(x,u|s,k) \in \mathcal{Q}} \min_{P_{Y|X}(y|x) \in \mathcal{A}(\mathcal{Q})} \{I(U;Y|K) - I(U;S|K)\}, \tag{5.10}$$

U is a random variable defined over an alphabet $\mathcal{U}$ of cardinality $|\mathcal{U}| \leq |\mathcal{X}||\Omega|$, and the random variables U, S, K, X, Y are jointly distributed as
$P_{U,S,K,X,Y}(u,s,k,x,y) = P_{S,K}(s,k)P_{X,U|S,K}(x,u|s,k)P_{Y|X}(y|x)$, i.e., $(U,S,K) \longrightarrow X \longrightarrow Y$ forms a Markov chain.

It can be easily seen, that in the special case $K = S$, i.e., when the host signal is also available to the decoder (private information embedding), the hiding-capacity is given by

$$C = \max_{P_{X|S}(x|s)} \min_{P_{Y|X}(y|x)} I(X;Y|S), \tag{5.11}$$

where the maximization and minimization are subject to the distortion and $\mathcal{B}$ constraints. Theorem 5 considers only memoryless attack channels. The results for the private information embedding and the public information embedding settings, were extended in [109] and [108], respectively, to any attack channel $P_{Y^N|X^N}(\cdot|\cdot)$ that satisfies a distortion constraint. These extended results agree with (5.10) and (5.11) when we consider only memoryless attack channels under the expected distortion constraints given by (5.6).

In Theorem 5, the cardinality of the auxiliary random variable U was mistakenly upper bounded by $|\mathcal{U}| \leq |\mathcal{X}||\Omega|$. The minimization over the attack channel in Theorem 5 means that in our setting, unlike Gel'fand and Pinsker's setting, the attack channel in not a known fixed one, and is not necessarily memoryless. Therefore, there is no obvious upper bound on the cardinality of the auxiliary random variable U, and one cannot apply Carathéodory's theorem straightforwardly, but rather to a finite set of conditional distributions $P_{Y|X}(\cdot|\cdot)$ (whose size depend on a number $l > 0$) that approximate the set of attack channels $\{P_{Y|X}(y|x) : x \in \mathcal{X}, y \in \mathcal{Y}\}$ which satisfy (5.6). Instead of the infinitely many channels in $\{P_{Y|X}(y|x) : x \in \mathcal{X}, y \in \mathcal{Y}\}$ there are $(l+1)^{|\mathcal{X}||\mathcal{Y}|}$ channels to which we can apply Carathéodory's theorem. The results in [108] quantify the loss, in terms of rate, when the encoder employs an auxiliary random variable with cardinality $(l+1)^{|\mathcal{X}||\mathcal{Y}|}$, where the rate loss tends to zero as l tends to infinity (see the result of [108]).

So far we have considered information embedding schemes, for which the embedder distorts the host signal. In [68], the problem of information hiding without the distortion of the host signal is considered, i.e., the host signal can be recovered reliably by the decoder. This type of information hiding is called reversible information hiding. Reversible information hiding is important in applications where no degradation of the original host is allowed. These applications include medical imagery, military imagery and multimedia archives of valuable original works.

In reversible information hiding scheme, the decoder has to produce an estimate of the host sequence as well as an estimate of the embedded message, such that $P_e \triangleq P_r(\hat{S}^N \neq S^N \bigcup \hat{m} \neq m)$, is small. As in standard information embedding, the embedder is subjected to a distortion constraint between the host and the watermarked signals $D_1 \geq \sum_{x,s} P_{X,S}(x,s)d(x,s)$. For the case of noiseless channel $P_{Y|X}(\cdot|\cdot)$ (i.e., no attack channel) the capacity is given by [68]

$$C = \max H(X) - H(S), \tag{5.12}$$

where the maximization is over all $P_{X,S}(x,s)$ satisfying the distortion constraint D_1. In a similar manner, if an attack channel $P_{Y|X}(\cdot|\cdot)$ is present, the capacity for a distortion D_1 is given by

$$C = \max I(Y;X) - H(S), \tag{5.13}$$

where the maximization is taken over all joint distributions satisfying

$$P_{X,Y,S}(x,y,s) = P_S(s)P_{Y|X}(y|x)P(X|S)(x|s). \tag{5.14}$$

Although the embedder distorts the host signal, transmitting at rates below (5.13) results in small P_e. Therefore, the receiver can estimate the host signal reliably.

The reversible information embedding problem can be extended to multi-user channels. In [73], Kotagiri and Leneman have found the reversible information embedding capacity for the two-user multiple access channel, two-user degraded broadcast channel and the two-user degraded relay channel. The capacity for the single-user scenario given in (5.13), is a special case of the capacity of the multiple access channel. Reversible information embedding for the two-user degraded broadcast channel is discussed in Subsection 6.6.2.

Due to the requirement to produce an estimate of the host sequence at the decoder, in cases where the host entropy is larger than the channel capacity, no communication can take place under a complete reconstruction requirement. In [133] a relaxed version of the problem is considered, where instead of an exact reconstruction, the decoder is required to reconstruct the host within a distortion level D_2. However, for very small distortion levels D_2 at the destination, there is a high penalty to pay in the embedding rate. In [113], Steinberg suggested as a possible solution to this problem to provide the decoders, a priori, with a compressed version of the host. Steinberg considered the problem of reversible information embedding for the degraded broadcast channel where a compressed host data is available, before transmission, at the decoders. This model corresponds to the scenario where the composite data is subjected to several stages of attack.

In [80] the problem of joint information hiding and lossy compression is considered. Consider the model depicted in Fig. 5.3. Due to

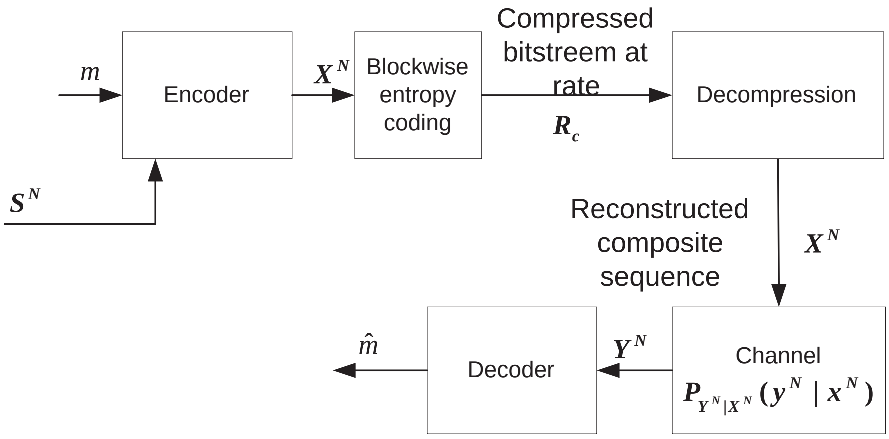

Fig. 5.3 Joint information hiding and compression Model.

bandwidth or storage constraints, we are interested in the compression of the watermarked signal before transmission. The model depicted in Fig. 5.3 could be used for the following scenario. A high-resolution image (e.g., a satellite image) is watermarked for copyright protection and compressed before it is transmitted through a noiseless channel (e.g., the Internet) to a costumer. The costumer decompresses the image and use it. If illicit copies are made from this authorized decompressed image, all copies will carry the watermark message, thus identifying the original owner of this image.

The difference between this model and the model presented in Fig. 5.2 is the absence of the auxiliary random variable K and the compression of the watermarked signal X^N. The information hider, in this setting, conveys S^N and the message m through an encoding function f, by producing the watermarked signal $X^N = f(S^N, m)$. Here, the watermarked signal X^N is entropy-coded, i.e., compressed in a blockwise manner using the optimum lossless code and the corresponding watermarked signal rate is defined by

$$\frac{H(f(S^N, m))}{N} \tag{5.15}$$

and should not exceed a prescribed value R_c. The compressed watermarked signal is sent to the decoder.

Let $\mathcal{A}$ denote the set of all triples (U, S, X) of random variables taking values in the finite sets $\mathcal{U}, \mathcal{S}, \mathcal{X}$, where $\mathcal{U}$ is an arbitrary finite alphabet of size $|\mathcal{U}| \leq |\mathcal{S}||\mathcal{X}|$, and the joint probability distribution of (U, S, X),$P_{U,S,X}(u, s, x)$, is such that the marginal distribution of S is $P_S(\cdot)$, and $\sum_{s^N \in \mathcal{S}^N} \sum_{m \in \mathcal{M}} \frac{1}{|\mathcal{M}|} P_{S^N}(s^N) d_1^N(s^N, f(s^N, m)) \leq D_1$. For any triple (U, S, X), there exists a related quadruple (U, S, X, Y), with Y taking values in $\mathcal{Y}$, such that

$$P_{U,S,X,Y}(u, s, x, y) = P_{U,S,X} P_{Y|X}(y|x). \tag{5.16}$$

Let $R(D_1)$ denote the rate-distortion function of the source $P_S(\cdot)$ [36]. The information hiding capacity for a discrete memoryless covertext S, a memoryless attack channel $P_{Y^N|X^N}(y^N|x^N)$ and $R_c \geq R(D_1)$ is given by

$$C = \max_{(U,S,X) \in \mathcal{A}} \min \{I(U; Y) - I(U; S), R_c - I(S; U, X)\}. \tag{5.17}$$

Here an alternative coding scheme to Gel'fand and Pinsker's coding scheme is proposed, which takes into account the compression. This coding scheme achieves the capacity of (3.14), and also allows us to characterize the tradeoff between the information hiding and the compression rates.

First, we will describe the code generation. For each message m, generate 2^{NR_0} codewords (or a bin) $U^N(m,j) \in \{u^N(m,1),...,u^N(m,2^{NR_0})\}$, where $m \in \mathcal{M}$ is the bin index ($|\mathcal{M}| = 2^{NR}$) and $j \in \{1,...,2^{NR_0}\}$ is the codeword index inside the bin m, i.i.d. according to the distribution $P_U(\cdot)$. For each codeword $U^N(m,j) = u^N(m,j)$, generate 2^{NR_x} composite sequences $X^N(m,j,k) \in \{x^N(m,j,1),...,x^N(m,j,2^{NR_x})\}$, where k is the composite sequence index inside the set indexed by (m,j), i.i.d. according to the distribution $P_{X|U}(\cdot|\cdot)$. Let us denote this set by $\mathcal{C}(m,j) = \{x^N(m,j,1),...,x^N(m,j,2^{NR_x})\}$.

Encoding is done in the following way: given the message m and the state sequence s^N, the encoder seeks a codeword in bin m that is jointly typical with s^N, say $u^N(m,j)$. The first composite sequence found in $\mathcal{C}(m,j)$ that is jointly typical with $(s^N, u^N(m,j))$, say $x^N(m,j,k)$, is chosen for transmission. If there exist more than one such sequence, the described above process is applied to the first matching $u^N(m,j)$ found in a bin's list. If no such $u^N(m,j)$ exists such that $(s^N, U^N(m,j), x^N(m,j,k))$ are jointly typical declare an encoding error.

The decoder finds $\hat{m}$ and $\hat{j}$ such that $u^N(\hat{m},\hat{j})$ is jointly typical with the channel output sequence y^N. If there exist more than one such pair $(\hat{m},\hat{j})$, or no such pair exit at all, declare a decoding error. The probability of encoding failure goes to zero as long as $R_0 \geq I(U;S)$ and $R_x \geq I(S;X|U)$, and the probability of decoding failure goes to zero as long as $R + R_0 \leq I(U;Y)$. Thus, the overall probability of error goes to zero as long as $R \leq I(U;Y) - I(U;S)$ and $R_x \geq I(S;X|U)$. Now, since the compression procedure applied to the composite sequences is lossless, it satisfies $R_c \geq R + R_x \geq R + I(S;U,X)$. Therefore, $R \leq \min\{I(U;Y) - I(U;S), R_c - I(S;U,X)\}$.

In [84], a similar problem with causal covertext was analyzed. This setting is closely related to Shannon's channel with causal CSIT [105] as the non-causal covertext setting is related to Gel'fand and Pinsker's channel [56].

Let $\mathcal{T}$ denote the set of functions $t : \mathcal{S} \longrightarrow \mathcal{X}$. The information hiding capacity for a causal discrete memoryless covertext is given by [84]

$$C = \max_{T:Ed_1(s,t(s))\leq D_1} \min[I(T;Y), R_c - H(T(S)|T)], \tag{5.18}$$

where the joint distribution of S, T, Y is given by

$$P_{S,T,Y}(s,t,y) = P_S(s)P_T(t)P_{Y|X}(y|t(s)). \tag{5.19}$$

The non-causal covertext setting was extended in [83], where the problem of joint information hiding, compression and encryption was considered. The encryption is done in order to protect the secrecy of the watermark against an unauthorized party, which has no access to a secret key shared by the legitimate parties. In the attack-free case, if the key is independent of the covertext, a separation principle was shown to exist. This separation principle, tells us that asymptotically, for long block codes, there is no optimality loss by first applying a rate-distortion code to the watermark source, then encrypting the compressed codeword, and finally, embedding it into the covertext. If we add an attack channel to the problem, this separation principle is no longer valid, as the key may play an additional role of side information used by the embedder (for more details see [83]).

We next consider the case of watermarking in Gaussian attack channels with Gaussian covertext. Consider the case of a Gaussian S and the squared-error distortion measure $d(x,y) \triangleq d_1(x,y) = d_2(x,y) = (x-y)^2$. Here $\mathcal{S} = \mathcal{X} = \mathcal{Y} = \mathbb{R}$, and $S \sim \mathcal{N}(0,\sigma^2)$. The class of attack channels is $\mathcal{A}(\mathcal{Q}, D_2)$. This case was studied in [89],[31] and is particularly interesting. We will see that we can use Costa's dirty-paper model for this case. Therefore, the capacity of the private and public versions of the problem are the same and it becomes possible to explicitly compute the distributions that achieve capacity.

We start with the public version of the problem. In this case, the host signal is also available to the decoder, and the capacity is given in Theorem 6.

Theorem 6. [89] Let $\mathcal{S} = \mathcal{X} = \mathcal{Y} = \mathbb{R}$ and $d(x, y) = (x - y)^2$ be the squared-error distortion measure. Assume that $K = S$. Let a be the maximizer of the function

$$f(a) = \frac{[(2a-1)\sigma^2 - D_2 + D_1][D_1 - (a-1)^2\sigma^2]}{[D_1 + (2a-1)\sigma^2]D_2} \tag{5.20}$$

in the interval $(a_{inf}, 1 + \sqrt{D_1}/\sigma)$, where $a_{inf} = \max(1, \frac{\sigma^2+D_2-D_1}{2\sigma^2})$. Then we have

(i) If $D_2 \geq (\sigma + \sqrt{D_1})^2$, the hiding capacity is $C = 0$.

(ii) If S is non-Gaussian with mean zero and standard deviation $\sigma > \sqrt{D_2} - \sqrt{D_1}$, the hiding capacity is upper-bounded by

$$C = \frac{1}{2}\log\left(1 + \frac{[(2a-1)\sigma^2 - D_2 + D_1][D_1 - (a-1)^2\sigma^2]}{[D_1 + (2a-1)\sigma^2]D_2}\right). \tag{5.21}$$

(iii) If $S \sim \mathcal{N}(0, \sigma^2)$ and $D_2 < (\sigma + \sqrt{D_1})^2$, the hiding-capacity is given by (5.21). The optimal covert channel is given by $X = aS + Z$, where $Z \sim \mathcal{N}(0, D_1 - (a-1)^2\sigma^2)$ is independent of S. The optimal attack is the Gaussian channel

$$P_{Y|X}(y|x) = \mathcal{N}(\beta^{-1}x, \beta^{-1}D_2), \tag{5.22}$$

where $\beta = \frac{\sigma_x^2}{\sigma_x^2 - D_2}$, and $\sigma_x^2 = D_2 + (2a-1)\sigma^2$.

In the asymptotic case of small distortions $D_1, D_2 \longrightarrow 0$, we have

$$\begin{aligned} a &\sim 1 + \frac{D_1 D_2}{3\sigma^4}, \\ \beta &\sim 1, \\ C &\sim \frac{1}{2}\log\left(1 + \frac{D_1}{D_2}\right). \end{aligned} \tag{5.23}$$

The additive white Gaussian noise attack channel $P_{Y|X}(y|x) \sim \mathcal{N}(x, D_2)$ is asymptotically optimal in this case.

For the private version of the problem, i.e., when the host signal is not available to the decoder. The capacity is given in Theorem 7:

Theorem 7. Let $\mathcal{S} = \mathcal{X} = \mathcal{Y} = \mathbb{R}$ and $d(x, y) = (x - y)^2$ be the squared-error distortion measure. Assume that S is $\mathcal{N}(0, \sigma^2)$ and $D_2 <$

$(\sigma+\sqrt{D_1})^2$. Let a be the maximizer of the function (5.20) in the interval $(a_{inf}, 1+\sqrt{D_1}/\sigma)$. Then the following distribution yields the maxmin solution of the game (5.10): $X = aS + Z$ and $U = \alpha + Z$, where $Z \sim \mathcal{N}(0, D_1 - (a-1)^2\sigma^2)$ is independent of S. The optimal attack $P_{Y|X}(y|x)$ is the Gaussian channel (5.22). Here $\beta = \frac{(2a-1)\sigma^2+D_1}{(2a-1)\sigma^2-(D_2-D_1)}$ and $\alpha = \frac{D_1-(a-1)^2\sigma^2}{D_1-(a-1)^2\sigma^2+\beta D_2}$.
The hiding capacity is the same as (5.21) in the private watermarking game.

As in Costa's dirty-paper problem, the capacity of the private and public versions of the problem are the same. Costa's result is discussed in Section 4.1. The optimal distributions that achieve capacity in this case, are the same as Costa's distributions in the dirty-paper problem. The additive white Gaussian noise attack $P_{Y|X}(y|x) = \mathcal{N}(x, D_2)$ is suboptimal but is asymptotically optimal for $\sigma^2 >> D_1, D_2$ (small distortion case). In the small distortion case, $\beta \longrightarrow 1$ and $\alpha \sim \frac{D_1}{D_1+D_2}$.

Costa showed that we can achieve capacity in this case using the binning technique. We will present coding techniques for the watermarking problem as well as for the dirty-paper problem, in Section 7.

6

Other Related Problems

Thus far, we have presented many applications and models, for our main problem of channel coding in the presence of side information. In this section, we consider several other problems which are related to the models discussed in the previous sections. In particular, we present the source coding problem with decoder side information, where we discuss both the lossless Slepian-Wolf problem and the lossy Wyner-Ziv problem. Next, we study the duality between the source coding problem with decoder side information and the problem of channel coding with transmitter side information. We also consider the problem of joint source channel coding for the Wyner-Ziv source and the Gel'fand-Pinsker channel and present, as an application for this model, a backward-compatible upgrading technique of an existing communication system. Then we present the problem of achievable tradeoff between message and state information rates, and also discuss a different aspect of the problem, where the state information is considered as undesirable information. Finally, we consider multiuser channels with states, and in particular, the multiple access channel and the broadcast channel, both are controlled by a state process.

This subsection by no means includes a complete coverage of all problems related to the models presented in this paper. We have chosen to include only several selected problems, as examples to the importance of the ideas discussed in this paper.

6.1 The Slepian-Wolf Problem

We consider the problem of lossless source coding of correlated sources. This problem is known as the Slepian-Wolf problem [107]. Slepian and Wolf showed that separate source coding of correlated sources (with increased complexity at the joint decoder) is as efficient as lossless joint source coding in terms of total rate. Example for an applications for the Slepian-Wolf problem is the distributed sensor network presented in [137]. In this example, we want to compress multiple correlated sensors, that do not communicate with each other, and send the sensors compressed outputs to a base station for joint decoding. These sensors could be, for example, low-cost video cameras or microphones for survivance applications. Lossless source coding with side information at the decoder, is a special case of the Slepian-Wolf problem, and is the lossless version of the Wyner-Ziv problem. This particular case is depicted in Fig. 6.1.

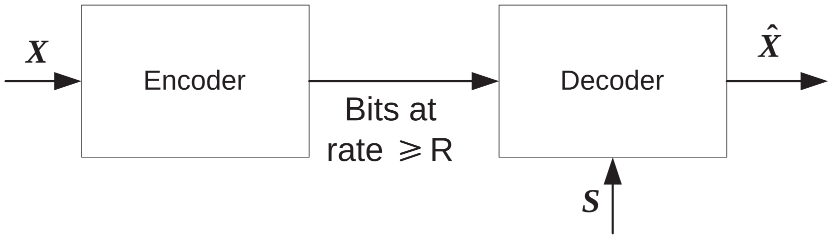

Fig. 6.1 Source coding with side information at the decoder.

Let $\{X_i, S_i\}_{i=1}^{\infty}$ be a sequence of i.i.d. copies of a pair of RV's $(X, S) \sim P_{X,S}(\cdot,\cdot)$. The sequences are available at different locations and therefore have to be separately lossless source coded. This problem is therefore also known as distributed source coding (DSC). We define a source code and an achievable source code as

Definition 10. An $\left(N, \left(2^{NR_1}, 2^{NR_2}\right)\right)$ distributed source code for the joint source (X, S) consists of two encoder maps,

$$\begin{aligned} f_1 : \mathcal{X}^N &\longrightarrow \left\{1, 2, ..., 2^{NR_1}\right\} \\ f_2 : \mathcal{S}^N &\longrightarrow \left\{1, 2, ..., 2^{NR_2}\right\} \end{aligned} \tag{6.1}$$

and a decoder map,

$$g : \left\{1, 2, ..., 2^{NR_1}\right\} \times \left\{1, 2, ..., 2^{NR_2}\right\} \longrightarrow \mathcal{X}^N \times \mathcal{S}^N. \tag{6.2}$$

Definition 11. A rate pair (R_1, R_2) is said to be achievable for a distributed source if there exists a sequence of $\left(N, \left(2^{NR_1}, 2^{NR_2}\right)\right)$ distributed source codes with probability of error $P_e \longrightarrow 0$, where the probability of error is given by

$$P_e = P_r\left(g\left(f_1\left(X^N\right), f_2\left(S^N\right)\right) \neq \left(X^N, S^N\right)\right). \tag{6.3}$$

The achievable rate region is the closure of the set of achievable rates.

Slepian and Wolf's main result is the following theorem,

Theorem 8. [107] For the distributed source coding problem of the source (X, S) drawn i.i.d. $\sim P_{X,S}(\cdot, \cdot)$, the achievable rate region is given by

$$\begin{aligned} R_1 &\geq H(X|S) \\ R_2 &\geq H(S|X) \\ R_1 + R_2 &\geq H(X, S). \end{aligned} \tag{6.4}$$

The achievable rate for the lossless source coding problem, with side information at the decoder, is by Theorem 8 given by

$$R \geq H(X|S). \tag{6.5}$$

The achievability of the Slepian-Wolf theorem, is proved using the random binning concept [36]. First, independently assign every $x^N \in \mathcal{X}^N$ to one of 2^{NR_1} bins according to a uniform distribution on $\{1, ..., 2^{NR_1}\}$. Similarly, randomly assign every $s^N \in \mathcal{S}^N$ to one of 2^{NR_2} bins. Reveal the assignments f_1 and f_2 to both the encoder and decoder. Given a source sequences $x^N \in \mathcal{X}^N$ and $s^N \in \mathcal{S}^N$, sender 1 sends the index of the bin to which x^N belongs, and sender 2 sends the index of the bin to which s^N belongs. Upon receiving the bin index pair (i, j), the receiver declare $(\hat{x}^N, \hat{s}^N) = (x^N, s^N)$, if there is one and only one pair of sequences (x^N, s^N) such that $f_1(x^N) = i$, $f_2(s^N) = j$ and x^N, s^N are jointly typical. Otherwise declare an error.

