Tech. Report EE-S-97-10: Capacity and Coding for Symmetric Channels with Side Information at the Transmitter

Uri Erez and Ram Zamir
Dept. of Elect. Eng. - Systems, Faculty of Engineering,
Tel Aviv University, Tel Aviv 69978, ISRAEL

October 5, 1997

Abstract

The capacity of a discrete memoryless channel (DMC) $P_S(y|x)$, with (random) state information $S \in \mathcal{S}$ available at the transmitter, is given by the capacity of an equivalent DMC with an extended input space $\mathcal{X}^{|\mathcal{S}|}$ commonly interpreted as a space of strategies. This well known solution, due to Shannon, is in general hard to compute and apply. We consider the special case of a symmetric (modulo-additive noise) channel $Y = X + Z_S$, where the transmitter (but not the receiver) has access to the state S that governs the distribution of the noise Z_S. For this channel we find a concise symmetric form for the equivalent DMC: $Y = \widetilde{X} + t(S) + Z_S$, where $t : \mathcal{S} \to \mathcal{X}$ is a function to be optimized. Consequently, the capacity is given by $C = \log|\mathcal{X}| - \min_{t:\mathcal{S}\to\mathcal{X}} H(t(S) + Z_S)$, and the optimal transmitter has a simple modular structure, consisting of a conventional (i.e., state independent) encoder at rate C followed by a memoryless "shifting" device $t^*(S)$. We further generalize our result to channels whose state sequence has memory.

Key words: Side information, channel capacity, optimum transmitter, minimum entropy prediction.

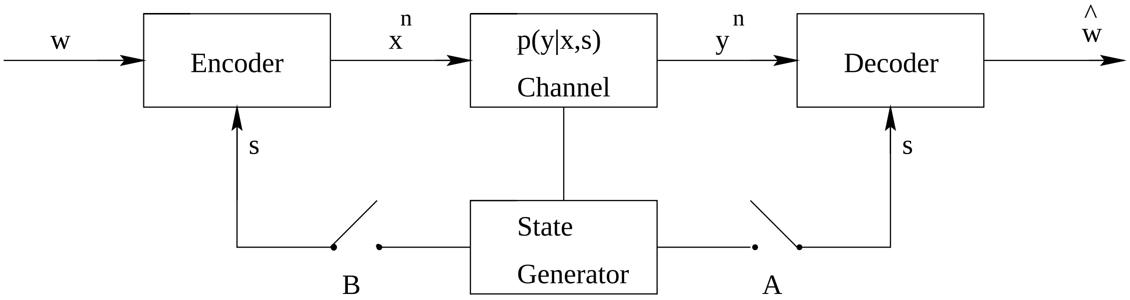

Figure 1: Channel configuration

1 Introduction

We begin with a brief review of discrete memoryless channels with side information , along with the appropriate definitions. Consider the channel depicted in Fig.1 . The channel is discrete with input alphabet $\mathcal{X}$, output alphabet $\mathcal{Y}$, and state alphabet $\mathcal{S}$, all of which are finite sets. The channel states are i.i.d. with distribution p(s) and independent of the input sequence. Furthermore, given the states, the channel is memoryless with transition distribution $p(y|x,s)$. Hence, the conditional distribution of $\mathbf{y} = y_1 \dots y_n$ and $\mathbf{s} = s_1 \dots s_n$ given $\mathbf{x} = x_1 \dots x_n$ can be written as:

$$p(\mathbf{y},\mathbf{s}|\mathbf{x}) = p(\mathbf{s})p(\mathbf{y}|\mathbf{x},\mathbf{s}) = \prod_{i=1}^{n} p(s_i)p(y_i|x_i,s_i).$$

The state sequence in this model plays the role of 'side information' (SI). The respective states of switches A,B in Fig. 1 give rise to the following distinct cases:

Case I - SI not available: In this case the channel is regarded as an ordinary DMC with transition probability

$$p(y|x) = \sum_s p(s)p(y|x,s).$$

We denote the resulting capacity by $C_{NO\ SI}$.