This is another example of the importance of the random binning concept which is the fundamental tool for proving the achievability of the main information-theoretic problems explored in this paper. We note here that in a chronological order, the Slepian-Wolf problem, was the first problem which used the random binning concept, back in 1973.

6.2 The Wyner-Ziv Problem

In this subsection, we consider the problem of lossy source coding (rate-distortion) with decoder side information under a distortion constraint. There is a large number of applications for this problem. These applications include distributed sensor networks and sensor arrays, as in the Slepian-Wolf problem, and also, digital upgrade of analog television signals, and communication in ad-hoc networks. Wyner and Ziv [135] found the rate-distortion function for this problem. We will see in Subsection 6.3 that this problem and the problem of channel coding with side information are information-theoretic duals of each other.

Consider the problem of rate-distortion optimal lossy encoding of a source X with side information S available to the decoder as shown in Fig 6.1. X and S are correlated random variables with joint distribution $P_{X,S}(x, s)$ which take values in the finite sets $\mathcal{X}$, $\mathcal{S}$, respectively. Let $d(x, \hat{x})$ be a distortion measure between the source X and it's reconstruction $\hat{X} \in \hat{\mathcal{X}}$. We say that rate R is achievable at distortion level D if there exist an encoding function $f : \mathcal{X}^N \longrightarrow \{1, 2, ..., 2^{NR}\}$, and a decoding function $g : \{1, 2, ..., 2^{NR}\} \times \mathcal{S}^N \longrightarrow \hat{\mathcal{X}}^N$ such that $E[d(X^N, \hat{X}^N)] \leq D$. The rate-distortion function $R(D)$ is the infimum of the achievable rates with distortion D. Wyner and Ziv [135] found that the rate-distortion function is given by

$$R(D) = \min_{P_{U|X}(u|x) P_{\hat{X}|U,S}(\hat{x}|u,s)} [I(U; X) - I(U; S)] \tag{6.6}$$

where the minimum is over all $P_{U|X}(u|x)P_{\hat{X}|U,S}(\hat{x}|u,s)$ such that $U \longrightarrow X \longrightarrow S$ and $\hat{X} \longrightarrow (U,S) \longrightarrow X$ form a Markov chains, and

$$\sum_{x,\hat{x},u,s} P_{X,S}(x,s)P_{U|X}(u|x)P_{\hat{X}|U,S}(\hat{x}|u,s)d(x,\hat{x}) \leq D, \tag{6.7}$$

where U is an auxiliary random variable with alphabet cardinality satisfying $|\mathcal{U}| \leq |\mathcal{X}| + 1$. Similarly to the Gel'fand-Pinsker problem, it has been shown in [135] that there is no loos of performance in restricting $P_{\hat{X}|U,S}(\cdot|u,s)$ to a deterministic function $\widetilde{f} : \mathcal{S}^N \times \mathcal{U}^N \longrightarrow \hat{\mathcal{X}}^N$ (see the converse proof in [135]).

The achievability of Wyner and Ziv's result, is proved using the random binning technique. Generate 2^{NR_1} i.i.d. codewords $U^N(i) \sim \prod_{n=1}^{N} P_U(u_n)$, and index them by
$i \in \left\{1,...,2^{NR_1}\right\}$. Randomly assign the indices $i \in \left\{1,...,2^{NR_1}\right\}$ to one of 2^{NR_2} bins using a uniform distribution over the bins. Let $B(j)$ denote the indices assigned to bin j. Given a source sequence $x^N \in \mathcal{X}^N$, the encoder looks for a codeword that is jointly typical with x^N, say $u^N(i)$. If there is no such $u^N(i)$, the encoder declares an encoding failure. If there is more than one such i, the encoder uses the lowest i. The encoder sends the index of the bin in which i belongs, i.e., the bin index j is sent if $i \in B(j)$. Upon receiving the bin index j, the decoder looks in bin j for a codeword $u^N(i)$, $i \in B(j)$ that is jointly typical with s^N. If he finds a unique i, he calculates $\hat{x}^N = \widetilde{f}\left(s^N, u^N(i)\right)$. If it does not find any such i or more than one such i, the decoder declares a decoding failure. With $R_1 = I(X;U)$ and $R_2 = I(X;U) - I(S;U))$ the probability of encoding or decoding failures goes to zero. Hence, with high probability, the decoder will produce $\hat{x}^N$ such that (6.7) is satisfied. The probability that (6.7) is not satisfied, is called the probability of distortion violation and is denoted by P_{dv}. Therefore, this probability goes to zero in the binning scheme.

We can also consider a causal version of the Wyner-Ziv problem, which is the source-coding dual of the Shannon channel model with causal side information at the transmitter. Usually when a side information signal is available to the decoder, we consider it as a non-causal side information, because the decoder can wait to the end of transmission and then decode the messages. However, in some situations, the source coding system is constrained to operate with no, or with limited delay. This causal version of the Wyner-Ziv problem was considered in [55].

Let $D_{min} = E\left[\min_{\hat{x}} d(x,\hat{x})\right]$, where $D < D_{min}$ is not achievable at any rate. The causal rate-distortion function for distortion level $D > D_{min}$ is given by

$$R(D) = \min I(U;X) \tag{6.8}$$

where the minimum is over all functions $\widetilde{f} : \mathcal{U} \times \mathcal{S} \longrightarrow \hat{\mathcal{X}}$, $|\mathcal{U}| \leq |\mathcal{X}|+1$, and $P_{U|X}(\cdot|x)$ such that $E\left[d\left(x, \widetilde{f}\,(u,s)\right)\right] \leq D$. Here, similarly ro Shannon's channel coding problem with causal side information, there is no binning scheme.

The difference between (6.8) and (6.6) is the subtracted term $I(U;S)$. If we look at the direct part of Wyner and Ziv's proof presented above, this term is a consequence of the binning technique which partitions an $(N, 2^{NI(U;X)})$ code into bins with $2^{NI(U;S)}$ codewords, and allows the encoder to send only the bin index instead of the codeword index. The minimization for the two versions of the problem is over exactly the same set. This causal version of the problem, is the dual problem to Shannon's channel coding problem with causal side information. We discuss this duality in Subsection 6.3. Next, we turn back to the non-causal scenario.

Wyner and Ziv have also treated the case of a joint Gaussian source and squared-error distortion measure. Let $X^N \sim \mathcal{N}(0, \sigma_x^2 I)$. The decoder has access to a noisy observation of the source $S^N = a(X^N + Z^N)$, where a is a positive number, and $Z^N \sim \mathcal{N}(0, \sigma_z^2 I)$ is Gaussian noise. We use the squared-error distortion measure between X and $\hat{X}$, i.e., $d(x,\hat{x}) = (x-\hat{x})^2$. The rate-distortion function is given by

$$R(D) = \begin{cases} \frac{1}{2}\log_2\left(\frac{\sigma_z^2\sigma_x^2}{D(\sigma_z^2+\sigma_x^2)}\right), & 0 \leq D \leq \frac{\sigma_z^2\sigma_x^2}{\sigma_z^2+\sigma_x^2}; \\ 0, & D > \frac{\sigma_z^2\sigma_x^2}{\sigma_z^2+\sigma_x^2}. \end{cases} \tag{6.9}$$

Wyner showed in [136], that if X and S are jointly Gaussian, then the rate-distortion function takes the form

$$R(D) = \frac{1}{2}\log_2\left(\frac{B}{D}\right) \tag{6.10}$$

where $B = \sigma^2_{x|s}$ is the conditional variance of X given S.

Similarly to Costa's result for the dirty-paper problem, the rate distortion in (6.10) is the same rate-distortion function has if the side information S^N is known to both the decoder and encoder or if it was not present at all.

An interesting application for this problem was given in [117]. Consider combining images from a space-based telescope and ground-based observatory. Both simultaneously observe the same object in space. X corresponds to the image at the telescope, which encounters no atmospheric interference, and S to the image at the observatory suffering atmospheric interferences. The telescope transmits information at rate R to the observatory, which computes the reconstructed image $\hat{X}$. Wyner and Ziv's result means that we can transmit at a lower rate than conventional lossy source codding without sacrificing image quality.

6.3 The Duality Between Source Coding and Channel Coding with Side Information

In this subsection, we consider the duality between source coding problem with receiver side information (SCRSI) under a distortion constraint and channel coding problem with transmitter side information (CCTSI) under a power constraint. We will characterize the conditions under which there is a functional duality between these problems. A functional duality means that given an optimal source (respectively, channel) coding scheme, this scheme is a functional dual to a channel (source) coding scheme in the sense that the optimal encoder mapping for one problem is functionally identical to the optimal decoder mapping for the other problem and the input-output joint distribution is the same with some renaming of variables. This duality is used for solving practical source (respectively, channel) coding problems by solving the dual problem of channel (respectively, source) coding. It is used to extend the Gaussian Wyner-Ziv result (6.10) via establishing a duality to the Costa problem.

The capacity of the channel coding problem with transmitter side information under a given power constraint Γ is given by (3.14) and the rate-distortion function for the source coding problem with receiver side information is given by (6.6). We start by stating the correspondence between channel capacity and rate-distortion and between the variables involved in the two coding problems:

$$\left\{\begin{array}{rcl} SCRSI & \longleftrightarrow & CCTSI \\ R(D) & \longleftrightarrow & C \\ minimization & \longleftrightarrow & maximization \\ source\ input\ X & \longleftrightarrow & Y\ channel\ output \\ side\ information\ S & \longleftrightarrow & S\ side\ information \\ auxiliary\ variable\ U & \longleftrightarrow & U\ auxiliary\ variable \\ source\ reconstruction\ \hat{X} & \longleftrightarrow & X\ channel\ input \end{array}\right\}. \quad (6.11)$$

The formula duality is evident, the maximization of achievable data rates given by (3.14) has a formula dual to the minimization of data rates given by (6.6). Next, we will use the notation of the SCRSI problem for the CCTSI problem to characterize the conditions for a functional duality between the two coding problems.

From the definition of the SCRSI problem, there are two Markov chains in this problem: $U \longrightarrow X \longrightarrow S$ and $\hat{X} \longrightarrow (U,S) \longrightarrow X$. Similarly in the definition of the CCTSI problem, there are two Markov chains: $U \longrightarrow (\hat{X},S) \longrightarrow X$ and $\hat{X} \longrightarrow (U,S) \longrightarrow X$ (where these two Markov chains are written here using the notation of the SCRSI problem). The second Markov chain in the CCTSI problem, follows since the optimal $P_{\hat{X}|U,S}(\cdot|u,s)$ in (3.14) can be restricted to a deterministic function. In the SCRSI problem, $P_{\hat{X}|U,S}(\cdot|u,s)$ can also be restricted to a deterministic function, however, this restriction follows from the definitions of the SCRSI problem (see the converse proof in [135]).

Using the first Markov chain in the SCRSI problem, it can be seen that $I(U;X) - I(U;S) = I(U;X|S)$ and therefore, (6.6) can be rewritten as

$$R(D) = \min_{P_{U|X}(u|x)P_{\hat{X}|U,S}(\hat{x}|u,s)} \left[I(U;X|S)\right].$$

The CCTSI problem, cannot be rewritten in the same manner as the SCRSI problem.

For a functional duality between the two problems, the minimization in the SCRSI problem given in (6.6) must satisfy an additional Markov chain $U \longrightarrow (\hat{X},S) \longrightarrow X$. The Gaussian case is an example for a subset of this problem, for which all three Markov chains are satisfied and there is a duality between the two problems.

So far we have considered for the SCRSI problem, the distortion measure $d : \mathcal{X} \times \hat{\mathcal{X}} \longrightarrow \mathbb{R}_+$, and for the CCTSI problem the cost measure $\phi : \hat{\mathcal{X}} \longrightarrow \mathbb{R}_+$. In order to formulate the duality between the two problems, we will have to generalize the distortion measure to $d : \mathcal{X} \times \hat{\mathcal{X}} \times \mathcal{S} \longrightarrow \mathbb{R}_+$, and the cost measure to $\phi : \hat{\mathcal{X}} \times \mathcal{S} \longrightarrow \mathbb{R}_+$. These extensions are straightforward. The new distortion measure can be interpreted as the distortion between x and $\hat{x}$ when the outcome of the side information is s. The new cost measure is interpreted in a similar way.

The following theorem characterizes the conditions under which for a SCRSI problem there is a dual CCTSI problem.

Theorem 9 ([95]). For a SCRSI, with a given source $P_{X|S}(x|s)$, side information $P_S(s)$, input, side information, and reconstruction alphabets $\mathcal{X}$, $\mathcal{S}$, and $\hat{\mathcal{X}}$, respectively, a distortion measure $d : \mathcal{X} \times \hat{\mathcal{X}} \times \mathcal{S} \longrightarrow \mathbb{R}_+$, and a distortion constraint D, let the conditional distribution achieving the rate-distortion optimality $R(D)$ be given by

$$\left\{ \begin{array}{l} P^*_{U|X}(u|x) \\ P^*_{\hat{X}|U,S}(\hat{x}|u,s) \end{array} \right\} \triangleq \\ \arg\min_{\left\{ \begin{array}{c} \left\{ \begin{array}{l} P_{U|X}(u|x) \\ P_{\hat{X}|U,S}(\hat{x}|u,s) \end{array} \right\} : \\ \left\{ \begin{array}{l} E[d(\hat{X},X,S)] \le D \\ (U \longrightarrow X \longrightarrow \ S) \\ (\hat{X} \longrightarrow (U,S) \longrightarrow X) \end{array} \right\} \end{array} \right\}} I(U;X) - I(U;S) \tag{6.12}$$

inducing the following distributions:

$P^*_{X,S,\hat{X},U}(x,s,\hat{x},u) = P_S(s)P_{X|S}(x|s)P^*_{U|X}(u|x)P^*_{\hat{X}|U,S}(\hat{x}|u,s)$ and

$$\begin{aligned} P^*_{X|\hat{X},S}(x|\hat{x},s) &\triangleq \frac{\sum_u P^*_{X,S,\hat{X},U}(x,s,\hat{x},u)}{\sum_{u,x} P^*_{X,S,\hat{X},U}(x,s,\hat{x},u)} \\ P^*_{U|S}(u|s) &\triangleq \frac{\sum_{x,\hat{x}} P^*_{X,S,\hat{X},U}(x,s,\hat{x},u)}{P_S(s)} \\ P^*_{\hat{X}|S}(\hat{x}|s) &\triangleq \frac{\sum_{u,x} P^*_{X,S,\hat{X},U}(x,s,\hat{x},u)}{P_S(s)}, \end{aligned} \tag{6.13}$$

where $P^*_{U|X}(\cdot|x)$, $P^*_{U|S}(\cdot|s)$, $P^*_{\hat{X}|U,S}(\cdot|u,s)$, $P^*_{X,S,\hat{X},U}(\cdot,\cdot,\cdot,\cdot)$, $P^*_{X|\hat{X},S}(\cdot|\hat{x},s)$ and $P^*_{\hat{X}|S}(\cdot|s)$ are the optimal distributions and the optimal induced distributions which achieve the minimum in (6.12). If $P^*_{X,S,\hat{X},U}(x,s,\hat{x},u)$ is such that $U \longrightarrow (\hat{X},S) \longrightarrow X$, then $\exists$ a dual CCTSI, for a channel $P^*_{X|\hat{X},S}(x|\hat{x},s)$, having side information $P_S(s)$, input, side information, and output alphabets $\hat{\mathcal{X}}$,$\mathcal{S}$, and $\mathcal{X}$ respectively, a cost measure $\phi : \hat{\mathcal{X}} \times \mathcal{S} \longrightarrow \mathbb{R}_+$, and a cost constraint Γ, such that

- the rate-distortion bound with receiver side information $R(D)$ is equal to the capacity bound with transmitter side information C under a power constraint Γ, i.e.,

$$\begin{aligned} &\min_{\left\{\begin{array}{l} P_{U|X}(u|x), P_{\hat{X}|U,S}(\hat{x}|u,s) : \\ \left\{\begin{array}{l} E[d(\hat{X},X,S)] \le D \\ (U \longrightarrow X \longrightarrow S) \\ (\hat{X} \longrightarrow (U,S) \longrightarrow X) \end{array}\right\} \end{array}\right\}} I(U;X) - I(U;S) \\ = &\max_{\left\{\begin{array}{l} P_{U|S}(u|s), P_{\hat{X}|U,S}(\hat{x}|u,s) : \\ \left\{\begin{array}{l} E[\phi(\hat{X},S)] \le \Gamma \\ (U \longrightarrow (\hat{X},S) \longrightarrow X) \\ (\hat{X} \longrightarrow (U,S) \longrightarrow X) \end{array}\right\} \end{array}\right\}} I(U;X) - I(U;S); \end{aligned} \tag{6.14}$$

- the conditional distributions $P^*_{U|S}(u|s), P^*_{\hat{X}|U,S}(\hat{x}|u,s)$ achieves the capacity. i.e.,

$$\left\{\begin{array}{l} P^*_{U|S}(u|s) \\ P^*_{\hat{X}|U,S}(\hat{x}|u,s) \end{array}\right\} \triangleq$$
$$\arg\max \left\{ \begin{array}{c} \left\{\begin{array}{l} P_{U|S}(u|s) \\ P_{\hat{X}|U,S}(\hat{x}|u,s) \end{array}\right\} : \\ \left\{\begin{array}{l} E[\phi(\hat{X},S)] \leq \Gamma \\ (U \longrightarrow (\hat{X},S) \longrightarrow \ X) \\ (\hat{X} \longrightarrow (U,S) \longrightarrow X) \end{array}\right\} \end{array} \right\} I(U;X) - I(U;S) \tag{6.15}$$

where the cost measure and the cost constraint are given, respectively, by

$$\begin{aligned} \phi(\hat{x},s) &\triangleq c_1 D(P^*_{X|\hat{X},S}(x|\hat{x},s)||P_{X|S}(x|s)) + \theta(s) \\ \Gamma &\triangleq E_{P_S(s)P^*_{\hat{X}|S}(\hat{x}|s)}(\phi(\hat{X},S)) \end{aligned} \tag{6.16}$$

for arbitrary $c_1 > 0$ and $\theta(s)$.

The conditions under which for a CCTSI problem there is a dual SCRSI problem are given in a similar theorem in [95].

Theorem 9 states that for a given SCRSI, there is a dual CCTSI such that the optimal joint distributions for the SCRSI $P^*_{U|S}(u|s)$, $P^*_{\hat{X}|U,S}(\hat{x}|u,s)$, are identical to the optimal joint distributions of the CCTSI, with the appropriate choice of the channel cost measure, and the rate-distortion bound is equal to the capacity bound.

If we look at the direct part of the coding theorems given in [56], [135], we have partitioned a codebook with $2^{NI(X;U)}$ into bins (cosets). In SCRSI, this is a partition of source codebook for quantizing X into bins with codewords serving as a channel codebook for the fictitious channel between U and S. In CCTSI, this is a partition of a channel codebook for the fictitious channel between U and X into bins with codewords serving as a source codebook for quantizing S. The jointly typical encoding operation in SCRSI is identical to the jointly typical decoding operation in CCTSI. The jointly typical decoding operation in SCRSI is identical to the jointly typical encoding operation in CCTSI.

Thus, the encoder (decoder, respectively) of SCRSI and the decoder (encoder, respectively) of CCTSI are functionally identical.

We can also consider the operational duality between the Gel'fand-Pinsker and the Wyner-Ziv problems. This operational duality was explored for these problems in [129]. We consider a deterministic binning scheme instead of the random binning scheme, and evaluate the performance on the corresponding problems.

Consider a fixed joint probability distribution $P_{X,\hat{X},S,U}(\cdot,\cdot,\cdot,\cdot)$ on the random variables X, $\hat{X}$, S and U having the following properties.

(1) The value of the conditional distribution $P_{\hat{X}|S,U}(\cdot|s,u)$ can only be 0 or 1 and is determined by a deterministic function $f : \mathcal{U} \times \mathcal{S} \longrightarrow \hat{\mathcal{X}}$.
(2) $U \longrightarrow (\hat{X}, S) \longrightarrow X$ forms a Markov chain.
(3) $U \longrightarrow X \longrightarrow S$ forms a Markov chain.

All the marginal and conditional probabilities are generated from this joint probability distribution $P_{X,\hat{X},S,U}(\cdot,\cdot,\cdot,\cdot)$. This joint probability distribution with the above properties defines a pair of Gel'fand-Pinsker and Wyner-Ziv problems.

Using a greedy algorithm as presented in [129], we are able to construct deterministic binning scheme. For this deterministic binning scheme, each bin contains an $\left(N, 2^{NI(U;S)}\right)$ code from U to S and the codewords in all the bins form an $\left(N, 2^{NI(U;X)}\right)$ code. For SCRSI, this is a partition of source codebook for quantizing X, with distortion violation probability satisfying $P_{dv} \leq \epsilon_1$, into bins with codewords serving as a channel codebook, with error probability satisfying $P_e \leq \epsilon_2$. In CCTSI, this is a partition of a channel codebook, with error probability satisfying $P_e \leq \epsilon_1$, for the fictitious channel between U and X, into bins with codewords serving as a source codebook for quantizing S, with distortion violation probability satisfying $P_{dv} \leq 1 - \epsilon_2$. This deterministic binning scheme, allows a rate $I(U;X) - I(U;S)$ to be achievable for both the side information problems with an error probability, for the CCTSI problem, satisfying

$$P_e \leq 1 - \epsilon_2 + (\epsilon_1)^{\frac{1}{4}}\,, \tag{6.17}$$

and with a distortion violation probability, for the SCRSI problem, satisfying

$$P_{dv} \leq 1 - \epsilon_1 + (\epsilon_2)^{\frac{1}{4}} . \qquad (6.18)$$

Using Theorem 9, we can generalize the Gaussian SCRSI case considered by Wyner and Ziv (see Subsection 6.2). In Subsection 4.1, we have presented generalizations to Costa's result, where we have seen that there is no rate loss, relative to the case where both the encoder and decoder have access to the side information, even if the side information is not Gaussian. We can use Theorem 9 and prove that the more general dirty-paper problem is dual to the Wyner-Ziv problem. Therefore, the Wyner-Ziv result is generalized to the case where Z is Gaussian, but X and S are arbitrary.

Thus far, we have only considered the non-causal version of the duality between the SCRSI and the CCTSI problems. There is also a duality between the causal versions of the problems, i.e., the causal Wyner-Ziv problem presented in Subsection 6.2, and the causal Shannon's model presented in Subsection 3.1. If we look at both the causal Wyner-Ziv rate-distortion function (6.8) and Shannon's capacity formula (3.17), and rewrite (3.17) using the notation of the SCRSI problem, i.e.,

$$C = \max_{P_U(\cdot), f:\mathcal{U}\times\mathcal{S}\longrightarrow\hat{\mathcal{X}}} I(U;X), \qquad (6.19)$$

we can see the same term $I(U;X)$ in both (6.8) and (6.19). We remember that the difference between the causal and non-causal versions of the Wyner-Ziv problem, is the subtracted term $I(U;S)$, where both of the problems are minimized over the same set. In the channel coding problems, i.e., the Gel'fand-Pinsker and Shannon problems, the difference between the capacity formulas is again the term $I(U;S)$. However, there is a difference between the maximizations in the two problem. This is due to the fact that in the causal version, U and S are restricted to be independent.