Case II - SI available at receiver only (A closed,B open): The capacity , which we denote by $C_{SI@REC}$, is defined in the ordinary Shannon sense, with the the decoding function being a function of the state vector as well as the received output vector, i.e. $\hat{W} = g(\mathbf{Y}, \mathbf{S})$, where $W \in \{1 \ldots 2^{nR}\}$ is the message sent, $\hat{W} \in \{1 \ldots 2^{nR}\}$ is the decoded message, and R is the transmission rate in bit per channel use. The capacity is given by (see e.g. [8])

$$C_{SI@REC} = \max_{p(x)} I(X; S, Y) = \max_{p(x)} I(X; Y|S). \quad (1)$$

Case III - SI available at transmitter and receiver alike(A,B closed): With the suitable definition of a code for this case, the capacity is given by (see e.g. [8])

$$C_{SI@BOTH} = \sum_{s \in \mathcal{S}} p(s) \max_{p(x|s)} I(X; Y|S = s) \triangleq \sum_{s \in \mathcal{S}} p(s) C_s \quad (2)$$

where C_s is the capacity at state s.

Case IV - SI available at transmitter only (A open,B closed): The encoder maps the message $w \in \{1, 2, ...2^{nR}\}$ into $\mathcal{X}^n$ using functions

$$x_i = f_i(w, \mathbf{s}^i) \quad 1 \le i \le n$$

where $\mathbf{s}^i = (s_1, s_2, ...s_i)$ is the state information at the encoder before the ith transmission. Namely, the encoder operates *causally* with respect to the state sequence. The receiver decodes the message W from the received vector $\mathbf{y}$ as $\hat{W} = g(\mathbf{y})$, where $\mathbf{y}$ denotes the *whole* received sequence $\mathbf{y}^n$. This constitutes a code. The notions of achievable rates of transmission and capacity are defined analogously with the ordinary DMC.

Shannon [7] showed that this capacity ,which we denote by $C_{SI@TR}$, is equal to the regular capacity of the derived DMC shown in Fig.2. The input alphabet of the derived channel, denoted $\mathcal{T}$, is the set of all possible mappings

$$t : \mathcal{S} \longrightarrow \mathcal{X}$$

which we refer to as *strategies* or *strategy functions*. We may describe each strategy $t_j(s) \in \mathcal{T}$ by the $|\mathcal{S}|$-tuple $(x_j^1, x_j^2, \ldots, x_j^{|\mathcal{S}|})$, i.e., $t_j(s) = x_j^s$ for

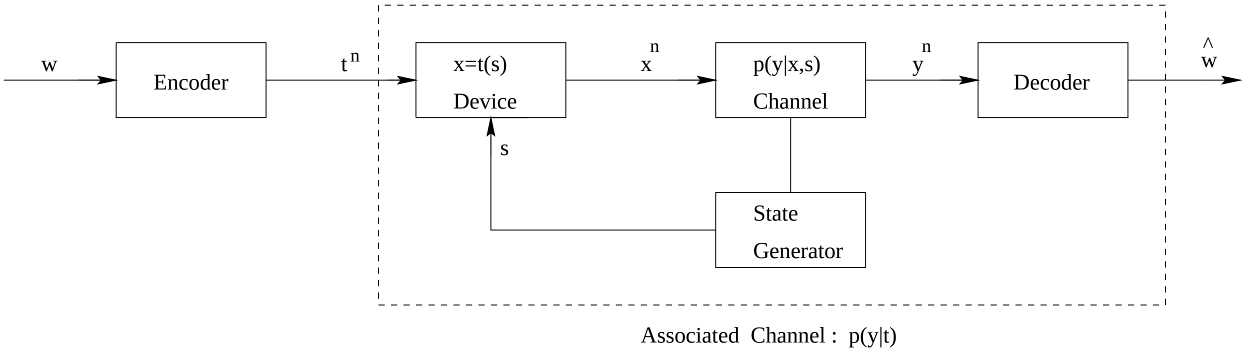

Figure 2: Associated channel

$s = 1, \ldots, |\mathcal{S}|$. Therefore $|\mathcal{T}| = |\mathcal{X}|^{|\mathcal{S}|}$. The device shown in Fig.2, called "transducer", just takes t and s as inputs and produces x as its output by computing $x = t(s)$. Therefore the boxed section constitutes a DMC defined by

$$p(y|t) = \sum_s p(s)p\Big(y\Big|x = t(s), s\Big) \tag{3}$$

and also

$$p(\mathbf{y}|\mathbf{t}) = \prod_{i=1}^{n} p(y_i|t_i). \tag{4}$$

Thus, the capacity with SI at the transmitter is given by, [7],

$$C_{SI@TR} = \max_{p(t)} I(T;Y) \tag{5}$$

where the maximization is taken over the distribution $p(t)$ of the random strategy variable $T \in \mathcal{T}$.

Two remarks are in order:

1. The strategies employed by the derived channel are functions of the *current* state alone. The operational meaning of this structure is that it is possible to achieve capacity using a code of the form $x_i = f(w, s_i)$.

2. At most $|\mathcal{Y}|$ of the strategies need be given positive probability in order to achieve capacity. Therefore if the cardinalities of $\mathcal{X}$ and $\mathcal{Y}$ are equal

(in particular if $\mathcal{X} = \mathcal{Y}$) then at most $|\mathcal{X}|$ of the strategies need be given positive probability.

In principle, then, the derived channel can be reduced to a $\mathcal{Y}$-input / $\mathcal{Y}$-output channel. However, in the general case one does not know in advance which of the strategies are to be used to achieve capacity. In the next section we treat a class of channels, namely, the class of symmetric channels, for which it is easy to identify this set of 'active' strategies. This will allow us to find a simple expression for the capacity $C_{SI@TR}$, and to characterize the structure of an optimal transmitter. In Section 3 we discuss the properties of our solution and illustrate it with a few examples, and in Section 4 we generalize our result to a class of symmetric channels with memory.

The subject of channels with SI at the transmitter is rarely discussed in the literature. Jelinek [3, 4] extended Shannon's work to channels with memory. Salehi [5] considered a certain mixture of Cases II - IV above. Shamai and Lapidoth [6] consider a continuous setting with a power constraint.

2 Additive Noise Channels and the Main Result

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

$$Y = X + Z$$

where Z is statistically independent of X, and addition (and in the sequel subtraction) is understood to be performed modulo-$|\mathcal{X}|$. Denoting by Z_s the distribution of the noise given the side information we have

$$p_Z(z) = \sum_s p(s) p_{z_s}(z).$$

Here the SI may be viewed either as the channel "state" or as a "noisy version" of the additive noise Z.

We shall now show that finding the capacity in this setting reduces to finding the shift function $t(s)$ which minimizes the entropy of $Z + t(S)$. More explicitly, we seek to find the vector $(\alpha_1, \ldots, \alpha_{|\mathcal{S}|})$ which minimizes the entropy of the distribution $\tilde{p}(z) = \sum_s p(s) p_{Z_s}(z - \alpha_s)$. Denoting this minimum entropy by H_{min}, we shall prove that the capacity is given by

$$C_{SI@TR} = \log |\mathcal{X}| - H_{min}.$$

Before stating this result formally in Theorem 1 below, we need to make a few definitions. We partition the set $\mathcal{T}$ into

$$\frac{|\mathcal{T}|}{|\mathcal{X}|} = |\mathcal{X}|^{|\mathcal{S}|-1}$$

classes as follows:

Definition (class of strategies): *Two strategies, t_i and t_j, belong to the same class iff $t_i(s) - t_j(s) = const.$, i.e. iff their difference (modulo-$|\mathcal{X}|$) is independent of s.*

Clearly, this definition induces a partition of $\mathcal{T}$ into $|\mathcal{T}|/|\mathcal{X}|$ classes, each class composed of $|\mathcal{X}|$ elements, corresponding to the the $|\mathcal{X}|$ possible values of *const.*. Denote these classes $\mathcal{T}_k, k = 1, \ldots, |\mathcal{X}|^{|\mathcal{S}|-1}$. Arbitrarily choosing some strategy $t_{k,0} \in \mathcal{T}_k$, we may index the strategies within the class $\mathcal{T}_k$ by

$$t_{k,j}(s) = t_{k,0}(s) + j \quad\quad j = 1, \ldots, |\mathcal{X}|. \tag{6}$$

Example: Let $\mathcal{X} = \{1, 2, 3, 4\}$ and $\mathcal{S} = \{1, 2\}$. Then the set of all strategies, $\mathcal{T} = \{1, 2, 3, 4\} \times \{1, 2, 3, 4\}$, is partitioned into the four classes:

$$\mathcal{T}_0 = \{(1,1), (2,2), (3,3), (4,4)\}$$

$$\mathcal{T}_1 = \{(1,2), (2,3), (3,4), (4,1)\}$$

$$\mathcal{T}_2 = \{(1,3), (2,4), (3,1), (4,2)\}$$

$$\mathcal{T}_3 = \{(1,4), (2,1), (3,2), (4,3)\}$$

where the first strategy in each class is $t_{k,0}$, the second is $t_{k,1}$, the third is $t_{k,2}$, and the last is $t_{k,3}$, for $k = 0, 1, 2, 3$.