Another similarity to the non-causal setting, is that feedforward does not improve the rate-distortion function in the causal/non-causal SCRSI, and feedback does not improve the capacity in the causal/non-causal CCTSI. The problem of a channel with a feedback is considered in Subsection 3.4.

In Subsection 7.2, we preset nested codes for practical implementation of the binning scheme for both the Gaussian Wyner-Ziv (in Subsection 7.2.4) and dirty-paper (is Subsection 7.2.3) problems. As it can be seen in these subsections, the same concepts and practical coding schemes (with some variations) are used for the two problems.

The duality in the Gaussian case has been explored in several papers [6], [95], [116], [117]. In these papers, a geometric interpretation for both the dirty-paper and the Wyner-Ziv problems was given. A sphere-packing versus sphere-covering interpretation of the duality has been formulated to illustrate the functional duality of these problems. In the dirty-paper problem, encoding has a sphere-covering interpretation, and decoding has a sphere-packing interpretation. In the Wyner-Ziv problem, it's the other way round. This interpretation emphasizes the fact that because there is a duality between these problems, the role of the encoder in one problem, match the role of the decoder in the other problem, and vice versa. A geometric derivation of the achievable rate-distortion for the Wyner-Ziv problem, and achievable rate for the dirty-paper problem, as given in [95], provides intuition for practical code constructions.

6.4 Joint Source-Channel Coding for the Wyner-Ziv Source and the Gel'fand-Pinsker Channel

In this subsection, we consider the problem of joint source channel coding for the Wyner-Ziv source and the Gel'fand-Pinsker channel. This combined model may be used to model the problem of a backward-compatible upgrading of an existing communication system [97], [85], which will be elaborated on, in this subsection.

Merhav and Shamai studied this combined model in [85], the model that was used to describe this problem is presented in Fig. 6.2. They assumed for this model, that the Wyner-Ziv source generates independent copies, $\{K_i, V_i\}_{i=1}^{\infty}$, of a pair of dependent, finite-alphabet RV's $(K, V) \in \mathcal{K} \times \mathcal{V}$, at a rate of ρ_s symbols pairs per second. Let $N = \rho_s T$, where T is the duration of the block in seconds. Similarly, the Gel'fand-Pinsker channel operates at the rate of ρ_c channel uses per second and $n = \rho_c T$. The channel input is subject to a cost constraint

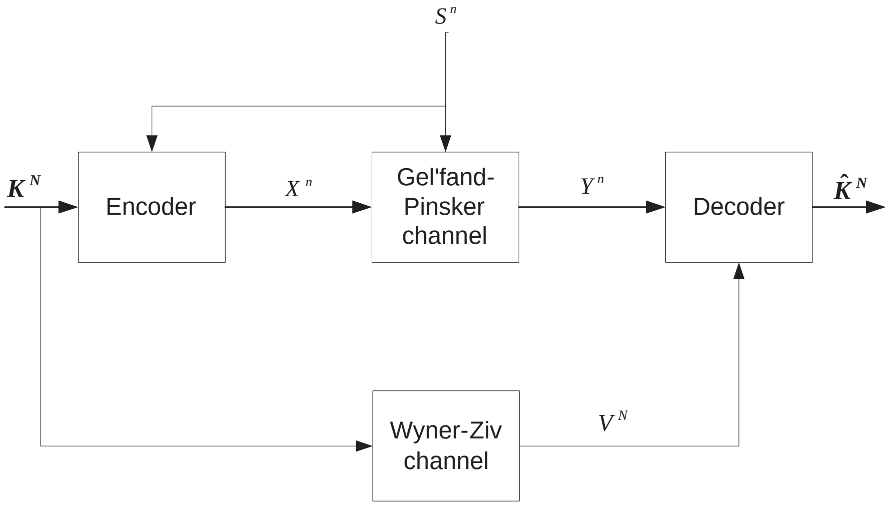

Fig. 6.2 Combined model of Wyner-Ziv and Gel'fand-Pinsker.

$E\{\sum_{i=1}^{n}\phi(X_i)\} \leq n\Gamma$, where ϕ is a given function from the set $\mathcal{X}$ to R^+ and $\Gamma \geq 0$ is a prescribed value.

The joint source-channel encoder implements a function $x^N = f(k^N, s^n)$, and the decoder is defined by a deterministic function $\hat{k}^N = g(v^N, y^n)$. The quality of the decoder output, $\hat{K}$, is judged with respect to the distortion measure $d(K^N, \hat{K}^N) = \sum_{i=1}^{N} d(K_i, \hat{K}_i)$. The conditional probability of V^N given K^N,

$$P_{V^N|K^N}(v^N|k^N) = \prod_{i=1}^{N} P_{V|K}(v_i|k_i), \tag{6.20}$$

is referred to as the Wyner-Ziv channel. It is assumed that $V^N \longrightarrow K^N \longrightarrow Y^N$ is a Markov chain, guaranteeing independence between the Wyner-Ziv channel and the Gel'fand-Pinsker channel.

Their main result is the following separation theorem for this model:

Theorem 10. Under the assumptions which are specified above, a necessary and sufficient condition for D being an achievable distortion level (of the Wyner-Ziv rate-distortion function) is

$$\rho_s R_{WZ}(D) \leq \rho_c C_{GP}(\Gamma).$$

Where

$$R_{WZ}(D) = \min[I(K;Z) - I(V;Z)]$$

where Z is an auxiliary RV and the minimum is under a D distortion constraint, and

$$C_{GP}(\Gamma) = \max[I(U;Y) - I(U;S)].$$

This separation theorem asserts that there is no loss in asymptotic optimality in applying first, an optimal Wyner-Ziv source code and then, an optimal Gel'fand-Pinsker channel code. This result was extended by Winshtok and Steinberg [134] for a Wyner-Ziv source, depending on an arbitrary varying state, known non-causally at the encoder, which is transmitted over an arbitrary varying Gel'fand-Pinsker channel. The separation principle presented above holds for this extended model.

As an example for the combined model presented here, we consider the problem of a backward-compatible upgrading of an existing communication system [85], [97]. An existing analog transmission system is digitally upgraded with the goal of optimizing (under a squared-error metric) the delivered analog and digital signal quality under a fixed transmission power constraint Γ. The proposed upgrading system does not increase the transmitted spectrum and is back-compatible with the existing analog receivers.

We consider the system depicted in Fig 6.3. The source K is an i.i.d. Gaussian source characterized by the probability density function $K \sim \mathcal{N}(0,\sigma^2)$, and Z is an AWGN distributed as $Z \sim \mathcal{N}(0,B)$. In order to enhance the performance for the overall system, the transmitted

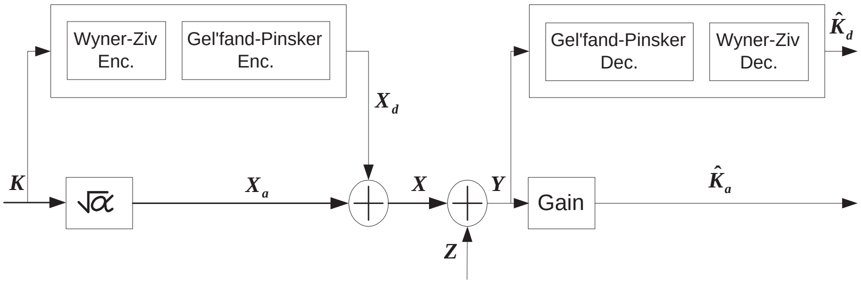

Fig. 6.3 A back-compatible digital upgrade system.

power Γ is split between "analog power", $\alpha\sigma^2$ $(0 \leq \alpha \leq \Gamma/\sigma^2)$ and "digital power" $\Gamma - \alpha\sigma^2$. Now, the "digital power" constitutes noise for the existing analog receiver. The achievable distortion between the source and the analog reconstruction $\hat{K}_a$ is $D_a = \frac{\sigma^2(\Gamma-\alpha\sigma^2+B)}{\Gamma+B}$, for an optimal gain factor at the decoder. The encoding procedure for the digital part motivated by Theorem 10, combines Wyner-Ziv source coding and Gel'fand-Pinsker channel coding. The Wyner-Ziv encoder encodes the source K treating $\hat{K}_a$ as side information with rate that can be communicated across Gel'fand-Pinsker channel which treats X_a as side information in which data is embedded (as in information embedding scheme) and Z as the channel noise. The Gel'fand-Pinsker capacity is given by $C_{GP} = 1/2\log_2(1 + \frac{\Gamma-\alpha\sigma^2}{B})$ [34] (there is a mistake in the Gel'fand-Pinsker capacity given in [97]). The Wyner-Ziv encoder encode the source at rate C_{GP} which result in a distortion $D_d = D_a 2^{-2C_{GP}} = \sigma^2 B/(\Gamma + B)$ between the source and the digital reconstruction $\hat{K}_d$[1]. The Gel'fand-Pinsker decoder decodes the bits generated by the Wyner-Ziv encoder and the Wyner-Ziv decoder uses this bits and the analog reconstruction $\hat{K}_a$ to reconstruct $\hat{K}_d$.

In this hybrid system we get an efficient digital upgrade of the existing analog system without compromising the digital quality compared to a completely digital system.

6.5 Channel Capacity and State Estimation for State-Dependent Channels

In this subsection, we consider the problem of achievable tradeoff between message and state information rates. In this problem, the sender has access to the channel state information and wishes to send both the message information and the state information across the channel. As an example to such a communication system, we consider the problem of writing to a memory with defective cells. If the defect information will be available to either or to both the encoder and the decoder, the memory utilization and it's reliability will increase. Therefore, if the encoder has access to the defect information, it will prove useful to si-

[1] A Gaussian source is Wyner-Ziv refinable without rate loss, for the digital part, we can treat it as a digital refinement of an equivalent source described by variance equal to D_a.

multaneously record information in the memory and deliver the defect information to the decoder. Another example, which was considered in Subsection 6.4, is the problem of a backward-compatible upgrading of an existing communication system. In this example, the analog signal serve as the channel state and the receiver generates an estimate of this analog signal as well as detecting the digital signal embedded in the analog signal.

Sutivong, Chiang, Cover and Kim [118], [119], [120] found an achievable tradeoff region between pure information rate and state estimation error for a discrete memoryless channel with an arbitrary state distortion measure, and characterized the optimal tradeoff for the binary channel and also to the Gaussian channel.

This problem also corresponds to the problem of joint source channel coding for the Wyner-Ziv source and the Gel'fand-Pinsker channel presented in Subsection 6.4. If we consider the case $\rho_c = \rho_s$, the combined channel $P_{Y,V|X,K,S}(y,v|x,k,s) = P_{Y|X,S}(y|x,s)P_{V|K}(v|k)$ can be thought of as a Gel'fand-Pinsker channel whose input is X, its state is (K,S), and its output is (Y,V). This case of the combined Wyner-Ziv and Gel'fand-Pinsker model, corresponds to the extreme case of tradeoff considered here, where the desired coding rate is $R = 0$ and the distortion measure of our new state (K,S) depends only on the component K.

The achievable tradeoff region for the discrete memoryless channel with arbitrary state distortion measure presented in Section 2 is given in Theorem 11 [120].

Theorem 11. For a discrete memoryless channel $P_{Y|X,S}(y|x,s)$ with state $s^N \sim \prod_{i=1}^{N} P_S(s_i)$ non-causally known at the transmitter and a state distortion measure $d(s,\hat{s})$, an achievable (R,D) tradeoff region is the closure of the convex hull of the set of all (R,D) pairs, $R \geq 0$, satisfying

$$R \leq I(U;Y) - I(U;S) - I(V;S|U) + I(V;Y|U)$$

$$D \geq E\{d(s,\hat{s})\}$$

for some distribution

$$P_{U|S}(u|s)P_{V|U,S}(v|u,s)P_{X|U,S}(x|u,s)P_{\hat{S}|V,U,S}(\hat{s}|v,u,s),$$

where V and U are auxiliary random variables.

This achievable rate region can be rewritten in a more compact way, by noticing that

$$I(U;Y) - I(U;S) - I(V;S|U) + I(V;Y|U) = I(U,V;Y) - I(U,V;S).$$

Next, we can combine (U,V) to one auxiliary RV by defining $W = (U,V)$. Therefore, the achievable rate region is given by

$$\begin{aligned} R &\leq I(W;Y) - I(W;S) \\ D &\geq E\{d(S,\hat{S})\}, \end{aligned} \tag{6.21}$$

which is the Gel'fand-Pinsker capacity formula with a distortion constraint between S to $\hat{S}$, instead of a power constraint.

The capacity region for this general scenario is still an open problem. The proposed achievable tradeoff region can be shown to be optimal for the additive Gaussian channel model with mean-squared distortion constraint $\frac{1}{N}\sum_{i=1}^{N} E(S_i - \hat{S}_i)^2 \leq D$ presented in Subsection 4.1. The optimal tradeoff region for this channel model is given in Theorem 12 [118].

Theorem 12. For the state-dependent additive Gaussian channel $Y^N = X^N + S^N + Z^N$, the optimal (R,D) tradeoff region is given by the closure of the convex hull of all (R,D) pairs satisfying

$$R \leq \frac{1}{2}\log\left(1 + \frac{\gamma\Gamma}{B}\right) \tag{6.22}$$

$$D \geq Q\frac{(\gamma\Gamma + B)}{\left(\sqrt{Q} + \sqrt{(1-\gamma)\Gamma}\right)^2 + \gamma\Gamma + B} \tag{6.23}$$

for some $0 \leq \gamma \leq 1$.

The variable γ is the power allocation parameter. By varying $0 \leq \gamma \leq 1$, we can tradeoff between the pure information rate R and the mean-squared estimation error D. The end points of this optimal tradeoff region are given by

$$(R,D) = \left(0, Q\frac{B}{\left(\sqrt{Q} + \sqrt{\Gamma}\right)^2 + B}\right)$$

for the case where the transmitter send only the state information which corresponds to the case where $\gamma = 0$, and

$$(R, D) = \left(\frac{1}{2}\log\left(1 + \frac{\Gamma}{B}\right), Q\frac{\Gamma + B}{Q + \Gamma + B}\right)$$

for the case where the transmitter send only pure information which correspond to the case where $\gamma = 1$.

In [86] a different aspect of the problem was considered. In [86], the state sequence is considered as undesired information transferred to the receiver. This undesired state information could be, for example, a secret analog information that should be concealed from the receiver, or a codeword belonging to another user that should be concealed from other users. Here, instead of transmitting both pure information and state information, the transmitter send to the receiver pure information and mask the state information in order to conceal it from the receiver. The amount of information that the receiver retrieves about the state sequence is measured by blockwise mutual information (equivocation). We would like to find the achievable tradeoff between reliable coding at rate R, which we would like to keep as large as possible, and an equivocation level, $\frac{1}{N}I(S^N; Y^N)$, which we would like to make smaller than some prescribed equivocation constraint $\frac{1}{N}I(S^N; Y^N) \leq E$.

The achievable tradeoff region for the discrete memoryless channel is given in Theorem 13 [86].

Theorem 13. The achievable region $\mathcal{A}$ is the set of pairs $\{(R, E)\}$ for which there exists a random variable U that satisfies the following conditions at the same time:

(1) $U \longrightarrow (X, S) \longrightarrow Y$ is a Markov chain.
(2) $E[\phi(x)] \leq \Gamma$.
(3) $R \leq I(U; Y) - I(U; S)$.
(4) $E \geq I(S; U, Y)$.

There are a few differences between this result and the Gel'fand-Pinsker capacity result (3.14). First, the cardinality of the auxiliary RV U is by two letters larger than in the ordinary Gel'fand-Pinsker coding because of the additional equivocation and power constraints. Second, the channel $P_{X|U,S}(x|u,s)$ is not necessarily deterministic as in the Gel'fand-Pinsker problem.

We can also consider the problem of the achievable tradeoff region for the causal channel. The causal achievable region, which is the set of all achievable pairs $\{(R,E)\}$, is the same as for the non-causal setting, with the additional constraint that U is independent of S. Therefore, $I(U;S) = 0$ in the rate inequality, and $I(S;Y,U) = I(S;Y|U)$ in the equivocation inequality, which are given in Theorem 13. This result is not surprising, because the causal setting considered by Shannon, is a special case of the non-causal setting considered by Gel'fand and Pinsker, as was argued in Subsection 3.3.

Theorem 13, may be used in order to find the achievable region for the additive Gaussian channel presented in Subsection 4.1. For this additive Gaussian channel with $E[XS] = \rho\sqrt{Q\Gamma}$, where ρ is the correlation coefficient between X S, we denote by $\hat{S}$, the optimal linear estimation of S based on Y. The variance of this estimator is given by

$$\hat{Q} = \frac{\left(Q + \rho\sqrt{Q\Gamma}\right)^2}{Q + 2\rho\sqrt{Q\Gamma} + \Gamma + B}. \tag{6.24}$$

The achievable rate, from Theorem 13, is bounded by

$$\begin{aligned} R &\leq I(U;Y) - I(U;S) \\ &\leq \frac{1}{2}\log\left(1 + \frac{\Gamma(1-\rho^2)}{B}\right). \end{aligned} \tag{6.25}$$

Next, we find the minimum achievable per-symbol masking mutual information for a given rate R. From Theorem 13, E is bounded by

$$\begin{aligned} E &\geq I(S;Y,U) \\ &\geq I(S;Y) \\ &\geq \frac{1}{2}\log\left(\frac{Q}{Q-\hat{Q}}\right). \end{aligned} \tag{6.26}$$

Let

$$R < \log\left(1 + \frac{\Gamma}{B}\right), \tag{6.27}$$

for a rate R, satisfying (6.27), the correlation coefficient as a function of R (by (6.25)) is given by

$$\varrho(R) = \sqrt{1 - (2^{2R} - 1)\,\frac{B}{\Gamma}}. \tag{6.28}$$

The minimum achievable per-symbol masking mutual information for a given rate R, satisfying (6.27), is given by

$$E = \min_{\varrho(R) > 0} \frac{1}{2} \log \frac{Q}{Q - \hat{Q}}, \tag{6.29}$$

as a function of ρ across the interval $\rho \in [-\varrho(R), +\varrho(R)]$, where $\hat{Q}$ is given in (6.24).

We note that the same choice of U simultaneously maximizes $I(U;Y) - I(U;S)$ and minimizes $I(S;Y,U)$. Let $\widetilde{X} \triangleq X - aS$, where aS stands for the best linear estimation of X given S and $a = \rho\sqrt{\Gamma}/\sqrt{Q}$. Thus, $\widetilde{X}$ is uncorrelated with S, $E[\widetilde{X}^2] = \Gamma(1 - \rho^2)$ and the channel output is now given by

$$Y = \widetilde{X} + (1 + a)S + Z. \tag{6.30}$$

Similarly to Costa we chose $U = \widetilde{X} + c(1 + a)S$, where $c = \frac{\Gamma(1-\rho^2)}{\Gamma(1-\rho^2)+B}$. Therefore, (6.25) is achievable. With this choice of U, it can be shown that the optimum linear estimation of S given Y and U, is independent of U, thus $I(S;Y,U) = I(S;Y)$.

We can see that in the case of weak interference, where a small part of the power is used to cancel the state and the reminder of the power is used to convey information, that the achievable rate is $R = \frac{1}{2}\log\left(1 + \frac{\Gamma - Q}{B}\right)$. We also note that if we transmit pure information at rate $R > 0$, $P_{X|U,S}(\cdot|\cdot,\cdot)$ is deterministic, i.e., $X = U - (c(1 + a) - a)\,S$ with c and a as given above.

6.6 Multiple User Channels with Side Information

6.6.1 The Multiple Access Channel with Side Information

We consider the multiple access channel (MAC) controlled by a state process with side information available to all the encoders, to one encoder and/or to the receiver. The MAC model is a very important model in mobile wireless communication, where several users are transmitting to the base station and due to the mobility of the users, the transmitted signals, which suffers from multipath, shadowing and propagation losses, are degraded. Therefore, we consider for this problem a model of a MAC controlled by a state process.

The capacity for a class of time varying MAC, when varying degrees of causal side information concerning the condition of the channel are provided to the transmitters and the receiver, was determined by Das and Narayan in [40]. Das and Narayan, studied the most general model, i.e., general channel statistics and general state process, which are not necessarily stationary or ergodic. For this general model, they have found the capacity region which is expressed as a union of a region characterized by limits of mutual information. These capacity expressions, as noted by Das and Narayan, do not lead to any useful insights, because in most of the cases they cannot be significantly simplified and be given as a single-letter expressions. However, there is one case for which these expressions can be simplified and be given as a single-letter formula. This case is considered here.

We consider the case of a memoryless two-user channel given by

$$P_{Y^N|X^N(1),X^N(2),S^N}(y^N|x^N(1),x^N(2),s^N) = \prod_{n=1}^{N} P_{Y|X(1),X(2),S}(y_n|x_n(1),x_n(2),s_n), \tag{6.31}$$

where $X(1) \in \mathcal{X}(1)$, $X(2) \in \mathcal{X}(2)$, $Y \in \mathcal{Y}$ and $S \in \mathcal{S}$ are the first user input, the second user input, the output and state, respectively. This channel is depicted in Fig. 6.4. We assume that the state process is stationary and ergodic, and that the receiver has perfect causal side information (the switch in Fig. 6.4 is closed). Let $E(1) \in \mathcal{E}(1)$, $E(2) \in \mathcal{E}(2)$ be the side information signals available to the encoders, and let

$$J_{e,1} : \mathcal{S} \longrightarrow \mathcal{E}(1) \quad J_{e,2} : \mathcal{S} \longrightarrow \mathcal{E}(2)$$

be mappings which are used to describe the causal side information available to the two senders. Therefore, in our setting, only partial casual side information is available to the encoders. We also restrict the encoders to use only strategies which depend on a k-window of the respective side information, i.e., S^k, where k is a fixed positive integer.

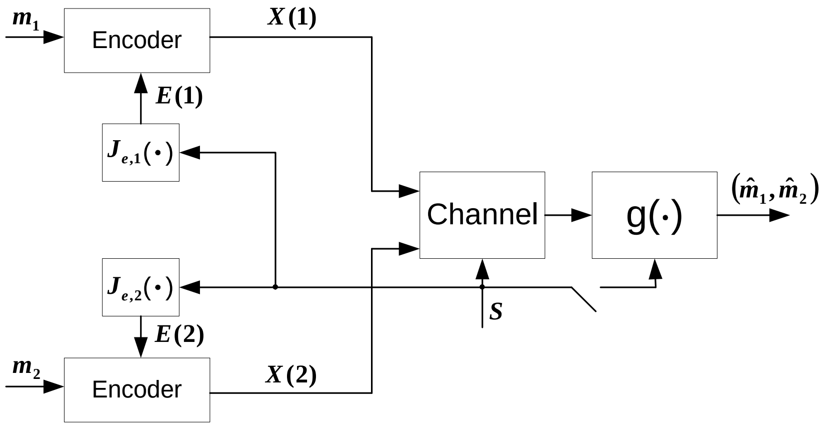

Fig. 6.4 A two-user Mac with states.