Due to the additivity of the original channel we have

$$p\Big(y|x = t(s) + c, s\Big) = p\Big(y - c|x = t(s), s\Big) = p_{Z_s}\Big(y - t(s) - c\Big) \quad \forall s \in \mathcal{S}. \quad (7)$$

Averaging over s, we have for the derived channel defined in (3)

$$p(y|t + c) = p(y - c|t) \tag{8}$$

where $t + c$ should be read as the strategy function $g(s) = t(s) + c \;\; s \in \mathcal{S}$. Define the random variable $\widetilde{Z}_k = Z + t_{k,0}(S)$ whose distribution satisfies

$$p_{\widetilde{Z}_k}(y) = \sum_s p(s) p_{Z_s}(y - t_{k,0}(s)) = p(y|t_{k,0}). \tag{9}$$

By (6), (8) and (9) we have

$$p(y|t_{k,j}) = p(y - j|t_{k,j} - j) = p(y - j|t_{k,0}) = p_{\widetilde{Z}_k}(y - j). \tag{10}$$

Thus the output conditional distributions, $p(y|t_{k,j})$, $j = 1, \ldots, |\mathcal{X}|$, differ in a shift alone. It follows that if we restrict the strategies employed at the input of the derived channel of Fig.2 to some single class $\mathcal{T}_k$, and identify the input alphabet letter $t_{k,j}$ with the output letter j, then we may view the resulting channel as an *additive noise channel* with noise $\tilde{Z}_k$. We can write this channel as

$$Y = \widetilde{X} + \widetilde{Z}_k = \widetilde{X} + t_{k,0}(S) + Z \tag{11}$$

where the channel input $\widetilde{X}$, representing the index $j \in \mathcal{X}$ of the strategy within $\mathcal{T}_k$, is independent of (S, Z). Hence $H(Y|t)$ is constant within each class, $H(Y|t) = H(\widetilde{Z}_k)$ for $t \in \mathcal{T}_k$, and putting a uniform distribution within any single class but zero elsewhere results in a uniform distribution on $\mathcal{Y}$, i.e., $H(Y) = \log|\mathcal{X}|$.

Let us define

$$H_{min} \triangleq \min_{t \in \mathcal{T}} H(Y|t) \triangleq \min_k H(\widetilde{Z}_k) \quad \text{and} \quad \mathcal{T}^* = \arg\min H(Y|t)$$

i.e., $\mathcal{T}^*$ is the class whose strategies achieve H_{min} (if there is more than one such class then choose one arbitrarily).

Theorem 1 *In the additive noise channel the capacity $C_{SI@TR}$ is achieved using strategies belonging solely to the class $\mathcal{T}^*$. Furthermore, within this class all the elements are given equal probability, and*

$$C_{SI@TR} = \log|\mathcal{X}| - H_{min}. \tag{12}$$

Proof: Since

$$H(Y) \le \log|\mathcal{X}| \quad \text{and} \quad H(Y|t) \ge \min_{t\in\mathcal{T}} H(Y|t) = H_{min},$$

we have $I(T;Y) = H(Y) - H(Y|T) \le \log|\mathcal{X}| - H_{min}$ for any distribution on $\mathcal{T}$. Choosing T having a uniform distribution on $\mathcal{T}^*$ and zero elsewhere, results in equality in both of the above inequalities. Thus the theorem follows from (5). ■

We define $\widetilde{Z}$ to be the induced minimum entropy noise, i.e. $\tilde{Z} \triangleq Z + t_{min}(S)$, where t_{min} is any member of $\mathcal{T}^*$. The function $t_{min}(s)$ may be viewed as an optimal "shifting device". Since the receiver does not know S, it sees a channel with an effective noise $\widetilde{Z} = Z + t_{min}(S)$. Shifting Z by $t_{min}(S)$ thus makes the effective noise the least harmful for the un-informed receiver. Note also that since $I(S;\widetilde{Z}) = H(\widetilde{Z}) - H(\widetilde{Z}|S) = H(\widetilde{Z}) - H(Z|S)$, minimizing $H(\widetilde{Z})$ is equivalent to minimizing the information carried by the effective noise about the channel state.