An $(N, 2^{NR_1}, 2^{NR_2})$ code for this channel consists of the following:

(1) A set of N encoding functions $f_n(i) : \{1, ..., 2^{NR_i}\} \times \mathcal{E}^n(i) \longrightarrow \mathcal{X}$.
(2) A decoding function $g : \mathcal{Y}^N \times \mathcal{S}^N \longrightarrow \{1, ..., 2^{NR_1}\} \times \{1, ..., 2^{NR_2}\}$

where i, $i = 1, 2$ is the user index.

The average probability of error is given by

$$P_e = \frac{1}{2^{N(R_1+R_2)}} \sum_{m_1=1}^{2^{NR_1}} \sum_{m_2=1}^{2^{NR_2}} \sum_{s^N \in \mathcal{S}^N} P_{S^N}(s^N) \\ \times \textstyle\sum_{y^N : g(y^N, s^N) \neq (m_1, m_2)} P_{Y^N|X^N(1), X^N(2), S^N}(y^N | x^N(1), x^N(2), s^N), \tag{6.32}$$

where $x^N(1)$ and $x^N(2)$ depend on the inputs to the encoders, as described above.

A rate pair (R_1, R_2) is achievable if there exists a sequence of $(N, 2^{NR_1}, 2^{NR_2})$ codes with $P_e \longrightarrow 0$ as $N \longrightarrow 0$.

The capacity of this MAC is the convex closure of the region

$$\bigcup_{\substack{X(1)|E^k(1),X(2)|E^k(2)\\ S^k \to (E^k(1),E^k(2)) \to (X(1),X(2))}} \left\{ (R_1,R_2): \begin{array}{c} 0 \le R_1 \le I(X(1);Y|X(2),S^k), \\ 0 \le R_2 \le I(X(2);Y|X(1),S^k), \\ R_1+R_2 \le I(X(1)X(2);Y|S^k) \end{array} \right\} \tag{6.33}$$

where the joint probability distribution function of $X(1)$, $X(2)$, S^k and Y is given by

$$\begin{aligned} P_{X(1),X(2),S^k,Y}(x(1),x(2),s^k,y) = \\ P_{S^k}(s^k)P_{Y|X(1),X(2),S}(y|x(1),x(2),s) \\ \times \prod_{i=1}^{2} \left[\sum_{e(i)^k \in \mathcal{E}^k(i)} \delta\left(J_{e,i}(s^k),e^k(i)\right) P_{X(i)|E^k(i)}(x(i)|e^k(i))\right], \end{aligned} \tag{6.34}$$

where i, $i = 1,2$, is the user index. If the state process is memoryless, then there is no need to restrict the coding strategies to a finite k-window.

A special case of this model, was considered in [106], where perfect causal side information is available to the encoders, and no side information is available to the receiver (the switch in Fig. 6.4 is open). An outer bound for the capacity region is the convex closure of the region

$$\bigcup_{\substack{X(1)|E(1),X(2)|E(2)\\ S \to (E(1),E(2)) \to (X(1),X(2))}} \left\{ \begin{array}{c} (R_1,R_2): \\ 0 \le R_1 \le I(X(1);Y|X(2)), \\ 0 \le R_2 \le I(X(2);Y|X(1)), \\ R_1+R_2 \le I(X(1)X(2);Y) \end{array} \right\}. \tag{6.35}$$

If we restrict $X(1)|E(1)$ to be independent of $X(2)|E(2)$, (6.35) is also an inner bound to the capacity region.

Next, we consider a variation of this model, where the state process is memoryless, with perfect non-causal side information for one of the users, no side information for the other user, and no side information for the decoder. This model was studied in [72], and it applies to the mobile wireless communication scenario, discussed in the beginning of this subsection. We shall also consider a Gaussian version of this problem, and see that both users benefit in terms of achievable rate from availability of the side information to only one of the users.

An achievable rate region for this discrete memoryless model is characterized in Theorem 14 [72].

Theorem 14. An achievable rate region for a discrete memoryless MAC with state information $S^N \sim P_{S^N}(s^N) = \prod_{n=1}^{N} P_S(s_n)$, which is available non-causally at only encoder 1, is the closure of the convex hull of all pairs (R_1, R_2) satisfying

$$\begin{aligned} R_1 &< I(U;Y|X(2)) - I(U;S), \\ R_2 &< I(X(2);Y|U), \\ R_1 + R_2 &< I(U,X(2);Y) - I(U;S) \end{aligned} \tag{6.36}$$

where U is a finite alphabet auxiliary RV, and all the mutual informations in the above equations are calculated using joint distribution of the form

$$\begin{aligned} &P_{S,U,X(1),X(2),Y}(s,u,x(1),x(2),y) = \\ &\quad P_S(s)P_{U,X(1)|S}(u,x(1)|s)P_{X(2)}(x(2))P_{Y|X(1),X(2),S}(y|x(1),x(2),s). \end{aligned} \tag{6.37}$$

We can use Theorem 14 to characterize the achievable rate region of the Gaussian MAC. The output of the channel is given by

$$Y^N = X^N(1) + X^N(2) + S^N + Z^N, \tag{6.38}$$

where Z^N is an additive Gaussian noise, with probability distribution $Z \sim \mathcal{N}(0,B)$, S^N is the channel state or interference, with probability distribution $S \sim \mathcal{N}(0,Q)$, and $X^N(1)$, $X^N(2)$ are the channel inputs, which satisfy power constraints $\frac{1}{N}\sum_{n=1}^{N} X_n^2(i) \leq \Gamma_i$, where i, $i = 1,2$ is the user index. The achievable rate region for this channel is characterized in Theorem 15 [72].

Theorem 15. An achievable rate region for the above Gaussian MAC with state information known at encoder 1 is given by

$$\bigcup_{\beta} \bigcup_{\alpha \in \mathcal{A}} \{(R_1, R_2) : R_1 < r_1(\beta,\alpha), R_2 < r_2(\beta,\alpha), R_1 + R_2 < r_3(\beta,\alpha)\} \tag{6.39}$$

where

$$\begin{aligned} r_1(\beta,\alpha) &= \\ &\tfrac{1}{2}\ln\left(\tfrac{\overline{\beta}\Gamma_1[\overline{\beta}\Gamma_1+(\sqrt{Q}-\sqrt{\beta\Gamma_1})^2+B]}{\overline{\beta}\Gamma_1(\sqrt{Q}-\sqrt{\beta\Gamma_1})^2(1-\alpha)^2+B(\overline{\beta}\Gamma_1+\alpha^2(\sqrt{Q}-\sqrt{\beta\Gamma_1})^2)}\right) \\ r_2(\beta,\alpha) &= \\ &\tfrac{1}{2}\ln\left(1+\tfrac{\Gamma_2[\overline{\beta}\Gamma_1+\alpha^2(\sqrt{Q}-\sqrt{\beta\Gamma_1})^2]}{\overline{\beta}\Gamma_1(\sqrt{Q}-\sqrt{\beta\Gamma_1})^2(1-\alpha)^2+B(\overline{\beta}\Gamma_1+\alpha^2(\sqrt{Q}-\sqrt{\beta\Gamma_1})^2)}\right) \\ r_3(\beta,\alpha) &= \\ &\tfrac{1}{2}\ln\left(\tfrac{\overline{\beta}\Gamma_1[\overline{\beta}\Gamma_1+\Gamma_2+(\sqrt{Q}-\sqrt{\beta\Gamma_1})^2+B]}{\overline{\beta}\Gamma_1(\sqrt{Q}-\sqrt{\beta\Gamma_1})^2(1-\alpha)^2+B(\overline{\beta}\Gamma_1+\alpha^2(\sqrt{Q}-\sqrt{\beta\Gamma_1})^2)}\right) \end{aligned} \tag{6.40}$$

and $0 \leq \beta < \min(1, \frac{Q}{\Gamma_1})$, $\overline{\beta} = (1-\beta)$, $\mathcal{A} = \{x : x \in \mathbb{R}, r_1(\beta,x) \geq 0, r_2(\beta,x) \geq 0, r_3(\beta,x) \geq 0\}$.

A coding scheme which achieves the above rate region, divides the power available to transmitter 1, between state cancelation scheme and pure information transmission, with β being the power sharing parameter, i.e., a power of $\beta\Gamma_1$ is used for state canceling, and $(1-\beta)\Gamma_1$ is used for pure information transmission using a dirty-paper scheme. The second user, now sees a channel with a weaker interference $\widetilde{S} \sim \mathcal{N}(0, (\sqrt{Q} - \sqrt{\beta\Gamma_1})^2)$, and uses all its power Γ_2 to overcome both this interference and the additive noise.

From the above discussion, comes our main conclusion, that even though side information is available to only one user, both of the users benefit from it.

We can also consider another Gaussian MAC where both of the users have non-causal side information S^N. In this case, explored in [57], [71], it turns out that the capacity region of this channel is equal to the capacity region when S^N is available at the transmitters and the receiver, which is, in turn, equal to the capacity region of the standard

Gaussian MAC without additive interference S^N. This result, is an extension of Costa's result to a Gaussian MAC.

The models and result presented in this subsection can be extended to the case where there are more that two users.

6.6.2 The Broadcast Channel with Side Information

In this subsection, we consider an extension to the broadcast channel model presented in Subsection 5.1, to a broadcast channel controlled by a state process with side information available to the encoder. Gel'fand and Pinsker [57] presented in 1984 a two-user extension of Costa's single-user dirty-paper model. They have shown that the capacity of a two-user Gaussian MAC or broadcast channel with states is equal to the capacity of a MAC or broadcast channel without the states respectively.

The broadcast channel with states, is the complementary model to the MAC model, which was also controlled by a state process, presented in Subsection 6.6.1. The model of a broadcast channel with states may describe a communication scenario in which a base station sends independent information to multiple non-cooperating mobile users. Due to the mobility of the users, the transmitted signal, which suffers from multipath, shadowing and propagation losses, is degraded. This example, is the reverse link of the communication scenario, considered as an example for the MAC model in Subsection 6.6.1.

A memoryless two-user broadcast channel with states, depicted in Fig. 6.5, is given by

$$\begin{aligned} &P_{Y^N(1),Y^N(2)|S^N,X^N}(y^N(1),y^N(2)|s^N,x^N) = \\ &\quad \textstyle\prod_{n=1}^N P_{Y(1),Y(2)|S,X}(y_n(1),y_n(2)|s_n,x_n), \end{aligned} \tag{6.41}$$

where $Y(1) \in \mathcal{Y}(1)$, $Y(2) \in \mathcal{Y}(2)$, $X \in \mathcal{X}$ and $S \in \mathcal{S}$ are the first user output, the second user output, the input and state, respectively.

We denote the marginals of the channel by $P_{Y(1)|S,X}(\cdot|s,x)$ and $P_{Y(2)|S,X}(\cdot|s,x)$. A broadcast channel $P_{Y(1),Y(2)|S,X}(\cdot,\cdot|s,x)$ is said to be physically degraded if we can write

$$\begin{aligned} &P_{Y(1),Y(2)|S,X}(y(1),y(2)|s,x) = \\ &\quad P_{Y(1)|S,X}(y(1)|s,x)P_{Y(2)|Y(1)}(y(2)|y(1)) \end{aligned} \tag{6.42}$$

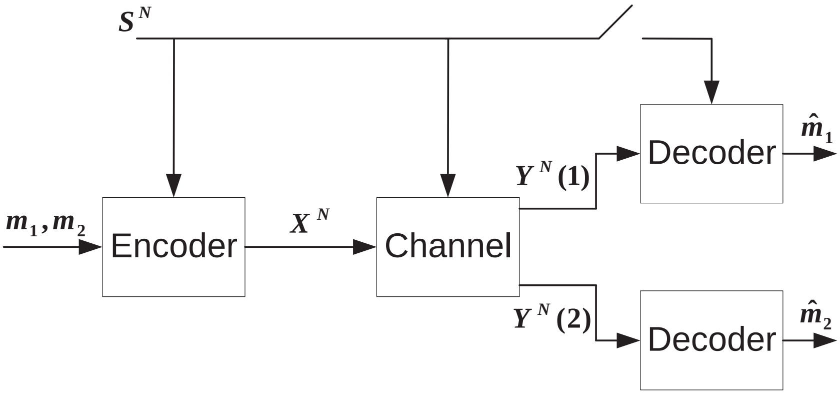

Fig. 6.5 A broadcast channel with states.

for some probability distribution $P_{Y(2)|Y(1)}(\cdot|y(1))$. From the definition of a physically degraded broadcast channel, it can be seen that the state process controls only the non-degraded channel, i.e., $P_{Y(1)|S,X}(\cdot|s,x)$, whereas the degraded channel, i.e., $P_{Y(2)|Y(1)}(\cdot|y(1))$, is independent of the state process.

We also define a stochastically degraded broadcast channel. A broadcast channel
$P_{Y(1),Y(2)|S,X}(\cdot,\cdot|s,x)$ is said to be stochastically degraded if there exists some probability distribution $\widetilde{P}_{Y(2)|Y(1)}(\cdot|y(1))$ such that

$$P_{Y(2)|S,X}(y(2)|s,x) = \sum_{y\in\mathcal{Y}} P_{Y(1)|S,X}(y(1)|s,x)\widetilde{P}_{Y(2)|Y(1)}(y(2)|y(1)). \tag{6.43}$$

First, we consider a broadcast channel with non-causal side information available to the encoder, and no side information is available to the decoders (the switch in Fig 6.5 is open).

An $\left(N, 2^{NR_1}, 2^{NR_2}\right)$ non-causal code for this channel consists of the following:

(1) An encoding function

$$f : \{1,...,2^{NR_1}\} \times \{1,...,2^{NR_2}\} \times \mathcal{S}^N \longrightarrow \mathcal{X}^N. \tag{6.44}$$

(2) A pair of decoding functions

$$g_i : \mathcal{Y}^N(i) \longrightarrow \{1, ..., 2^{NR_i}\}, \tag{6.45}$$

where i, $i = 1, 2$, is the user index.

The average probability error, in the non-causal case, is given by

$$P_e = \frac{1}{2^{N(R_1+R_2)}} \sum_{m_1=1}^{2^{NR_1}} \sum_{m_2=1}^{2^{NR_2}} \sum_{s^N \in \mathcal{S}^N} \sum_{(y^N(1), y^N(2)) \notin \mathcal{A}_{m_1,m_2}} \\ \times P_{S^N}(s^N) P_{Y^N(1), Y^N(2)|X^N, S^N}(y^N(1), y^N(2)|x^N, s^N), \tag{6.46}$$

where $\mathcal{A}_{m_1,m_2} \triangleq \{(y^N(1), y^N(2)) : g_1(y^N(1)) = m_1, g_2(y^N(2)) = m_2\}$, $\mathcal{A}_{m_1,m_2} \subseteq \mathcal{Y}^N(1) \times \mathcal{Y}^N(2)$, is the correctly decoded outputs set, and where x^N depends on the input to the encoder as described above.

A rate pair (R_1, R_2) is said to be achievable if there exists a sequence of $(N, 2^{NR_1}, 2^{NR_2})$ codes with $P_e \longrightarrow 0$ as $N \longrightarrow 0$. The closure of all achievable rate pairs is the capacity region.

An $(N, 2^{NR_1}, 2^{NR_2})$ causal code for this channel consists of the following:

(1) A set of N encoding functions

$$f_n : \{1, ..., 2^{NR_1}\} \times \{1, ..., 2^{NR_2}\} \times \mathcal{S}^n \longrightarrow \mathcal{X}. \tag{6.47}$$

(2) A pair of decoding functions as (6.45).

The average probability error, for the causal case, is given by (6.46), where x^N depends on the input to the encoder as described above.

We start by giving an inner and outer bounds for the capacity region of the non-causal case, and then we turn our attention to the causal case. The bound for the two cases were given by Steinberg in [110].

Let $\mathcal{P}$ stands for the collection of random variables $(\widetilde{K}, S, X, Y(1), Y(2))$, where $\widetilde{K}$ take values in the finite alphabet $\widetilde{\mathcal{K}}$. The joint probability distribution of these random variables satisfy

$$P_{\widetilde{K},S,X,Y(1),Y(2)}(\widetilde{k}, s, x, y(1), y(2)) = \\ P_{\widetilde{K},S,X}(\widetilde{k}, s, x) P_{Y(1),Y(2)|X,S}(y(1), y(2)|x, s), \tag{6.48}$$

and

$$\sum_{\widetilde{k},x} P_{\widetilde{K},S,X}(\widetilde{k},s,x) = P_S(s). \tag{6.49}$$

Therefore, the following

$$\widetilde{K} \longrightarrow (S,X) \longrightarrow (Y(1),Y(2)) \tag{6.50}$$

is a Markov chain.

An inner bound to the capacity region of any discrete memoryless degraded broadcast channel, with states and non-causal side information is the set of all pairs (R_1, R_2) such that

$$\begin{aligned} R_1 &\leq I(U;Y(1)|K) - I(U;S|K) \\ R_2 &\leq I(K;Y(2)) - I(K;S), \end{aligned} \tag{6.51}$$

for some $((K,U),X,S,Y(1),Y(2)) \in \mathcal{P}$, where K and U are auxiliary random variables which take values in the finite alphabets $\mathcal{K}$ and $\mathcal{U}$, respectively. The alphabet cardinality of these auxiliary random variables satisfy

$$\begin{aligned} |\mathcal{K}| &\leq |\mathcal{S}||\mathcal{X}| + 1, \\ |\mathcal{U}| &\leq |\mathcal{S}||\mathcal{X}|\left(|\mathcal{S}||\mathcal{X}| + 1\right). \end{aligned} \tag{6.52}$$

In order to exhaust this inner bound, it is sufficient to take X to be a deterministic function of the triple (K,U,S). The proof of this inner bound is given in [110], and is based on the code construction for the degraded broadcast channel [36] and the binning construction discussed in this paper.

An outer bound to the capacity region of any discrete memoryless degraded broadcast channel, with states and non-causal side information is the set of all pairs (R_1, R_2) such that

$$\begin{aligned} R_1 &\leq I(U;Y(1)|K,V) - I(U;S|K,V), \\ R_2 &\leq I(K;Y(2)) - I(K;S), \\ R_1 + R_2 &\leq I(K,V,U|Y(1)) - I(K,V,U|S), \end{aligned} \tag{6.53}$$

for some $((K,V,U),S,X,Y(1),Y(2)) \in \mathcal{P}$, where K, V and U are auxiliary random variables which take values in the finite alphabets

$\mathcal{K}$, $\mathcal{V}$ and $\mathcal{U}$, respectively. The alphabet cardinality of these auxiliary random variables satisfy

$$\begin{aligned}
&|\mathcal{K}| \leq |\mathcal{S}||\mathcal{X}| + 2,\\
&|\mathcal{V}| \leq |\mathcal{S}||\mathcal{X}|\left(|\mathcal{S}||\mathcal{X}| + 2\right) + 1,\\
&|\mathcal{U}| \leq |\left(\mathcal{S}||\mathcal{X}|\left(|\mathcal{S}||\mathcal{X}| + 2\right) + 1\right)|\mathcal{S}||\mathcal{X}|\left(|\mathcal{S}||\mathcal{X}| + 2\right) + 1.
\end{aligned} \tag{6.54}$$

In Subsection 3.4, we have presented Cover and Chiang's extension to the Gel'fand-Pinsker model, where side information was also available to the receiver, and we have seen that this case follows directly form Gel'fand-Pinsker model by incorporating the receiver side information as part of the corresponding output. Therefore, a broadcast channel with side information available to the transmitter and to the first user, or to the second user, or to both, is a special case of our model.

For the case where only the first user has side information (the switch in Fig. 6.5 is closed), it turns out that the inner bound, given in (6.51), is tight and gives the capacity region. We will not consider the case where only the second user has side information, because it conflicts with the physically degraded broadcast channel definition in (6.42).

The capacity region of any discrete memoryless degraded broadcast channel, with states and non-causal side information at the encoder, and at the first user, is the set of all pairs (R_1, R_2) such that

$$\begin{aligned}
R_1 &\leq I(X;Y(1)|K,S),\\
R_2 &\leq I(K;Y(2)) - I(K;S),
\end{aligned} \tag{6.55}$$

for some $(K,S,X,Y(1),Y(2)) \in \mathcal{P}$, where K is an auxiliary random variable which take values in the finite alphabet $\mathcal{K}$. The alphabet cardinality of this auxiliary random variable satisfy

$$|\mathcal{K}| \leq |\mathcal{S}||\mathcal{X}| + 1. \tag{6.56}$$

An example to a possible application of this model that was given in [110], is a watermarking system where the transmitter encodes watermarks to both the first user with the side information, and the second user without the side information. In this case, both of the users receive the same channel output, and one of them has access to the side information, whereas the other does not. The user which has access to the side information, say user $Y(1)$, is called the private user, and the user without the side information, user $Y(2)$, is called the public user. In our model the channel from $Y(1)$ to $Y(2)$, $P_{Y(2)|Y(1)}(\cdot|\cdot)$, is an identity channel. Another scenario in which the discrete memoryless degraded broadcast channel model with states and non-causal side information applies is a watermarking problem, where the encoded message is supposed to pass several stages of attacks, thus resulting in a degraded channel model.

The second example is connected to the problem of reversible information embedding with several stages of attack discussed in Section 5.2. In [73], Kotagiri and Leneman have found the reversible information embedding capacity region for the two-user degraded broadcast channel with non-causal CSIT. In this model, the encoder in (6.44) is subjected to the distortion constraint given by

$$E\left[d\left(s^N, f\left(s^N, m_1, m_2\right)\right)\right] \leq D, \tag{6.57}$$

where $d : \mathcal{S}^N \times \mathcal{X}^N \longrightarrow \mathbb{R}_+$ is the distortion measure between S^N and X^N. The decoders produce both message estimates and host sequence estimates, i.e.,

$$g_i : \mathcal{Y}^N(i) \longrightarrow \left(\mathcal{S}^N, \{1, ..., 2^{NR_i}\}\right), \tag{6.58}$$

where i, $i = 1, 2$, is the user index. The reversible information embedding capacity region of the degraded broadcast channel is the closure of rate pairs (R_1, R_2) satisfying

$$\begin{aligned} R_1 &< I(X; Y(1)|K, S) \\ R_2 &< I(K, S; Y(2)) - H(S) \end{aligned} \tag{6.59}$$

for the set of conditional distributions $\left\{P_{K,X|S}(\cdot, \cdot|\cdot)\right\}$ satisfying the distortion constraint (6.57), where K is an auxiliary random variable which take values in the finite alphabet $\mathcal{K}$.

Due to the requirement to produce an estimate of the host sequence at the decoder, in cases where the host entropy is larger than the channel capacity, no communication can take place under a complete reconstruction requirement. In [113], Steinberg suggested as a possible solution to this problem to provide the decoders, a priori, with a compressed version of the host. Steinberg considered the problem of reversible information embedding for the degraded broadcast channel where a compressed host data is available, before transmission, at the decoders. The model for this problem is depicted in Fig. 6.6. The compressed host data V is generated by the state encoder which is given by

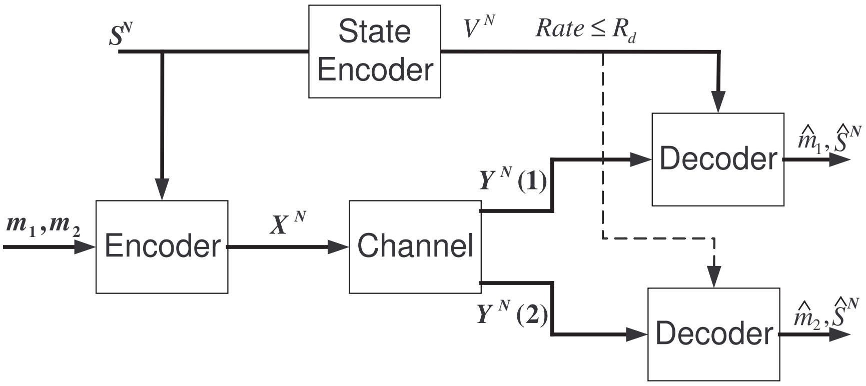

Fig. 6.6 Reversible information embedding with compressed host at the decoders.

$$f_s : \mathcal{S}^N \longrightarrow \mathcal{V}, \tag{6.60}$$

where $\mathcal{V} = \{1, ..., 2^{NR_d}\}$. The decoders which have access to the compressed host data are given by

$$g_i : \mathcal{Y}^N(i) \times \mathcal{V}^N \longrightarrow \left(\mathcal{S}^N, \{1, ..., 2^{NR_i}\}\right), \tag{6.61}$$

where i, $i = 1, 2$, is the user index. The probability of error in decoding the messages and reproducing the state S are given by

$$P_e = \frac{1}{2^{N(R_1+R_2)}} \sum_{m_1=1}^{2^{NR_1}} \sum_{m_2=1}^{2^{NR_2}} \sum_{(y^N(1),y^N(2),s^N) \notin \mathcal{A}_{m_1,m_2,s}} \times P_{S^N}(s^N) P_{Y^N(1),Y^N(2)|X^N,S^N}(y^N(1), y^N(2)|x^N, s^N), \tag{6.62}$$

where

$$\mathcal{A}_{m_1,m_2,s} \triangleq \left\{ (y^N(1), y^N(2), s^N) : \begin{array}{c} g_1(f_s(s^N), y^N(1)) = (s^N, m_1) \\ g_2(f_s(s^N), y^N(2)) = (s^N, m_2) \end{array} \right\}, \quad (6.63)$$

$\mathcal{A}_{m_1,m_2,s} \subseteq \mathcal{S}^N \times \mathcal{Y}^N(1) \times \mathcal{Y}^N(2)$, is the correctly decoded outputs and state set, and where x^N depends on the input to the encoder as described above.

The reversible information embedding rate-distortion region of the degraded broadcast channel with compressed host data available to the decoders is the collection of all rate-distortion quadruplets (R_1, R_2, R_d, D) satisfying

$$\begin{aligned} R_1 &\leq I(X; Y(1)|K, S_d, S) && (6.64) \\ R_2 &\leq I(K, S; Y(2)|S_d) - H(S|S_d) && (6.65) \\ R_d &\geq I(S_d; S) - H(S_d; Y(2)) && (6.66) \\ D &\geq E\left[d(S, X)\right] && (6.67) \end{aligned}$$

for some $((K, S_d), X, S, Y(1), Y(2)) \in \mathcal{P}$, where S_d and K are auxiliary random variables with alphabets bounded by

$$\begin{aligned} |\mathcal{S}_d| &\leq |\mathcal{S}||\mathcal{X}| + 3, && (6.68) \\ |\mathcal{K}| &\leq |\mathcal{S}_d||\mathcal{S}||\mathcal{X}| + 2. && (6.69) \end{aligned}$$

Next, we turn our attention to the causal case, and give the capacity region for this case.

Let $\mathcal{P}_c$ stands for the collection of random variables $(\widetilde{K}, S, X, Y(1), Y(2))$ such that

$$(\widetilde{K}, S, X, Y(1), Y(2)) \in \mathcal{P} \quad (6.70)$$

and

$$P_{\widetilde{K},S}(\widetilde{k}, s) = P_{\widetilde{K}}(\widetilde{k}) P_S(s). \quad (6.71)$$

The capacity region of any discrete memoryless degraded broadcast channel, with states and causal side information at the encoder, is the set of all pairs (R_1, R_2) such that

$$\begin{aligned} &R_1 \leq I(U; Y(1)|K), \\ &R_2 \leq I(K; Y(2)), && (6.72) \end{aligned}$$

for some $((K,U), S, X, Y(1), Y(2)) \in \mathcal{P}$, where K and U are auxiliary random variables which take values in the finite alphabets $\mathcal{K}$ and $\mathcal{U}$, respectively. The alphabet cardinality of these auxiliary random variables satisfy

$$\begin{aligned} |\mathcal{K}| &\leq |\mathcal{S}||\mathcal{X}| + 1, \\ |\mathcal{U}| &\leq |\mathcal{S}||\mathcal{X}| \left(|\mathcal{S}||\mathcal{X}| + 1 \right). \end{aligned} \tag{6.73}$$

In order to exhaust this capacity region, it is sufficient to take X to be a deterministic function of the triple (K, U, S), i.e., $x = f(k, u, s)$.

As in the single-user channel, where Shannon's capacity formula can be obtained as a special case of Gel'fand and Pinker non-causal case (as seen in Subsection 3.3), the capacity of the casual broadcast channel with states, is a special case of the inner and outer bounds for the non-causal case (see the proof in [110], subsection 5.5). Another similarity to the single-user channel, is that the capacity region given in (6.72) can be expresses in terms of strategies. Denote by $\mathcal{T}$ the family of all Shannon strategies, i.e., mapping from $\mathcal{S}$ to $\mathcal{X}$. For fixed k and u, $x = f(k, u, \cdot)$. The rate region, in terms of strategies, is the collection of all pairs (R_1, R_2) over the distributions $P_K(\cdot), P_{T|K}(\cdot|\cdot)$ such that

$$\begin{aligned} R_1 &\leq I(T; Y(1)|K), \\ R_2 &\leq I(K; Y(2)), \end{aligned} \tag{6.74}$$

where K ia an auxiliary random variable, $P_{T|K}(\cdot|\cdot)$ is a conditional distribution on the set $\mathcal{T}$, conditioned on K, and the pair (T, K) is independent of S.

This general capacity region, is used in [110] to determine the capacity region of the degraded broadcast modulo additive noise channel. This model is an extension of the single-user model presented in Section 4.3. This degraded broadcast modulo additive noise channel can be described as

$$\begin{aligned} Y(1) &= X + Z_1 \\ Y(2) &= Y(1) + Z_2 = X + Z_1 + Z_2 = X + Z_3, \end{aligned} \tag{6.75}$$

where Z_1 is a state dependent noise, with conditional distribution $P_{Z_1|S}(\cdot|\cdot)$, Z_2 is an additive noise independent of Z_1 and of the state

S, and Z_3 is the sum of Z_1 and Z_2,i.e., $Z_3 \triangleq Z_1 + Z_3$ with conditional distribution $P_{Z_3|S}(\cdot|\cdot)$. As in the single-user model presented in Section 4.3, the alphabets of all the RV's are the same, i.e., $\mathcal{X} = \mathcal{Y}(1) = \mathcal{Y}(2) = \mathcal{Z}_1 = \mathcal{Z}_2 = \{0, 1, ..., |\mathcal{X}| - 1\}$ and addition or subtraction is preformed modulo $|\mathcal{X}|$. Using (6.76) the capacity of the degraded broadcast modulo additive noise channel is the collection of all pairs (R_1, R_2) over the distributions $P_K(\cdot), P_{T|K}(\cdot|\cdot)$ such that

$$\begin{aligned} &R_1 \leq H(Z_1 + T(S)|K) - H(Z_1 + T(S)|T), \\ &R_2 \leq H(Y(2)) - H(Z_3 + T(S)|K). \end{aligned} \tag{6.76}$$

Similarly to the single-user case, the strategy t can be interpreted as a noise predictor, which minimizes the noise entropy.

We now consider a general broadcast channel (not necessarily degraded) with states, and with non-causal side information available to the encoder. In [115], this problem was considered and an inner bound to the capacity region was given for this model. This bound is an extension of Marton's [82] achievable region to our model.

Let $\overline{\mathcal{P}}$ stands for the collection of random variables $(\widetilde{K}, S, X, Y(1), Y(2))$, where $\widetilde{K}$ take values in the finite alphabet $\widetilde{\mathcal{K}}$. The joint probability distribution of these random variables satisfy

$$\begin{aligned} &P_{\widetilde{K},S,X,Y(1),Y(2)}(\widetilde{k}, s, x, y(1), y(2)) = \\ &\quad P_{\widetilde{K},S,X}(\widetilde{k}, s, x) P_{Y(1)|X,S}(y(1)|x, s) P_{Y(2)|X,S}(y(2)|x, s), \end{aligned} \tag{6.77}$$

and both triples

$$\begin{aligned} &\widetilde{K} \longrightarrow (S, X) \longrightarrow Y(1) \\ &\widetilde{K} \longrightarrow (S, X) \longrightarrow Y(2) \end{aligned} \tag{6.78}$$

are Markov chains.

An inner bound to the capacity region of any discrete memoryless broadcast channel, with states and non-causal side information is the convex hull of the set of all pairs (R_1, R_2) such that

$$\begin{aligned} R_1 &\leq I(K,V;Y(1)) - I(K,V;S), \\ R_2 &\leq I(K,U;Y(2)) - I(K,U;S), \\ R_1 + R_2 &\leq -\left[\max\{I(K;Y(1)), I(K;Y(2))\} - I(K;S)\right]_+ \\ &\quad +I(K,V;Y(1)) - I(K,V;S) \\ &\quad +I(K,U;Y(2)) - I(K,U;S) \\ &\quad -I(U;V|K,S) \end{aligned} \tag{6.79}$$

for some $((K,V,U),X,S,Y(1),Y(2)) \in \mathcal{P}$, where K, V and U are auxiliary random variables which take values in the finite alphabets $\mathcal{K}$, $\mathcal{V}$ and $\mathcal{U}$, respectively, and where $[a]_+ \triangleq \max\{0,a\}$.

We can extend this model and also include common messages which are transmitted to both of the users, in addition to the separate messages intended to each one of the users [115]. This model applies to the wireless mobile users example, which was given in the start of this subsection. In this example, the base station sometimes transmits a common message to the users in addition to the separate messages. This common message could be, for example, a control message intended to all of the users.

The definitions of a code with a common messages set $\{1,...,2^{NR_0}\}$ is similar to that given for the non-causal code, except that the encoder is now

$$f : \{1,...,2^{NR_1}\} \times \{1,...,2^{NR_2}\} \times \{1,...,2^{NR_0}\} \times \mathcal{S}^N \longrightarrow \mathcal{X}^N,$$

and the decoders are replaced by

$$g(i) : \mathcal{Y}^N(i) \longrightarrow \{1,...,2^{NR_i}\} \times \{1,...,2^{NR_0}\}.$$

The probability of error is normalized to $2^{N(R_1+R_2+R_0)}$, and the definition of an achievable rate region, stays the same.

An inner bound to the capacity region of any discrete memoryless broadcast channel, with states, non-causal side information and transmitted common messages is the convex hull of the set of all triples

(R_1, R_2, R_0) such that

$$\begin{aligned}
R_0 &\leq \left[\min\left\{I(K;Y(1)), I(K;Y(2))\right\} - I(K;S)\right]_+ ,\\
R_1 + R_0 &\leq I(K,V;Y(1)) - I(K,V;S),\\
R_2 + R_0 &\leq I(K,U;Y(2)) - I(K,U;S),\\
R_1 + R_2 + R_0 &\leq -\left[\max\left\{I(K;Y(1)), I(K;Y(2))\right\} - I(K;S)\right]_+\\
&\quad +I(K,V;Y(1)) - I(K,V;S)\\
&\quad +I(K,U;Y(2)) - I(K,U;S)\\
&\quad -I(U;V|K,S) \qquad (6.80)
\end{aligned}$$

for some $((K,V,U),X,S,Y(1),Y(2)) \in \mathcal{P}$, where K, V and U are auxiliary random variables which take values in the finite alphabets $\mathcal{K}$, $\mathcal{V}$ and $\mathcal{U}$, respectively. The proof of this inner bound, follows directly from the proof of the inner bound without common messages.

Khisti, Erez and Wornell [70] considered a similar problem, in which only common messages are transmitted to the users. They have found an achievable common rate for the special case of binary channels [70]. The achievable common rate is also characterized in (6.80), as this problem is a special case of the problem considered above. Therefore, the achievable common rate is given by

$$R = \min\left\{R_1, R_2\right\} + R_0, \qquad (6.81)$$

where R_1, R_2 and R_0 are characterized in (6.80).

Next, we consider the non degraded Gaussian broadcast channel with states and side information available to the transmitter. A two-user Gaussian broadcast channel is given by

$$\begin{aligned}
Y(1) &= X + S_1 + Z_1,\\
Y(2) &= X + S_2 + Z_2, \qquad (6.82)
\end{aligned}$$

where $Z_1 \sim \mathcal{N}(0,B_1)$ and $Z_2 \sim \mathcal{N}(0,B_2)$ are additive Gaussian noises, for the first and second users, $S_1 \sim \mathcal{N}(0,Q_1)$ and $S_2 \sim \mathcal{N}(0,Q_2)$ are additive Gaussian states of interferences, both known non-causally at the transmitter. The Gaussian random variables Z_1, Z_2 and (S_1,S_2) are independent of each other, except of S_1 and S_2 that can be correlated. The input X is subjected to a power constraint $E\left[X^2\right] \leq \Gamma$.

We can use (6.79) in order to characterize the capacity region of this channel. We assume that $B_1 \geq B_2$. Define the state variable $S \triangleq (S_1, S_2)$, and set a parameter $\beta \in [0,1]$. We decompose X as $X = X_1 + X_2$, where X_1 and X_2 are independent of each other, with powers $\beta\Gamma$ and $(1-\beta)\Gamma$, respectively. Let K in (6.79), be a null variable and define

$$V = X_1 + \left(\frac{\beta\Gamma}{\Gamma + B_1}\right) S_1, \tag{6.83}$$

and

$$U = X_2 + \left(\frac{(1-\beta)\Gamma}{(1-\beta)\Gamma + B_2}\right)(S_2 + X_1). \tag{6.84}$$

With these definitions and by (6.79), the following rate pair is achievable

$$\begin{aligned} R_1 &= \frac{1}{2}\log\left(1 + \frac{\beta\Gamma}{B_1 + (1-\beta)\Gamma}\right) \\ R_2 &= \frac{1}{2}\log\left(1 + \frac{(1-\beta)\Gamma}{B_2}\right). \end{aligned} \tag{6.85}$$

Therefore, the capacity region of the Gaussian broadcast channel with additive states S_1, S_2 known non-causally at the encoder, is equal to the capacity region of the Gaussian broadcast channel without additive states, with the same input constraints.

This result is in the spirit of Costa's result, and was also noticed for the special case when $\widetilde{S} \triangleq S_1 = S_2$, considered in [57], [71]. For this special case, the capacity region of this Gaussian broadcast channel coincides with the capacity region of a Gaussian broadcast channel without states.

7

Algorithms

7.1 Numerical Computation of the Channel Capacity

In this subsection, we present numerical algorithms for the computation of the capacity of channels with causal or non-causal transmitter side information. Expressions for the capacity of channels with causal or non-causal CSIT, were given is Section 3. Evaluating the capacity of these channels, can be a very difficult task. In most cases, closed-form solutions are unavailable. However, this capacity can be evaluated numerically.

A numerical algorithm for the computation of the capacity for a DMC was given in [3], [12], [29]. This algorithm is known as the Arimoto-Blahut algorithm. We start by presenting the algorithm for the DMC to show how this type of algorithm is developed. We give the the following lemma, that will help us derive the algorithm for the DMC.

Lemma 2. [36] Let $P_X(x)P_{Y|X}(y|x)$ be a given joint distribution. Then the distribution $P_Y(y)$ that minimizes the relative entropy $D\left(P_X(x)P_{Y|X}(y|x)||P_X(x)P_Y(y)\right)$ is the marginal distribution $P_Y^*(y)$

corresponding to $P_{Y|X}(y|x)$, i.e.,

$$D\left(P_X(x)P_{Y|X}(y|x)||P_X(x)P_Y^*(y)\right) = \\ \min_{P_Y(\cdot)} D\left(P_X(x)P_{Y|X}(y|x)||P_X(x)P_Y(y)\right), \tag{7.1}$$

where $P_Y^*(y) = \sum_x P_X(x)P_{Y|X}(y|x)$. Also

$$\max_{P_{X|Y}(\cdot|\cdot)} \sum_{x,y} P_X(x)P_{Y|X}(y|x) \log \frac{P_{X|Y}(x|y)}{P_X(x)} = \\ \textstyle\sum_{x,y} P_X(x)P_{Y|X}(y|x) \log \frac{P_{X|Y}^*(x|y)}{P_X(x)}, \tag{7.2}$$

where

$$P_{X|Y}^*(x|y) = \frac{P_X(x)P_{Y|X}(y|x)}{\sum_x P_X(x)P_{Y|X}(y|x)}. \tag{7.3}$$

We then rewrite the definition of the DMC capacity using Lemma 2

$$\begin{aligned} C &= \max_{P_X(\cdot)} I(X;Y) \\ &= \max_{P_{X|Y}(\cdot|\cdot)} \max_{P_X(\cdot)} \sum_X \sum_Y P_X(x)P_{Y|X}(y|x) \log \frac{P_{X|Y}(x|y)}{P_X(x)} \\ &= \max_{P_{X|Y}(\cdot|\cdot)} \max_{P_X(\cdot)} F\left(P_{X|Y}(\cdot|\cdot), P_X(\cdot)\right), \end{aligned} \tag{7.4}$$

where $F\left(P_{X|Y}(\cdot|\cdot), P_X(\cdot)\right) = \sum_X \sum_Y P_X(x)P_{Y|X}(y|x) \log \frac{P_{X|Y}(x|y)}{P_X(x)}$. The concept of the algorithm is to regard $P_{X|Y}(x|y)$ and $P_X(x)$ as independent variables and to optimize alternately between the two. We start with a guess of the maximizing distribution $P_X(x)$ and find the best conditional distribution $P_{X|Y}^*(x|y)$, which is by Lemma 2 given by (7.3). For this conditional distribution the best input distribution $P_X^*(x)$ is found by solving the maximization problem using a Lagrange multiplier to constrain $\sum_x P_X^*(x) = 1$. Hence,

$$P_X^*(x) = \frac{\prod_y \left(P_{X|Y}(x|y)^{P_{Y|X}(y|x)}\right)}{\sum_x \prod_y \left(P_{X|Y}(x|y)^{P_{Y|X}(y|x)}\right)}. \tag{7.5}$$

A stopping criterion was also given in [12]. When the criterion is met, we get the capacity of the channel for a desired accuracy. In order to find a stopping criterion, we introduce an upper-bound to the capacity [12]

$$U\left(P_X(\cdot)\right) = \max_x \sum_Y P_{Y|X}(y|x) \log \frac{P_{Y|X}(y|x)}{\sum_x P_X(x) P_{Y|X}(y|x)} \geq C. \quad (7.6)$$

The stopping criterion is when $U\left(P_X(\cdot)\right) - F\left(P_{X|Y}(\cdot|\cdot), P_X(\cdot)\right) < \epsilon$ for any desired $\epsilon > 0$. The upper-bound (7.6) appears as a problem in [54]. We can summarize the different steps of the algorithm.

(1) Start the algorithm with a guess of the maximizing distribution $P_X(\cdot)$ (the algorithm will converge from any initial distribution).
(2) Calculate $P_{X|Y}(\cdot|\cdot)$ by using (7.3).
(3) Calculate $U\left(P_X(\cdot)\right)$ and $F\left(P_{X|Y}(\cdot|\cdot), P_X(\cdot)\right)$.
(4) If $U\left(P_X(\cdot)\right) - F\left(P_{X|Y}(\cdot|\cdot), P_X(\cdot)\right) < \epsilon$ the algorithm terminates and the capacity is given by $C = F\left(P_{X|Y}(\cdot|\cdot), P_X(\cdot)\right)$, else proceed to the next step.
(5) Calculate $P_X(\cdot)$ by using (7.5) and jump to step 2.

In [12], Blahut proved that lower bound to the capacity $F\left(P_{X|Y}(\cdot|\cdot), P_X(\cdot)\right)$ converges monotonically to the capacity of the DMC. Blahut's convergence proof suffers a mistake. A convergence proof for this algorithm is complicated and can be viewed in [3] and [29]. Since the algorithm monotonically increases the lower bound $F\left(P_{X|Y}(\cdot|\cdot), P_X(\cdot)\right)$ until the convergence to the capacity, and the upper bound is equal to the lower bound only when they converge to the capacity, we can use the stopping criterion presented in the algorithm.

Next, we consider Arimoto-Blahut algorithm for numerical computation of the capacity of a channel with non-causal CSIT. Arimoto-Blahut algorithm for this channel were given in [63], [42]. We start by describing these algorithms.

The capacity of a channel with non-causal CSIT is given by (3.15) in terms of an extended input alphabet. We can rewrite the mutual information in this equation and give it as

$$\begin{aligned} &I(T;Y) - I(T;S) = \\ &\quad \textstyle\sum_{s,y,t} P_S(s) P_{T|S}(t|s) P_{Y|T,S}(y|t,s) \log \left(\frac{P_{T|Y}(t|y)}{P_{T|S}(t|s)}\right). \end{aligned} \quad (7.7)$$

The concept of the algorithm proposed in [42], similarly to the algorithm for the DMC, is to regard $P_{T|Y}(t|y)$ and $P_{T|S}(t|s)$ as independent variables and to optimize alternately between the two. Since the objective function is concave in both $P_{T|Y}(t|y)$ and $P_{T|S}(t|s)$, and since both $P_{T|Y}(t|y)$ and $P_{T|S}(t|s)$ are convex, the alternating optimization scheme must converge to the global maximum from any initial distribution [63], [29]. The authors in [42], haven't given any stopping criterion for this algorithm, although, this algorithm can have a stopping criterion, as we shall see in an extended version of this algorithm [61].

In terms of complexity, the algorithm presented so far is an improved version of an earlier algorithm given in [63]. In the original algorithm of [63], we optimize the objective function

$$\begin{aligned} &I(U;Y) - I(U;S) = \\ &\quad \textstyle\sum_{s,u,x,y} P_S(s)P_{U|S}(u|s)P_{X|U,S}(x|u,s)P_{Y|X,S}(y|x,s)\log\left(\frac{P_{U|Y}(u|y)}{P_{U|S}(u|s)}\right), \end{aligned} \tag{7.8}$$

alternately between three independent variable: $P_{X|U,S}(x|u,s)$, $P_{U|S}(u|s)$ and $P_{U|Y}(u|y)$. Therefore, this algorithm is inferior to the first algorithm, because it includes an additional step. However, this algorithm requires less computer memory when it runs. The optimization in the first algorithm, is over an extended alphabet with cardinality $|\mathcal{X}|^{|\mathcal{S}|}$. Thus, the memory requirements for the first algorithm is exponential in the number of channel states.

A generalization of the first algorithm (for a channel with states) which also take into account an average power constraint, was proposed in [61]. The capacity of a channel with non-causal CSIT and with an average power constraint was given by (3.23), where the average power constraint is given by

$$E\left[T(S)^2\right] = \sum_{s,t} P_S(s)P_{T|S}(t|s)t^2(s) \leq \Gamma. \tag{7.9}$$

Let $C(\Gamma)$ be the capacity-cost curve given in (3.23) as a function of the power constraint Γ. $C(\Gamma)$ is a monotone and concave function of Γ [6]. Therefore, the computation of the capacity-cost function for a given value of Γ can instead be solved by computing the value of the capacity-cost function at a given slope a, i.e.

$$\begin{aligned} C(\Gamma) - a\Gamma = & \max_{P_{T|S}(t|s), P_{T|Y}(t|y)} \\ \{ \quad & \textstyle\sum_{s,t,y} P_S(s) P_{T|S}(t|s) P_{Y|T,S}(y|t,s) \log \left(\frac{P_{T|Y}(t|y)}{P_{T|S}(t|s)} \right) \\ & -a \textstyle\sum_{s,t} P_S(s) P_{T|S}(t|s) t^2(s) \}. \end{aligned} \tag{7.10}$$

Again we alternately optimize over $P_{T|S}(t|s)$ and $P_{T|Y}(t|y)$,

$$P^*_{T|S}(t|s) = \frac{\left(\prod_y P_{T|Y}(t|y)^{P_{Y|T,S}(y|t,s)}\right) e^{-at^2(s)}}{\sum_{t'} \left\{ \left(\prod_y P_{T|Y}(t'|y)^{P_{Y|T,S}(y|t',s)}\right) e^{-at'^2(s)} \right\}} \tag{7.11}$$

$$P^*_{T|Y}(t|y) = \frac{\sum_s P_S(s) P_{T|S}(t|s) P_{Y|T,S}(y|t,s)}{\sum_{s',t'} P_S(s') P_{T|S}(t'|s') P_{Y|T,S}(y|t',s')}. \tag{7.12}$$

This algorithm and the algorithm presented in [63], also have a stopping criterion. This criterion may shorten the time needed to compute the capacity when compared to the algorithm given in [42] without a stopping criterion. We define

$$\begin{aligned} & F\left(P_{T|S}(\cdot|\cdot), P_{T|Y}(\cdot|\cdot)\right) = \\ & \quad \textstyle\sum_{s,t} P_S(s) P_{T|S}(t|s) \left(\sum_y P_{Y|T,S}(y|t,s) \log \frac{P_{T|Y}(t|y)}{P_{T|S}(t|s)} - at^2(s) \right) \end{aligned} \tag{7.13}$$

$$\begin{aligned} & U'\left(P_{T|S}(\cdot|\cdot), \widetilde{P}_{T|S}(\cdot|\cdot)\right) = \\ & \quad \textstyle\sum_{s,t} P_S(s) P_{T|S}(t|s) \left(\sum_y P_{Y|T,S}(y|t,s) \log \frac{\widetilde{P}^*_{T|Y}(t|y)}{\widetilde{P}_{T|S}(t|s)} - at^2(s) \right) \end{aligned} \tag{7.14}$$

$$\begin{aligned} & U\left(P_{T|S}(\cdot|\cdot)\right) = \\ & \quad \textstyle\sum_s P_S(s) \max_t \left(\sum_y P_{Y|T,S}(y|t,s) \log \frac{P^*_{T|Y}(t|y)}{P_{T|S}(t|s)} - at^2(s) \right) \end{aligned} \tag{7.15}$$

where $\widetilde{P}^*_{T|Y}(t|y)$ is the conditional distribution given by (7.12) under $\widetilde{P}_{T|S}(t|s)$. Similarly to the upper-bound in the DMC case, we can see that

$$C(\Gamma) - a\Gamma = \max_{P_{T|S}(t|s)P_{T|Y}(t|y)} F\left(P_{T|S}(\cdot|\cdot), P_{T|Y}(\cdot|\cdot)\right) \leq U\left(\widetilde{P}_{T|S}(\cdot|\cdot)\right)$$

and $F\left(P_{T|S}(\cdot|\cdot), P_{T|Y}(\cdot|\cdot)\right) = U\left(\widetilde{P}_{T|S}(\cdot|\cdot)\right)$ if and only if F is maximized over $P_{T|S}(t|s)$ and $P_{T|Y}(t|y)$ and $P_{T|S}(t|s) = \widetilde{P}_{T|S}(t|s)$. Therefore, the stopping criterion in this case is when $U\left(P_{T|S}(\cdot|\cdot)\right) - F\left(P_{T|S}(\cdot|\cdot), P_{T|Y}(\cdot|\cdot)\right) < \epsilon$ for any desired ϵ.

The algorithm for the causal side information scenario, is similar to the algorithm for the non-causal scenario and is given in [61]. The equivalence of a channel with causal CSIT to a DMC with an expanded alphabet as was demonstrated in Subsection 3.1, means that we can use the standard Arimoto-Blahut algorithm given in [3], [12] when there is no power constraint.

In order to calculate the capacity-cost function for a continuous alphabet channel, we approximate the channel by a discrete one (quantizing) and then apply the algorithms presented here to finer and finer approximations until convergence to the continuous channel capacity is reached.

We note here that similar algorithms exist for the calculation of the rate-distortion function [12], [36].

7.2 Coding Schemes

In Section 3, we have seen that the random binning scheme is one of the key elements in the solution of the problems presented in this work. We next present (in Subsection 7.2.1) lattice codes and nested lattice codes that are used (in Subsection 7.2.3) as a coding scheme for the continuous version of the dirty-paper problem, and also (in Subsection 7.2.4) as a coding scheme for the continuous version of the Wyner-Ziv problem. We also present parity-check codes and nested parity-check codes (in Subsection 7.2.2) for the binary version of the dirty-paper problem, discussed in Subsection 7.2.3, and to both the binary version of the Wyner-Ziv and the Slepian-Wolf problems, which are discussed in Subsection 7.2.4.

The concept of nested codes has been the main theme in various works pertaining to practical coding techniques for these problems. Willems proposed a scalar version of a nested code for channels with CSIT [132]. Codes for the dirty-paper and dirty-tape problems, were discussed in [47], [15], [93], [7], [93], [28]. In [23], [24], [20], [21], [43], [88], [69], [87] coding schemes which are based on nested codes, were proposed to the information embedding problem. Shamai, Verdú and Zamir [103], suggested using nested codes for the Wyner-Ziv problem. Pradhan and Ramchandran [96] proposed using nested codes for both the Slepian-Wolf and the Wyner-Ziv problems.

This subsection by no means includes a complete coverage of all coding schemes for the problems presented in this work. We have chosen to include only the main ideas of nested codes, and we show how these codes are used for the main problems presented in this paper.

7.2.1 Nested Lattice Codes for Binning Schemes

The bins of the coding schemes of Section 3, were constructed at random. This random construction is not constructive. In this subsection, we present a nested lattice code, that may be used as a binning scheme.

We start by introducing the basic terms and properties of lattice codes and nested lattice codes [143]. An N-dimensional lattice Λ is defined by a set of N basis (column) vectors $g^N(1), ..., g^N(N)$ in $\mathbb{R}^N$. The lattice is composed of all integral combinations of the basis vectors, i.e.,

$$\Lambda = \{l^N = G \cdot i^N : i^N \in \mathbb{Z}^N\} \tag{7.16}$$

where $\mathbb{Z} = \{0, \pm 1, \pm 2, ...\}$, and the $N \times N$ generator matrix G is given by
$G = [g^N(1), g^N(2), ..., g^N(N)]$. The nearest neighbor quantizer $Q(\cdot)$ associated with Λ is defined by

$$Q(x^N) = l^N \in \Lambda \text{ if } \|x^N - l^N\| \le \|x^N - l'^N\|, \quad \forall l'^N \in \Lambda \tag{7.17}$$

where $\|\cdot\|$ denotes Euclidean norm. The basic Voronoi cell of Λ is the set of points in $\mathbb{R}^N$ closest to the zero codeword, i.e.,

$$\nu_0 = \{x^N : Q(x^N) = 0\}. \tag{7.18}$$

The Voronoi cell associated with each $l^N \in \Lambda$ is a shift of ν_0 by l^N. The error in quantizing x^N with respect to Λ is given by

$$x^N \mod \Lambda = x^N - Q(x^N). \tag{7.19}$$

The second moment of Λ is defined as the second moment per dimension of a uniform distribution over ν_0, i.e.,

$$\sigma^2 = \frac{1}{V} \cdot \frac{1}{N} \int_{\nu_0} \|x^N\| dx^N, \tag{7.20}$$

where $V = \int_{\nu_0} dx^N$ is the volume of ν_0. The normalized second moment of Λ is given by

$$G(\Lambda) = \frac{1}{V^{1+2/N}} \cdot \frac{1}{N} \int_{\nu_0} \|x^N\| dx^N = \sigma^2 / V^{2/N}. \tag{7.21}$$

The minimum possible value of $G(\Lambda)$ over all lattices in $\mathbb{R}^N$ is denoted by G_N, where $G_N \geq 1/2\pi e, \forall N$. The error probability when this lattice code is used as a channel code over an AWGN channel without a channel input constraint, is the probability that a white Gaussian noise vector Z^N exceeds the basic Voronoi cell

$$P_e = P_r\{Z^N \notin \nu_0\}. \tag{7.22}$$

An important result pertaining to lattice codes is the existence of asymptotically "good" codes. We consider two definitions of "good" lattice codes:

(1) *Good channel codes over AWGN channel*: For any $\epsilon > 0$ and sufficiently large N, there exists an N-dimensional lattice Λ whose cell volume $V < 2^{N(h(Z)+\epsilon)}$, where $h(z) = \frac{1}{2}\log(2\pi eB)$ and B are the differential entropy and the variance of the AWGN Z, respectively, such that $P_e < \epsilon$.

(2) *Good source codes under mean squared distortion measure*: For any $\epsilon > 0$ and sufficiently large N, there exists an N-dimensional lattice Λ with $\log(2\pi e G_n) < \epsilon$, i.e., the normalized second moment of good lattice codes approach the bound $1/2\pi e$ as N goes to infinity.

Such channel codes approach the capacity per unit volume of the AWGN channel, and are called “good AWGN channel B-codes”, and such source codes, scaled to second moment D, approach the quadratic rate-distortion function $R(D)$ at a high-resolution quantization conditions, and are called “good source D-codes”. The basic Voronoi cell of a good lattice code, when used as a channel code, approximates Euclidean ball of radius $\sqrt{NB}$, or $\sqrt{ND}$, when used as a source code with distortion D. Therefore, the volume of the Voronoi cells of good δ-codes satisfies asymptotically

$$\frac{1}{N}\log V \approx \frac{1}{2}\log(2\pi e\delta) \tag{7.23}$$

where δ corresponds to B or D, and $\approx$ means approximation in an exponential sense, i.e., the difference between the normalized logarithms is small. A lattice which is good in one sense is not necessarily good in the other.

Definition 12. A pair of N-dimensional lattices (Λ_1, Λ_2) if

$$\Lambda_2 \subset \Lambda_1, \tag{7.24}$$

i.e., each codeword of Λ_2 is also a codeword of Λ_1.

It can be shown that there exists corresponding generator matrices G_1 and G_2, such that $G_2 = G_1 \cdot J$ where J is an $N \times N$ integer matrix whose determinant is greater than one. Λ_1 is called the fine lattice and Λ_2 is called the coarse lattice. The volume of the Voronoi cells of Λ_1 and Λ_2 satisfy

$$V_2 = |J| \cdot V_1 \tag{7.25}$$

where $V_2 = Vol(\nu_{0,2})$ and $V_1 = Vol(\nu_{0,1})$ and $\nu_{0,i}$ is the basic Voronoi cell of $\Lambda_i, i = 1, 2$. We call $\sqrt[N]{|J|} = \sqrt[N]{V_2/V_1}$ the nesting ratio.

The points of the set

$$\{\Lambda_1 \mod \Lambda_2\} \triangleq \{\Lambda_1 \bigcap \nu_{0,2}\} \tag{7.26}$$

are called the coset leaders of Λ_2 relative to Λ_1. For each coset leader $v^N \in \{\Lambda_1 \mod \Lambda_2\}$, the shifted lattice $\Lambda_{2,v^N} = v^N + \Lambda_2$ is called a coset of Λ_2 relative to Λ_1. The cosets Λ_{2,v^N}, $v^N \in \{\Lambda_1 \mod \Lambda_2\}$ are disjoint. Therefore, there are $V_2/V_1 = |J|$ different cosets, whose union gives the fine lattice

$$\bigcup_{v^N \in \{\Lambda_1 \mod \Lambda_2\}} \Lambda_{2,v^N} = \Lambda_1. \tag{7.27}$$

In order to implement a binning scheme with nested lattice codes, we require that one of the lattices is a good channel code over an AWGN channel, and the other one is good for source coding under mean squared error distortion measure (or with a power constraint in the dirty-paper problem) as mentioned in Subsection 3.2. In a good binning scheme, each bin should contain a good collection of representative points which spread over the entire space. Hence, each bin plays the role of a good source code. The collection of all the codewords in all the bins, should play the role of a good channel code, in order to overcome the noise in the channel. If the fine lattice is a good δ_1-code and the coarse lattice is a good δ_2-code, $\delta_2 > \delta_1$, then the number of cosets of Λ_2 relative to Λ_1 is about

$$V_2/V_1 = (\delta_2/\delta_1)^{N/2}, \tag{7.28}$$

where for a good channel code component, δ_1 indicates the AWGN power, which is in general smaller than, or equal to the second moment of the lattice, and for a good source code component, δ_2 indicates the mean square distortion, which coincides with the second moment of the lattice.

7.2.2 Nested Parity-Check Codes for Binning Schemes

We present the binary counterpart of the nested lattice codes, called the nested parity-check codes. These codes are use as a binary binning scheme for the discrete case, e.g., coding to a memory with defective cells. Kuznetsov and Tsybakov introduced such a code for the noiseless problem of coding to a memory with defective cells [75], and Tsybakov [123] introduced such a code for the more general case of a noisy memory with defective cells.

As in Subsection 7.2.1, we start by introducing the basic terms and properties of parity-check codes and nested parity-check codes [143].

Let an (N, K) binary parity-check code be specified by the $(N-K) \times N$ binary parity-check matrix H. The code $\mathcal{C} = \{c^N\}$ contains all binary vectors c^N whose syndrome, defined by $s^{(N-K)} \triangleq Hc^N$, is equal to zero[1]. There are 2^K codewords in $\mathcal{C}$ as there are 2^K linearly independent rows of H, so the code rate is $\frac{\log|\mathcal{C}|}{N} = K/N$. The set of all vectors x^N satisfying $Hx^N = s^{(N-K)}$, where $s^{(N-K)}$ is some general syndrome $s^{(N-K)} \in \{0,1\}^{(N-K)}$, is called a coset, and is denoted by $\mathcal{C}_{s^{(N-K)}}$. We define a decoding function $f : \{0,1\}^{(N-K)} \longrightarrow \{0,1\}^N$, where $f(s^{(N-K)})$ is equal to the vector $v^N \in \mathcal{C}_{s^{(N-K)}}$ with the minimum Hamming weight, where ties are broken arbitrarily. The coset, is a shift of the code $\mathcal{C}$ by the vector v^N, i.e.,

$$\mathcal{C}_{s^{(N-K)}} \triangleq \left\{x^N : Hx^N = s^{(N-K)}\right\} = \left\{c^N + v^N : c^N \in \mathcal{C}\right\} \tag{7.29}$$

where the vector $v^N = f(s^{(N-K)})$ is called the coset leader.

As in the continuous case discussed in Subsection 7.2.1, maximum-likelihood decoding of parity-check code, over a binary symmetric channel with transition probability p (BSC(p))[2] $y^N = x^N + z^N$, where x^N, z^N and y^N are the channel input, the channel noise and the channel output, respectively, amounts to quantizing y^N to the nearest vector in $\mathcal{C}$ with respect to the Hamming distance. This vector, $\hat{c}^N$, is computed by the following procedure

$$\hat{c}^N = y^N + \hat{z}^N, \qquad \hat{z}^N = f(Hy^N). \tag{7.30}$$

As in the continuous case, the error in quantizing y^N by $\mathcal{C}$, is given by

$$\hat{z}^N = f(Hy^N) = y^N \mod \mathcal{C}. \tag{7.31}$$

The basic Voronoi set is defined as the set of vectors z^N closest to the zero vector, i.e.,

$$\left\{z^N : z^N + f(Hz^N) = 0\right\} = \Omega_o. \tag{7.32}$$

The quantizer presented above, may be viewed as a partition of $\{0,1\}^N$ to 2^K decision cells of size $2^{(N-K)}$ each, which are all shifted version of

[1] Multiplication and addition is the binary case are modulo 2.

[2] A binary symmetric channel has a binary input and output, and its output is equal to the input with probability $(1-p)$, and with probability p, on the other hand, a 0 is received as 1, and vice versa.

the basic Voronoi set Ω_0. Each of the $2^{(N-K)}$ members of Ω_0 is a coset leader for a different coset.

As in the continuous case, we consider two definitions of "good" codes:

(1) *Good channel codes over BSC(p) [36]*: For any $\epsilon > 0$ and sufficiently large N, there exists an (N, K) code of rate $K/N > C - \epsilon$, where $C = 1 - H(p)$ is the BSC(p) capacity, with a probability of decoding error smaller than ϵ

$$P_r\left\{\hat{Z}^N \neq Z^N\right\} = P_r\left\{f(HZ^N) \neq Z^N\right\} < \epsilon \tag{7.33}$$

where Z^N denotes the channel noise vector (a Bernoulli(p) vector), and $\hat{Z}^N$ denotes its estimation given by (7.30).

(2) *Good source codes under Hamming distortion [36]*: For any $0 \leq D \leq 1/2$, $\epsilon > 0$, and sufficiently large N, there exists an (N, K) code of rate $K/N < R(D) + \epsilon$, where $R(D) = 1 - H(D)$ is the rate-distortion function of a binary symmetric source (BSS) X^N, such that the expected quantization error Hamming weight satisfies

$$\frac{1}{N}E\left\{\omega_H\left(X^N + \hat{X}^N\right)\right\} = \frac{1}{N}E\left\{\omega_H\left(E^N\right)\right\} < D+\epsilon \tag{7.34}$$

where $\omega_H(\cdot)$ denotes the Hamming weight, $\hat{X}^N$ denotes the quantization of X^N by the code, and where $E^N = X^N + \hat{X}^N = f(HX^N)$ is the quantization error, which is uniformly distributed over Ω_0.

Such channel codes are called "good BSC p-codes", and such source codes are called "good BSS D-codes". The basic Voronoi cell of a good (K, N) parity-check code, approximate a Hamming ball of radius Np, when used as a channel code, or ND, when used as a source code under Hamming distortion D.

As in the continuous case, a nested code is a pair of linear codes $(\mathcal{C}_1, \mathcal{C}_2)$ satisfying

$$\mathcal{C}_2 \subset \mathcal{C}_1 \tag{7.35}$$

i.e., each codeword of $\mathcal{C}_2$ is also a codeword of $\mathcal{C}_1$, and as in the continuous case, we call $\mathcal{C}_1$ the fine code and $\mathcal{C}_2$ the coarse code. A pair

$\{(N, K_1), (N, K_2)\}$ of parity-check codes, $K_1 > K_2$, is nested in the sense of (7.35), if there exists corresponding parity-check matrices H_1 and H_2, such that

$$H_2 = \begin{bmatrix} H_1 \\ \cdots \\ \triangle H \end{bmatrix} \tag{7.36}$$

where H_1 is an $(N - K_1) \times N$ matrix, H_2 is an $(N - K_2) \times N$ matrix, and $\triangle H$ is a $(K_1 - k_2) \times N$ matrix. By (7.36), the syndromes $s_1^{(N-K_1)} = H_1 x^N$ and $s_2^{(N-K_2)} = H_2 x^N$ associated with some vector x^N are related as

$$s_2^{(N-K_2)} = \begin{bmatrix} s_1^{(N-K_1)} \\ \triangle s^{(K_1-K_2)} \end{bmatrix},$$

where $\triangle s^{(K_1-K_2)}$ is the syndrome vector associated with $\triangle H$. Therefore, $\mathcal{C}_1$ is partitioned into $2^{(K_1-K_2)}$ cosets of $\mathcal{C}_2$ by setting $s_1^{(N-K_1)} \equiv 0$, and varying $\triangle s^{(K_1-K_2)}$, i.e.,

$$\mathcal{C}_1 = \bigcup_{\triangle s^{(K_1-K_2)} \in \{0,1\}^{(K_1-K_2)}} \mathcal{C}_{2, s_2^{(N-K_2)}}, \text{where}$$

$$s_2^{(N-K_2)} = \begin{bmatrix} 0 \\ \cdot \\ \cdot \\ \cdot \\ 0 \\ \triangle s^{(K_1-K_2)} \end{bmatrix} \tag{7.37}$$

where for each $\triangle s^{(K_1-K_2)} \in \{0,1\}^{(K_1-K_2)}$ and therefore, each $s_2^{(N-K_2)}$ as in (7.37), $\mathcal{C}_{2, s_2^{(N-K_2)}}$ is called the coset of $\mathcal{C}_2$ relative to $\mathcal{C}_1$. We denote by $\Omega_{0,i}$ the basic Voronoi cell of $\mathcal{C}_i, i = 1, 2$.

If the fine code is a good δ_1-code and the coarse code is a good δ_2-code, for some δ_1, δ_2 such that $\delta_2 > \delta_1$, then the number of cosets in (7.37) is about

$$\frac{|\Omega_{0,2}|}{|\Omega_{0,1}|} = 2^{(K_1-K_2)} \approx 2^{N[H(\delta_2)-H(\delta_1)]}. \tag{7.38}$$

7.2.3 Dirty Paper Codes

In this subsection, we present coding schemes for both the continuous dirty-paper problem and the binary modulo-2 additive noise channel problem (see Subsection 4.1), called in this work the binary dirty-paper problem, which are based on a nested lattice and a nested parity-check codes. The nested lattice coding scheme, can also be adapted to be used in the dirty-tape problem where we use lattices with one dimension [48], [15], or for watermarking problems [23], [24], [20], [43]. The nested parity-check coding scheme, can be used for a memory with defective cells [123].

In both the binary and continuous cases, we use a pair of nested lattice/parity-check codes, and tune the fine code to the noise level, and the coarse code to the input constraint. In the continuous case, we use for the fine lattice Λ_1 a good channel $\frac{\Gamma B}{\Gamma+B}$-code, and for the coarse lattice Λ_2 a good source Γ-code, where Γ is the power constraint. In the binary case, we use for the fine code $\mathcal{C}_1$ a good channel p-code, and for the coarse code $\mathcal{C}_2$ a good source γ-code, where γ is the Hamming input constraint. Let the random vector K^N be uniform over the basic Voronoi cell of the coarse code, i.e., $\nu_{0,2}$ for the lattice case, and $\Omega_{0,2}$ for the binary case. K^N is a dither signal available to both the encoder and decoder. We also use, for the continuous alphabet case, the optimum estimation coefficient $\alpha = \frac{\Gamma}{\Gamma+B}$ defined in Subsection 4.1. A coding scheme for the dirty-paper problem is depicted in Fig. 7.1. We use the following coding scheme:

Message selection: identify each coset $\Lambda_{2,v^N}, v^N \in \{\Lambda_1 \bigcap \nu_{0,2}\}$, where v^N is the coset leader, with a unique message $m \in \mathcal{M}$. By (7.28), this amounts to $\log(V_2/V_1) \approx \frac{N}{2}\log(1+\frac{\Gamma}{B})$ bits per N-block.

Encoding: transmit the error vector between $\alpha s^N + k^N$ and the selected coset Λ_{2,v^N}, i.e.,

$$x^N = [v^N - \alpha s^N - k^N] \mod \Lambda_2 \tag{7.39}$$

where s^N is the interference vector and k^N is the dither. By the properties of dithered quantization, $\frac{1}{N}E\left\{\|X^N\|^2 \; |V^N = v^N, S^N = s^N\right\} = \Gamma$, independently of the value of v^N and s^N, where the expectation is over the dither K^N.

Decoding: reconstruct the message as the unique coset containing $Q_1(\alpha y^N + k^N)$, where $Q_1(\cdot)$ is the nearest neighbor quantizer associated with Λ_1. The leader of this coset can be computed as

$$\hat{v^N} = Q_1(\alpha y^N + k^N) \mod \Lambda_2 = Q_1(\tilde{y}^N) \mod \Lambda_2, \quad (7.40)$$

where $\hat{v^N}$ denotes the reconstructed coset leader from which we can estimate the transmitted message, and for a convenience reason, we denote the sum $\alpha y^N + k^N$, by $\tilde{y}^N \triangleq \alpha y^N + k^N$.

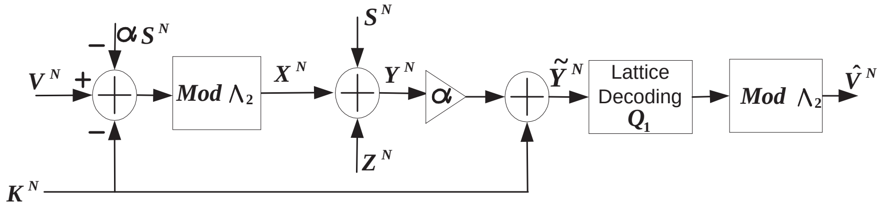

Fig. 7.1 Nested coding scheme for the dirty-paper problem.

In the binary case, we set $\alpha = 1$, replace Λ_2 by $\mathcal{C}_2$, and replace the nearest neighbor quantizer $Q_1(\cdot)$ associated with Λ_1, by the minimum Hamming distance decoding function (7.30) associated with $\mathcal{C}_1$ and use the same coding scheme. Using this coding scheme for the binary case, by (7.38) we have

$$\log \frac{|\Omega_{0,2}|}{|\Omega_{0,1}|} \approx N\,[H(\gamma) - H(p)] \quad (7.41)$$

bits of information, and a γ average Hamming weight of the codewords. By (4.4), $[H(\gamma) - H(p)]$ is exactly the maximal rate which can be reliably communicated over this channel.

Next, we will show that the decoding error probability for the continuous case given by

$$P_e = Pr\left(\hat{V}^N \neq V^N\right) \quad (7.42)$$

is small. We start by replacing the gain α in the signal path depicted in Fig. 7.1, by a short circuit and compensate for that by subtracting $(1-\alpha)y^N$ from $\tilde{y}^N$. We also add another mod-Λ_2 operation and denote its output by $\bar{y}^N$, i.e., $\bar{y}^N = [\alpha y^N + k^N] \mod \Lambda_2$. By adding the mod-$\Lambda_2$ operation, we do not affect the final result, and we can reconstruct the message by computing the leader of coset chosen by the encoder $\hat{v}^N = Q_1(\bar{y}^N) \mod \Lambda_2$, instead of (7.40). The resulting channel from V^N to $\bar{Y}^N$ is a modulo-Λ_2 additive noise channel described by the following lemma:

Lemma 3 (Inflated lattice lemma [48]). The channel from V^N to $\bar{Y}^N$, for K^N uniformly distributed over $\nu_{0,2}$, is equivalent in distribution to the channel

$$\bar{Y}^N = [V^N + Z_{eq}^N] \mod \Lambda_2, \tag{7.43}$$

where

$$Z_{eq}^N = [(1-\alpha)K^N + \alpha Z^N], \tag{7.44}$$

and where Z_{eq}^N is independent of V^N.

Substituting $\alpha = \frac{\Gamma}{\Gamma+B}$, the second moment per dimension of Z_{eq}^N is equal to $(1-\alpha)^2\Gamma + \alpha^2 B = \frac{\Gamma B}{\Gamma+B}$. If Z_{eq}^N were AWGN (a Gaussian random variable has the greatest entropy), then by using for the fine lattice a good channel $\frac{\Gamma B}{\Gamma+B}$-code, $Q_1(\bar{Y}^N) \mod \Lambda_2$ is equal to V^N with high probability, and therefore, P_e is small as desired.

We have assumed that Z_{eq}^N is AWGN. However, $(1-\alpha)K^N$ is not Gaussian, but rather uniform over $\nu_{0,2}$. This component of Z_{eq}^N is called the self noise component. Therefore, Z_{eq}^N deviates from an AWGN distribution. If we put an additional condition on the fine code, the probability of the decoding error will go to zero in spite of the slight deviation of Z_{eq}^N from AWGN. This condition extends the meaning of a good channel code, that was defined in Subsection 7.2.1:

(1) *Exponentially good channel codes over AWGN channel*: For any N and $\epsilon > 0$, there exists an N-dimensional lattice Λ whose cell volume $V < 2^{N(h(Z)+\epsilon)}$, where $h(z) = \frac{1}{2}\log(2\pi e B)$ and B are the differential entropy and the variance of the AWGN Z, respectively, such that $P_e = \{Z^N \neq \nu_0\} < e^{-N\cdot E(\epsilon)}$ where $E(\epsilon) > 0$.

The effect of the self noise component on the decoding error probability is subexponential in N relative to an AWGN noise with the same power [143]. Thus, if the fine lattice Λ_1 is an exponentially good channel $\frac{\Gamma B}{\Gamma+B}$-code, the effect of the equivalent noise Z_{eq}^N with the self noise component is asymptotically equivalent to AWGN.

In the coding scheme we have just presented, the basic cell of the coarse lattice, Λ_2, defines the region of the code where the codewords are points of the fine lattice Λ_1. Thus the coarse lattice determines the shaping gain while the fine lattice the coding gain. This coding scheme may be viewed as a generalization of Tomlinson-Harashima (TH) precoding [121], where the interference signal S plays the role of ISI, and the scalar modulo operation of TH amounts to the special case of a coarse lattice with one dimension. Therefore, this one dimension lattice means that there will be no shaping gain. In [142], [139] precoding schemes for the broadcast channel, which are viewed as a generalization of TH precoding, were suggested for the broadcast channel.

Most of the coding techniques which have been proposed for the dirty-paper problem (see Section 7.2 for a partial list), are based on the nested code scheme. A generalization of the nested lattice coding scheme, was given in [8] and is called superposition coding. In the nested lattice technique we have presented in Subsection 7.2.1, we require that the fine lattice Λ_1 be designed as a good channel code, while the coarse lattice Λ_2 be designed as a good source code. As mentioned in Subsection 7.2.1, a lattice which is good in one sense is not necessarily good in the other. Superposition of codes, enables independent selection of the two codes, where one code, which is called the auxiliary code (denoted by $\mathcal{C}_2$), should be a good channel code, and the other one, which is called the quantization code (denoted by $\mathcal{C}_1$), should be both a good channel and source code. We describe this coding scheme for the binary dirty-paper problem.

The quantization code $\mathcal{C}_1$ is randomly generated with uniform i.i.d. elements with rate R_1, while the auxiliary code $\mathcal{C}_2$ is randomly generated according to an i.i.d. Bernoulli-q probability distribution with rate R_2, where q is some constant $q \in [0, 1]$. The superposition code is defined as $\mathcal{C} \triangleq \mathcal{C}_1 + \mathcal{C}_2$, i.e., the codewords of the superposition code, are given by

$$\mathcal{C} \triangleq \left\{c^N = c_1^N + c_2^N : c_1^N \in \mathcal{C}_1, c_2^N \in \mathcal{C}_2\right\}. \tag{7.45}$$

Define the minimum Hamming distance decoding function for a code $\mathcal{C}$ (not necessarily a parity-chec code) as

$$Q_{\mathcal{C}}(x^N) = \arg\min_{c^N \in \mathcal{C}} d_H(x^N, c^N), \tag{7.46}$$

where $d_H(a^N, b^N)$ denotes Hamming distance between two length-N vectors a^N and b^N.

We use the following coding scheme:

Encoding: select a codeword $c_2^N \in \mathcal{C}_2$, where the codeword index is the message to be transmitted, and send the sequence

$$x^N = c_2^N + s^N \mod \mathcal{C}_1 = c_2^N + s^N + Q_{\mathcal{C}_1}(c_2^N + s^N) = c_2^N + s^N + c_1^N$$

where $c_1^N = Q_{\mathcal{C}_1}(c_2^N + s^N)$.

Decoding: reconstruct c_1^N from

$$y^N = x^N + s^N + z^N = c_1^N + c_2^N + z^N \tag{7.47}$$

by treating $z^N + c_2^N$ as noise, i.e.,

$$\hat{c}_1^N = Q_{\mathcal{C}_1}(c_1^N + c_2^N + z^N), \tag{7.48}$$

next, reconstruct c_2^N from $y^N + \hat{c}_1^N$, i.e.,

$$\hat{c}_2^N = Q_{\mathcal{C}_2}(y^N + \hat{c}_1^N). \tag{7.49}$$

The codeword c_2^N of $\mathcal{C}_2$ contains the transmitted data, and hence the rate of this scheme is R_2.

If R_1 and R_2 satisfy both

$$R_1 > 1 - H(\gamma) \tag{7.50}$$

and

$$\begin{aligned} R_2 &\le H(q(1-p) + p(1-q)) - H(p) \\ R_1 + R_2 &\le 1 - H(p), \end{aligned} \tag{7.51}$$

where we select q such that $q(1-p)+p(1-q)=\gamma$, then the average probability of encoding and decoding error of the above coding scheme, approaches zero with N (see [8]). Since these probabilities approaches zero with N, we are ensured that most codes in the random-coding ensemble are good in both senses.

We note here that similar superposition coding scheme for the continuous dirty-paper problem exist (for mores details see [8]).

7.2.4 Codes for the Wyner-Ziv and the Slepian-Wolf Problems

We start by presenting a coding scheme for the Wyner-Ziv problem introduced is Subsection 6.2. This coding scheme, as the coding scheme for the dirty-paper problem, is based on a nested lattice code.

As in the nested lattice code for the continuous dirty-paper problem, we use a nested lattice pair (Λ_1, Λ_2), and, contrary to the dirty-paper code, we tune the fine code to the distortion constraint, and the coarse code to the noise level by choosing for the fine lattice Λ_1 a good source D-code, where D is the distortion constraint, and for the coarse lattice Λ_2 a good channel B-code, where B is the noise variance. Let the random vector K^N be uniform over the basic Voronoi cell of the fine code (uniform over the coarse lattice in the dirty-paper setting), i.e., $\nu_{0,1}$. We also use the optimum estimation coefficient $\alpha = \sqrt{1-D/B}$. A coding scheme for the Wyner-Ziv problem is depicted in Fig. 7.2. We use the following coding scheme:

Encoding: quantize $\alpha x^N + k^N$ to the nearest point in Λ_1, i.e., $\tilde{x}^N = Q_1(\alpha x^N + k^N)$, then transmit an index which identifies $v^N = \tilde{x}^N \mod \Lambda_2$, the leader of the unique relative coset containing $\tilde{x}^N$. By (7.28), this index requires $\log(V_2/V_1) \approx \frac{N}{2}\log(B/D)$ bits.

Decoding: decode the coset leader v^N, and reconstruct x^N as

$$\hat{x}^N = y^N + \alpha\hat{w}^N, \quad \textit{where } \hat{w}^N = \left[v^N - k^N - \alpha y^N\right] \mod \Lambda_2, \tag{7.52}$$

and where y^N is the side information (denoted in Subsection 6.2 by s^N).

As seen in the above coding scheme, the coding rate of this scheme

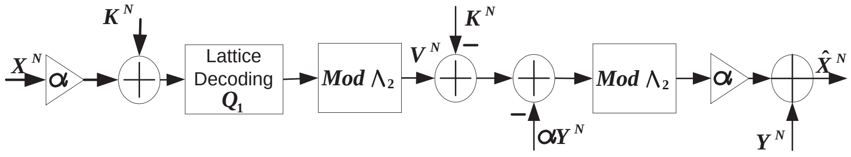

Fig. 7.2 Nested coding scheme for the Wyner-Ziv problem.

coincides with (6.10). The expected mean squared reconstruction error is $\frac{1}{N}E\left\|\hat{X}^N - X^N\right\| \leq D$, as shown in [143].

Here, similarly to the dirty-paper coding scheme presented in Subsection 7.2.3, the distribution of the error vector $\hat{x}^N - x^N$ is not Gaussian due to the self noise phenomenon [143]. Therefore, we need to put an additional condition on the coarse lattice code, and use for the coarse lattice Λ_2 an exponentially good channel B-code, that was defined in Subsection 7.2.3, instead of a good channel B-code.

Similarly to the dirty-paper setting, the Wyner-Ziv setting has a binary version, and the coding scheme for the continuous case presented here, is also suitable to be used in the binary case with the replacement of the nested lattice codes with nested parity-check codes and setting $\alpha = 1$ [143].

We now turn our attention to the Slepian-Wolf problem, presented in Subsection 6.1, which can model a lossless version of the Wyner-Ziv problem. In this setting, we choose a good BSC p-code, in the sense that was defined in Subsection 7.2.2, and use as bins its $2^{(N-K)} \approx 2^{NH(p)}$ cosets. A coding scheme for the Slepian-Wolf problem is depicted in Fig. 7.3. We use the following coding scheme:

Encoding: transmit the syndrome $s^{(N-K)} = Hx^N$, this requires $N - K \approx NH(p)$ bits.

Decoding: decode the coset leader v^N associated with $s^{(N-K)}$ by $v^N = f(s^{(N-K)})$, calculate the error between the side information y^N (denoted by s^N is Subsection 6.1) and the coset $\mathcal{C}_{v^N}$

by

$$\begin{aligned}\hat{z}^N &= (v^N + y^N) \mod \mathcal{C} \\ &= f(s^{(N-K)} + Hy^N), \qquad (7.53)\end{aligned}$$

and reconstruct x^N as

$$\hat{x}^N = y^N + \hat{z}^N,$$

where addition is modulo-2, and where $f(\cdot)$ is defined in Subsection 7.2.2.

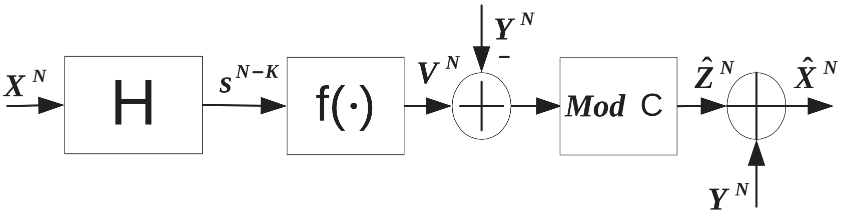

Fig. 7.3 A coding scheme for the Slepian-Wolf problem.

In this coding scheme, the probability of decoding error, by (7.33), is smaller than ϵ, $\epsilon > 0$ for a good BSC p-code.

8

Conclusion

In this paper we have reviewed information-theoretic aspects of the problem of coding to channels in the presence of side information. After describing the statistical model for this type of channels, we have given the main results for the capacity of a general channel controlled by a state sequence, and where a CSI signal is available at the transmitter non-causally (causally, respectively) or at the receiver. We have also introduced Arimoto-Blahut numerical algorithms for the computation of this capacity.

Various specific channel models of communication systems emerge from these general models. In particular, we have focused on the dirty-paper channel, the AWGN channel with fading and the modulo additive noise channel. These specific channels can serve for modeling in a wide range of problems, e.g., the Gaussian vector broadcast channel and the watermarking problem which were presented in this paper.

In the last few years, coding strategies that come close to the optimum have been widely studied. The nested lattice codes are shown as a practical binning strategy. The random binning strategy, was used for showing achievability in our general channel model.

The models and applications which where discussed in this paper, inspire results for other problems and models. We have given some related models and problems which are very similar to the models and applications presented in this paper.

Some of the models discussed in this paper still remain open problems. These models includes: the capacity of the general dirty-tape problem, the capacity of a discrete memoryless channel with states and rate-limited side information, the capacity region of the joint problem of pure information transmission and state estimation, analysis of the error exponents of the Gel'fand-Pinsker channel, and finally, coding schemes for the Gel'fand-Pinsker channel with feedback.

To conclude, we hope that this overview will resolve in better understanding of this research field and will help to attract interest to this field.

References

[1] R. Ahlswede, “Arbitrarily varying channels with states sequence known to the sender,” *IEEE Trans. Inform. Theory*, vol. IT-32, no. 5, pp. 621-629, Sep. 1986.

[2] M. Airy, A. Forenza, R. W. Heath. Jr, and S. Shakkottai, “Practical Costa precoding for the multiple antenna broadcast channel,” *Proc. IEEE Globecom 2004*, Dallas Nov. 2004.

[3] S. Arimoto, “An algorithm for computing the capacity of arbitrary discrete memoryless channel,” *IEEE Trans. Inform. Theory*, vol. IT-18, no. 1, pp. 14-20, Jan. 1972.

[4] M. E. Arutyunyan, “E-capacity of an arbitrary varying channel with an informed coder,” *Problemy peredachi informatsii*, vol. 26, no. 4, pp. 16-23, Oct.-Dec. 1990.

[5] M. E. Arutyunyan, “Bound on E-capacity of a channel with random parameters,” *Problemy peredachi informatsii*, vol. 27, no. 1, pp. 14-23, Mar. 1991.

[6] R. J. Barron, B. Chen and G. W. Wornell, “The duality between information embedding and source coding with side information and some applications,” *IEEE Trans. Inform. Theory*, vol. 49, no. 5, pp. 1159-1180, May 2003.

[7] B. Beferull-Lozano, and S. Diggavi, “Nested trellis codes and shaping for the transmitter side-information problem,” *Proc. Int. Symp. Inform. Theory (ISIT) 2003*, Yokohama, Japan, p. 183, Jul. 2003.

[8] A. Bennatan, D. Burshtein, G. Caire, and S. Shamai (Shitz), “Superposition coding for side-information channels,” *IEEE Trans. Inform. Theory*, vol. 52, no. 5, May 2006.

[9] P. P. Bergmans, “A simple converse for broadcast channels with additive white gaussian noise,” *IEEE Trans. Inform. Theory*, vol. IT-20, no. 2, pp. 279-280, Mar. 1974.

[10] Bhashyam, A. Sabharwal, and B. Aazhang, "Feedback gain in multiple antenna systems," *IEEE Trans. Commun*, vol. 50, pp. 785-798, May 2002.

[11] E. Biglieri, J. Proakis, and S. Shamai (Shitz), "Fading channels: information-theoretic and communications aspects," *IEEE Trans. Inform. Theory*, vol. 44, no. 6, pp. 2619-2692, Oct. 1998.

[12] R. Blahut, "Computation of channel capacity and rate-distortion functions," *IEEE Trans. Inform. Theory*, vol. IT-18, no. 4, pp. 460-472, Jul. 1972.

[13] G. Caire and S. Shamai (Shitz), "On the capacity of some channels with channel state information," *IEEE Trans. Inform. Theory*, vol. 45, no. 6, pp. 2007-2019, Sep. 1999.

[14] G. Caire and S. Shamai (Shitz), "On the multiple antenna broadcast channel," *36th Asilomar Conf. on Signals, Systems and Computers Asilomar*, Pacific Grove, CA, 3-6 Nov. 2001.

[15] G. Caire and S. Shamai (Shitz), "Writing on dirty tape with LDPC codes," *DIMACS Workshop on Signal Processing for Wireless Transmission*, Rutgers University, NJ, USA, 7-9 Oct. 2002.

[16] G. Caire and S. Shamai (Shitz), "On the achievable throughput of a multi-antenna Gaussian broadcast channel," *IEEE Trans. Inform. Theory*, vol. 49, pp. 1691-1706, Jul. 2003.

[17] Y. Cemal and Y. Steinberg, "Coding problems for channels with partial state information at the transmitter," submitted to *IEEE Trans. Inform. Theory*, 2005.

[18] Y. Cemal and Y. Steinberg, "Hierarchical and joint source-channel coding with coded state information at the transmitter," *Proc. Int. Symp. Inform. Theory (ISIT) 2005*, Adelaide, pp. 636-640, Sep. 2005.

[19] Y. Cemal and Y. Steinberg, "The multiple access channel with partial state information at the encoders," *IEEE Trans. Inform. Theory*, vol. 51, pp. 3992-4003, Nov. 2005.

[20] B. Chen and G. W. Wornell, "Digital watermarking and information embedding using dither modulation," *Proc. IEEE Workshop on Multimedia Signal Processing*, Redondo Beach, CA, Dec. 1998.

[21] B. Chen and G. W. Wornell, "Dither modulation: a new approach to digital watermarking and information embedding," *Proc. SPIE: Security and Watermarking of Multimedia Contents*, vol. 3657, pp. 342-353, San José, CA, Jan. 1999.

[22] B. Chen and G. W. Wornell, "Achievable performance of digital watermarking systems," *Proc. Int. Conf. Multimedia Computing and Systems (ICMCS-99)*, vol. 1, pp. 13-18, Florence, Italy, Jun. 1999.

[23] B. Chen and G. W. Wornell, "Preprocessed and postprocessed quantization index modulation methods for digital watermarking," *Proc. SPIE: Security and Watermarking of Multimedia Contents (EI'00)*, pp. 48-59, San Jose, CA, Jan. 2000.

[24] B. Chen and G. W. Wornell, "Quantization index modulation: a class of provably good methods for digital watermarking and informaion embedding," *IEEE Trans. Inform. Theory*, vol. 47, pp. 1423-1443, May 2001.

[25] M. Chiang and T. M. Cover, "Unified duality between channel capacity and rate distortion with state information," *Proc. Int. Symp. Inform. Theory (ISIT) 2002*, Lausanne, Switzerland, p. 226, Jul. 2002.

[26] J. Chou, S. Pradhan, L. El Ghaoui, and K. Ramchandran, "A robust optimization solution to the data hiding problem using distributed source coding principles," *Proc. SPIE conference*, San Jose, CA, vol. 3971, Jan. 2000.

[27] J. Chou, S. S. Pradhan, and K. Ramchandran, "On the duality between distributed source coding and data hiding," *Proc. 33rd Asilomar Conference on Signals, Systems and Computers,*pp. 1503-1507, Nov. 1999.

[28] J. Chou, S. S. Pradhan, and K. Ramchandran, "Turbo coded trellis-based constructions for data embedding: channel coding with side information," *Proc. Asilomar Conference on Signals, Systems and Computers*, Pacific Grove, USA, Oct. 2001.

[29] I. Csiszár and G. Tusnády, "Information geometry and alternating minimization procedures," *Statistics and Decisions*, Supplement Issue No. 1, pp. 205-237, 1984.

[30] A. S. Cohen, "Communication with side information," graduate Seminar 6.962, MIT, Cambridge, MA, Spring 2001, available on (`http://web.mit.edu/6.962/www/www_spring_2001/schedule.html`).

[31] A. S. Cohen and A. Lapidoth, "The Gaussian watermarking game," *IEEE Trans. Inform. Theory*, vol. 48, no. 6, pp. 1639-1667, Jun. 2002.

[32] A. S. Cohen and A. Lapidoth, "Generalized writing on dirty paper," *Proc. Int. Symp. Inform. Theory (ISIT) 2002*, Lausanne, Switzerland, p. 227, Jul. 2002.

[33] A. S. Cohen and R. Zamir, "Writing on dirty paper in the presence of difference set noise," *Proc. 41th Allerton Conference*, Illinois, USA, Oct. 2003.

[34] M. H. M. Costa, "Writing on dirty paper," *IEEE Trans. Inform. Theory*, vol. IT-29, no. 3, pp. 439-441, May 1983.

[35] T. M. Cover, "Broadcast channels," *IEEE Trans. Inform. Theory*, vol. IT-18, pp. 2-14, Jan. 1972.

[36] T. M. Cover and J. A. Thomas, *Elements of information theory*, John Wiley & Sons, 1991.

[37] T. M. Cover and M. Chiang, "Duality between channel capacity and rate distortion with two-sided state information," *IEEE Trans. Inform. Theory*, vol. 48, no. 6, pp. 1629-1638, Jun. 2002.

[38] I. J. Cox, M. L. Miller, and A. L. McKellips, "Watermarking as communications with side information," *Proc. of the IEEE*, vol. 87, no. 7, pp. 1127-1141, Jul. 1999.

[39] I. Csiszár and P. Narayan, "Arbitrary varying channels with constrained inputs and states," *IEEE Trans. Inform. Theory*, vol. 34, no. 1, pp. 27-34, Jan. 1998.

[40] A. Das and P. Narayan, "Capacities of time-varying multiple-access channels with side information," *IEEE Trans. Inform. Theory*, vol. 48, no. 1, pp. 4-25, Jan. 2002.

[41] R. L. Dobrushin, "General formulation of Shannon's main theorem in information theory," *Amer. Math. Soc. Trans.*, vol. 33, pp. 323-438, AMS, Providence, RI, 1963.

[42] F. Dupuis, W. Yu, and F. M. J. Willems, "Blahut-Arimoto algorithms for computing channel capacity and rate-distortion with side information," *Proc. Int. Symp. Inform. Theory (ISIT) 2004*, p. 179, Chicago, IL USA, Jun. 2004.

[43] J. J. Eggers, R. Bäuml, R. Tzschoppe, and B. Girod, "Scalar costa scheme for information embedding," *IEEE Trans. Signal Processing*, vol. 51, no. 4 , pp. 1003-1019, Apr. 2003.

[44] J. J. Eggers, J. K. Su, and B. Girod, "A blind watermarking scheme based on structured code books," *Proc. IEE Conference on Secure Images and Image Authentication*, vol. 4, pp. 1-6, London, U.K., Apr. 2000.

[45] J. J. Eggers, J. K. Su, and B. Girod, "Robustness of a blind image watermarking scheme," *Proc. IEEE International Conference on Image Processing, ICIP-2000*, Vancouver, Canada, Sep. 2000.

[46] J. J. Eggers, J. K. Su, and B. Girod, " Performance of a practical blind watermarking scheme," *Proc. SPIE: Electronic Imaging 2001, Security and Watermarking of Multimedia Contents III*, vol. 4314, San Jose, CA, USA, Jan. 2001.

[47] U. Erez and S. T. Brink, "Approaching the dirty paper limit for canceling known interference," *Proc. 41th Ann. Allerton Conf. On Commun., Control, and Computing*, Oct. 2003.

[48] U. Erez, S. Shamai (Shitz), and R. Zamir, "Capacity and lattice-strategies for canceling known interference," *Proc. Int. Symp. Information Theory and Applications(ISITA) 2000*, pp. 681-684, Honolulu, HI, Nov. 2000.

[49] U. Erez and R. Zamir, "Noise prediction for channels with side information at the transmitter: error exponents," *Proc. Information Theory Workshop*, p. 60, Metsovo, Greece, Jun. 1999.

[50] U. Erez and R. Zamir, "Noise prediction for channel coding with side-information at the transmitter," *IEEE Trans. Inform. Theory*, vol. 46, pp. 1610-1617, Jul. 2000.

[51] U. Erez and R. Zamir, "Error exponents of modulo additive noise channels with side information at the transmitter," *IEEE Trans. Inform. Theory*, vol. 47, pp. 210-218, Jan. 2001.

[52] U. Erez, R. Zamir, and S. Shamai (Shitz), "Additive noise channels with side information at the transmitter," *Proc. 21st IEEE Convention of the Electrical and ELectronic Engineers in Israel, 2000*, pp. 373-376, Apr. 2000.

[53] M. Feder and N. Shulman, "Source broadcasting with unknown amount of receiver side information," *Information Theory Workshop 2002, Proc. 2002 IEEE*, pp. 127-130, Bangalore, India, Oct. 2002.

[54] R. G. Gallager, *Information theory and reliable communication*, New York, NY: John Wiley & Sons, 1968.

[55] A. El Gamal, and T. Weissman, "Source coding with causal side information at the decoder," *Proc. 43rd Annu. Allerton Conf. Communication, Control, and Computing*, Monticello, pp. 28-30, Sep. 2005.

[56] S. I. Gel'fand and M. S. Pinsker, "Coding for channel with random parameters," *Problems of Control and Inform. Theory*, vol. 9, no. 1, pp. 19-31, 1980.

[57] S. I. Gel'fand and M. S. Pinsker, "On Gaussian Channels with Random Parameters," *The Sixth International Symposium on Information Theory* , Abstract of Papers, Part 1, pp. 247-250, Moscow, Tashkent, 1984.

[58] G. Ginis and J. Cioffi, "Vectored-DMT: a FEXT canceling modulation scheme for coordinating users," *Proc. ICC, 2001*, Helsinki, Jun. 2001.

[59] A. J. Goldsmith and M. Médard, "Capacity of time-varying channels with causal channel side information," submitted to *IEEE Trans. Inform. Theory*. available on `www.mit.edu/ medard/m.pdf`.

[60] A. J. Goldsmith and P. P. Varaiya, "Capacity of fading channels with channel side information ," *IEEE Trans. Inform. Theory*, vol. 43, pp. 1986-1992, Nov. 1997.

[61] D. Goldsmith, S. Shamai (Shitz), and Y. Steinberg, "On fading channels with side information at the transmitter," submitted to *IEEE Trans. Inform. Theory*.

[62] C. Heegard, "On the capapcity of permanent memory," *IEEE Trans. Inform. Theory*, vol. IT-31, no. 1, pp. 34-42, Jan. 1985.

[63] C. Heegard and A. El Gamal, "On the capacity of computer memories with defects," *IEEE Trans. Inform. Theory*, vol. IT-29, no. 5, pp. 731-739, Sep. 1983.

[64] N. Jindal and A. Goldsmith, "Capacity and dirty paper coding for gaussian broadcast channels with common information," available on `www.ece.umn.edu/users/nihar/common_info_final.pdf`.

[65] N. Jindal, S. Vishwanath, S. A. Jafar, and A. J. Goldsmith, "Duality, dirty paper coding, and capacity for multiuser wireless channels," *Proc. of the DIMACS workshop on Signal Processing for Wireless Transmission*, Rutgers, NJ, Oct. 2002.

[66] N. Jindal and A. Goldsmith, "Dirty paper coding vs. TDMA for MIMO broadcast channels," submitted to *IEEE Trans. Inform. Theory*, Jun. 2004.

[67] C. Jöngren, M. Skoglund, and B. Ottersten, "Combinning beamforming and orthogonal space-time block coding," *IEEE Trans. Inform. Theory*, vol. 48, pp. 611-627, Mar. 2002.

[68] T. Kalker and F. M. J. Willems, "Capacity bounds and code constructions for reversible data-hiding," *Proc. Electronic Imaging 2003, Security and Watermarking of Multimedia Contents V*, Santa Clara, California, Jan. 2003.

[69] M. Kesal, M. K. Mihcak, R. Koetter and P. Moulin, "Iteratively decodable codes for watermarking applications," *Proc. 2nd Int. Symp. on Turbo Codes and Related Topics*, Brest, France, Sep. 2000.

[70] A. Khisti, U. Erez and G. Wornell, "Writing on many pices of dirty paper at once: the binary case," *Proc. Int. Symp. Inform. Theory (ISIT) 2004*, p. 533, Chicago, IL USA, Jun. 2004.

[71] Y. H. Kim, A. Sutivong and S. Sihurjónsson, "Multiple user writing on dirty paper ," *Proc. Int. Symp. Inform. Theory (ISIT) 2004*, p. 534, Chicago, IL USA, Jun. 2004.

[72] S. Kotagiri and J. N. Lanaman, "Achievable rates for multiple access channels with state information known at one encoder," *Proc. Allerton Conf. Communications, Control, and Computing*, Monticello, IL, Oct. 2004.

[73] S. Kotagiri and J. N. Lanaman, “Reversible information embedding in multi-user channels,” *Proc. 43rd Annu. Allerton Conf. Communication, Control, and Computing*, Monticello, Sep. 2005.

[74] A. V. Kuznetsov, T. Kasami, and S. Yamamura, “An error correcting scheme for defective memory,” *IEEE Trans. Inform. Theory*, vol. IT-24, pp. 712-718, Nov. 1978.

[75] N. V. Kuznetsov and B. S. Tsybakov, “Coding in memories with defective cells,” *Problemy peredachi informatsii*, vol. 10, no. 2, pp. 52-60, 1974.

[76] A. V. Kuznetsov and J. H. Vinck, “On the general defective channel with informed encoder and capacities of some constrained memories,” *IEEE Trans. Inform. Theory*. vol. 40, pp. 1866-1871, Nov. 1994.

[77] A. Lapidoth and P. Narayan, “Reliable communication under channel uncertainty,” *IEEE Trans. Inform. Theory*, vol. 44, no. 6, pp. 2148-2177, Oct. 1998.

[78] A. Lapidoth and S. Shamai (Shitz), “Fading channels: how prefect need ”perfect side information” be?” *IEEE Trans. Inform. Theory*, vol. 48, no. 5, pp. 1118-1134, May 2002.

[79] A. Maor and N. Merhav, “On joint information embedding and lossy compression,” *IEEE Trans. Inform. Theory*, vol. 51, no. 8, pp. 2998-3008, Aug. 2005.

[80] A. Maor and N. Merhav, “On joint information embedding and lossy compression in the presence of a memoryless attack,” *IEEE Trans. Inform. Theory*, vol. 51, no. 9, pp. 3166-3175, Sep. 2005.

[81] E. Martinian, G. W. Wornell, and R. Zamir, “Source coding with distortion side information at the Encoder,” *Proc. Data Compression Conference 2004*, pp. 172-181.

[82] K. Marton, “A coding theorem for the discrete memoryless broadcast channel,” *IEEE Trans. Inform. Theory*, vol. IT-25, pp. 306-311, May 1979.

[83] N. Merhav, “On joint coding for watermarking and encryption,” *IEEE Trans. Inform. Theory*, vol. 52, no. 1, pp. 190-205, Jan. 2006.

[84] N. Merhav and E. Ordentlich, “On causal and semicausal codes for joint information embedding and source coding,” *IEEE Trans. Inform. Theory*, vol. 52, no. 1, pp. 213-226, Jan. 2006.

[85] 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*, vol. 49, no. 11, pp. 2844-2855, Nov. 2003.

[86] N. Merhav and S. Shamai (Shitz), “Information rates subjected to state masking,” submitted to *IEEE Trans. Inform. Theory*, Mar. 2006.

[87] M. L. Miller, “Watermarking with dirty-paper codes,” *Proc. IEEE International Conference on Image Processing*, vol. 2, pp. 528-541, Sep. 2001.

[88] M. L. Miller, G. J. Doerr, and I. J. Cox, “Dirty-paper trellis codes for watermarking,” *Proc. IEEE International Conference on Image Processing*, New York, USA, vol. 2, pp. 633-636, Sep. 2002.

[89] P. Moulin and J. A. O’Sullivan, “Information-theoretic analysis of information hiding,” *IEEE Trans. Inform. Theory*, vol. 49, pp. 563-593, Mar. 2003.

[90] P. Moulin and Y. Wang, “Error exponents for channel coding with side information,” preprint, Sep. 2004. available on `arxiv.org/PS_cache/cs/pdf/0410/0410003.pdf`.

[91] C. T. K. Ng, and A. J. Goldsmith, “Transmitter cooperation in ad-hoc wireless networks: does dirty-paper coding beat relaying,” *IEEE Inform. Theory Workshop*, San Antonio, Texas, Oct. 2004.

[92] C. B. Peel, “On dirty-paper coding,” *IEEE Signal Processing Magazine*, vol. 20, pp. 112-113. May 2003.

[93] T. Philosof, U. Erez, and R. Zamir, “Combined shaping and precoding for interference cancellation at low SNR,” *Proc. Int. Symp. Inform. Theory (ISIT) 2003*, Yokohama, Japan, p. 68, Jul. 2003.

[94] T. Philosof, U. Erez, and R. Zamir, “Precoding for interference cancellation at low SNR,” *Proc. 22nd Convention of Electrical and Electronics Engineers in Israel, 2002*, pp. 144-147, Dec. 2002.

[95] S. S. Pradhan, J. Chou, and K. Ramchandran, “Duality between source coding and channel coding and its extension to the side information case,” *IEEE Trans. Inform. Theory*, vol. 49, pp. 1181-1203, May 2003.

[96] S. S. Pradhan, and K. Ramchandran, “Distributed source coding using syndromes (DISCUS): design and construction,” *IEEE Trans. Inform. Theory*, vol. 49, no. 3, pp. 626-643, Mar. 2003.

[97] R. Puri, K. Ramchandran, and S. S. Pradhan, “On seamless digital upgrade of analog transmission systems using coding with side information,” *40th Allerton Conference on Communication, Control and Computing, Allerton*, IL, Oct. 2002.

[98] T. Rappaport, “Wireless communications,” Engelewood Cliffs, NJ: Prentice-Hall, 1996.

[99] A. Rosenzweig, Y. Steinberg, and S. Shamai (Shitz), “On channels with partial channel state information at the transmitter,” *IEEE Trans. Inform. Theory*, vol. 51, no. 5, pp. 1817-1830, May 2005.

[100] A. Sabharwal, E. Erkip, and B. Aazhang, “On channel state information in multiple antenna block fading channels,” *Proc. Int. Symp. Inform. Theory (ISIT) 2000*, pp. 116-119, Hawaii, Nov. 2000.

[101] M. Salehi, “Capacity and coding for memories with real-time noisy defect information at encoder and decoder,” *IEE Proceedings -1*, vol. 139, No. 2. pp. 113-117. Apr. 1992.

[102] H. Sato, “An outer bound on the capacity region of broadcast channels,” *IEEE Trans. Inform. Theory*, vol. IT-24, pp. 374-377, May 1978.

[103] S. Shamai (Shitz), S. Verdú, and R. Zamir, “Systematic lossy source/channel coding,” *IEEE Trans. Inform. Theory*, vol. 44, no. 2, pp. 564-579, Mar. 1998.

[104] S. Shamai (Shitz) and B. M. Zaidel, “Enhancing the cellular capacity via co-processing at the transmitting end,” *Vehicular Technology Conference 2001, IEEE VTS 53rd*, Greece, vol. 3, pp. 1745-1749, 2001.

[105] C. E. Shannon, “Channels with side information at the transmitter,” *IBM Journal Research and Development*, vol. 2, pp. 289-293, Oct. 1958.

[106] S. Sigurj*ó*nsson and Y. H. Kim, “On multiple user channels with state information at the transmitters,” *Proc. Int. Symp. Inform. Theory (ISIT) 2005*, pp. 72-76, Adelaide, Sep. 2005.

[107] D. Slepian and J. K. Wolf, “Noiseless coding of correlated information sources,” *IEEE Trans. Inform. Theory*, vol. IT-19, pp. 471480, Jul. 1973.

[108] A. Somekh-Baruch and N. Merhav, “On the capacity game of public watermarking systems,” *IEEE Trans. Inform. Theory*, vol. 50, no. 3, pp. 511-524, Mar. 2004.

[109] A. Somekh-Baruch and N. Merhav, “On the error exponent and capacity games of private watermarking systems,” *IEEE Trans. Inform. Theory*, vol. 49, no. 3, pp. 537-562, Mar. 2003.

[110] Y. Steinberg, “Coding for degraded broadcast channel with random parameters, with causal and non-causal side information,” *IEEE Trans. Inform. Theory*, vol. 51, no. 8, pp. 2867-2877, Aug. 2005.

[111] Y. Steinberg, “Coding for channels with rate-limited side information at the decoder, with applications,” submitted to *IEEE Trans. Inform. Theory*, Feb. 2006.

[112] Y. Steinberg, “Coding for channels with rate-limited side information at the decoder,” *Proc. Inform. Theory Workshop (ITW) 2006*, pp. 11-13, Punta Del Este, Uruguay, Mar. 2006.

[113] Y. Steinberg, “Reversible information embedding with compressed host at the decoder,” *Proc. Int. Symp. Inform. Theory (ISIT) 2006*, Seattle, Washington, USA, Jul. 2006.

[114] Y. Steinberg and N. Merhav, “On hierarchical joint source-channel coding,” *IEEE Trans. Inform. Theory*, vol. 52, no. 3, pp. 886-903, Mar. 2006.

[115] Y. Steinberg and S. Shamai (Shitz), “Achievable rates for the broadcast channel with states known at the transmitter,” *Proc. Int. Symp. Inform. Theory (ISIT) 2005*, pp. 2184-2188, Adelaide, Sep. 2005.

[116] J. K. Su, J. J. Eggers, and B. Girod, “Channel coding and rate distortion with side information: Geometric interpretation and illustration of duality,” submitted to *IEEE Trans. Inform. Theory*, May 2000.

[117] J. K. Su, J. J. Eggeres, and B. Girod, “Illustration of the duality between channel coding and rate distortion with side information,” *34th Asilomar Conf. on Signals, Systems, and Computers*, Asilomar, CA, USA, Oct. 29-Nov. 1, 2000.

[118] A. Sutivong, M. Chiang, T. M. Cover, and Y. H. Kim, “Channel capacity and state estimation for state-dependent Gaussian channels,” *IEEE Trans. Inform. Theory*, vol. 51, no. 4, pp. 1486-1495, Apr. 2005.

[119] A. Sutivong, T. M. Cover, and M. Chiang, “Tradeoff between message and state information rates,” *Proc. Int. Symp. Inform. Theory (ISIT) 2001*, Washington, DC, p. 303, Jun. 2001.

[120] A. Sutivong, T. M. Cover, M. Chiang, and Y. H. Kim, “Rate vs. distortion trade-off for channels with state information,” *Proc. Int. Symp. Inform. Theory (ISIT) 2002*, Lausanne, Switzerland, p. 226, Jun. 2002.

[121] M. Tomlinson, “New automatic equalizer employing modulo aritmetic,” *Electronic lett.*, vol. 7, pp. 138-139, 1971.

[122] B. S. Tsybakov, “Additive group codes for defect correction,” *Problemy Peredachi Informatsii*, vol. 11, no. 1, pp. 111-113, 1975.

[123] B. S. Tsybakov, “Defects and error correction,” *Problemy Peredachi Informatsii*, vol. 11, no. 3, pp. 21-30, 1975.

[124] Z. Tu and R. S. Blum, “Multiuser diversity for a dirty paper approach,” *IEEE Communications letters*, vol 7, no. 8, pp. 370-372, Aug. 2003.

[125] S. Verdú and T. S. Han, "A general formula for channel capacity," *IEEE Trans. Inform. Theory*, vol. 40, pp. 1147-1157, Jul. 1994.

[126] S. Vishwanath, N. Jindal, and A. Goldsmith, "Duality, achievable rates, and sum-rate capacity of gaussian MIMO broadcast channels," *IEEE Trans. Inform. Theory*, vol. 49, pp. 2658-2668, Oct. 2003.

[127] P. Viswanath and D. N. C. Tse, "Sum capacity of the vector gaussian broadcast channel and uplink-downlink duality," *IEEE Trans. Inform. Theory*, vol. 49, pp. 1912-1921, Aug. 2003.

[128] B. Vojcic and W. M. Jang, "Transmitter precoding in synchrounous multiuser communications," *IEEE Trans. Commun.*, vol. 46, no. 10, pp. 1346-1355, Oct. 1998.

[129] H. Wang and P. Viswanath, "Fixed binning schemes for channel and source coding problems: an operational duality," preprint, Sep. 2003. available on `www.ifp.uiuc.edu/ pramodv/talks/ciss_hua_talk.pdf`.

[130] H. Weingarten, Y. Steinberg and S. Shamai (Shitz), "The capacity region of the gaussian mimo broadcast channel," submitted to *IEEE Trans. Inform. Theory*, 2005.

[131] T. Weissman and N. Merhav, "Coding for the feedback Gel'fang-Pinsker channel and the feedforward Wyner-ziv source," submitted to *IEEE Trans. Inform. Theory*, Sep. 2005.

[132] F. M. J. Willems, "Signalling for the gaussian channel with side information at the transmitter," *Proc. Int. Symp. Inform. Theory (ISIT) 2000*, Sorrento, Italy, p. 348, Jun. 2000.

[133] F. M. J. Willems and T. Kalker, "Reversible embedding methods," *40th Allerton Conference on Communication, Control and Computing, Allerton*, IL, Oct. 2002.

[134] A. Winshtok and Y. Steinberg, "Join source-channel coding for arbitrarily varying Wyner-Ziv source and Gel'fand-Pinsker channel," *Proc. 44th Allerton Conference*, Illinois, USA, Sep. 2006.

[135] A. D. Wyner and J. Ziv, "The rate-distortion function for source coding with side information at the decoder," *IEEE Trans. Inform. Theory*, vol. IT-22, no. 1, pp. 1-10, Jan. 1976.

[136] A. D. Wyner, "The rate-distortion function for source coding with side information at the decoder-II: General sources," *Inform. Contr.*, vol. 38, pp. 60-80, 1978.

[137] Z. Xiong, A. D. Liveris, and S. Cheng, "Distributed source coding for sensor networks," *IEEE Signal Processing Magazine*, vol. 21, no. 5, pp. 80-94, Sep. 2004.

[138] Z. Xiong, V. Stankovic, S. Cheng, A. Liveris, and Y. Sun, "Source-channel coding for algebraic multiterminal binning," *Proc. ITW'04*, San Antonio, TX, Oct. 2004.

[139] W. Yu and J. M. Cioffi, "Trellis precoding for the broadcast channel," *Proc. IEEE GlobeCom 2001*, pp. 1344-1348, Nov. 2001.

[140] W. Yu and John M. Cioffi, "Sum capacity of Gaussian vector broadcast channels," *IEEE Trans. Inform. Theory*, vol. 50, no. 9, pp. 1875-1892, Sep. 2004.

[141] W. Yu, A. Sutivong, D. Julian, T. M. Cover, and M. Chiang, “Writing on colored paper,” *Proc. Int. Symp. Inform. Theory (ISIT) 2001*, p. 302, Washington, DC, 2001.

[142] W. Yu, D. P. Varodayan, and J. M. Cioffi, “Trellis and convolutional precoding for transmitter-based interference pre-subtraction,” *IEEE Trans. Commun.*, vol. 53, no. 7, pp. 1220-1230, Jul. 2005.

[143] R. Zamir, S. Shamai (Shitz), and U. Erez, “Nested linear/lattice codes for structured multiterminal binning,” *IEEE Trans. Inform. Theory*, vol. 48, no. 6, pp. 1250-1276, Jun. 2002.