Another useful interpretation for the function t_{min} follows from the structure of an optimal transmitter with side information. The transmitter consists of an ordinary (i.e., state independent) Shannon code for a symmetric channel, followed by the shifting device $t_{min}(s)$. Specifically, suppose a code $\mathcal{C} = \{(c_{w1}, \ldots, c_{wn})\}_{w=1}^{2^{nR}}$ is drawn at random, using a uniform distribution over $\mathcal{X}$, with rate $R < C_{SI@TR}$. Given a message $1 \le w \le 2^{nR}$, and a channel state s_i, the transmitter outputs at time instant i

$$x_i = c_{wi} + t_{min}(s_i) \qquad i = 1 \dots n.$$

The receiver decodes w from $\mathbf{Y}$ as if the channel was an ordinary symmetric DMC with additive noise $\widetilde{Z}$. It then follows from Theorem 1 and the forward channel coding theorem [1], that the decoding is reliable with high probability. In practice, we can replace the random code by any conventional "good" code for the additive noise channel.

3 Properties of the Optimal Shifting Function $t_{min}(s)$ and Examples

We make a short digression to present some results obtained in [2]. Elias arrives at this very solution in his treatment of entropy-coded predictive coding. Specifically, for a discrete stationary ergodic source X_i, one wishes to find the (prediction) function $f(x_{i-1}, x_{i-2}, \dots)$ such that the entropy of the error, $e_i \triangleq x_i - f(x_{i-1}, x_{i-2}, \dots)$, is minimized. This will minimize the rate assuming that e_i is encoded according to its *marginal* distribution (ignoring the possible residual memory in the sequence $e_1, e_2 \dots$). The analogy to the problem we study is clear. The past source samples have the role of the current state, and the conditional random variable $X_i | x_{i-1}, x_{i-2}, \dots$ replaces Z_s. With this done, the problems become identical.

In [2] the concept of an ordered average was introduced. Consider a set of n discrete distributions, each of m letters. Denote by $p_{i,j}$ the probability of the jth letter in the ith distribution. Let w_i be some averaging weight on the distributions:

$$w_i \geq 0 \qquad \sum_{i=1}^{n} w_i = 1$$

Define the ordered distribution $\tilde{p}_{i,j}$ of the ith distribution to be the set of probabilities $p_{i,j}$ indexed according to decreasing probability. The ordered average is defined by

$$p_j^{ord} = \sum_{i=1}^{n} w_i \tilde{p}_{i,j}$$

Thus p_1^{ord} is formed by taking w_1 times the largest probability in the first distribution, w_2 times the largest probability in the second distribution, and so on. It can be shown that the entropy of the ordered average distribution is less than that of any other average formed from the same distributions with the same weights, but with the terms of one or more of the distributions not arranged in order of decreasing probability.

Generally, the optimal shifts $\{\alpha_s\}$, depend on the probabilities of the corresponding states. Nevertheless, following Elias we shall now identify two simple cases of interest in which the optimal shifts are independent of $p(s)$:

a) p_{Z_s} is a symmetric and unimodal distribution for each s.

b) p_{Z_s} is zero to the left(right) of a point, $x_0(s)$, and is monotonically nonincreasing to the right(left) of $x_0(s)$:

$$p_{Z_s}(x) = 0 \quad for\ x < x_0(s)$$

$$p_{Z_s}(x) \geq p_{Z_s}(y) \quad for\ x_0(s) \leq x < y.$$

In both cases the assertion follows from the observation that choosing $\{\alpha_i\}$ so as to align the maxima together yields an ordered average irrespective of $p(s)$. We refer to these strategies as *state-weight independent strategies.*

Figure 3 depicts the strategies employed in order to achieve capacity (for the case $|\mathcal{S}| = 2$) , viewing the strategies as points as suggested by the vector interpretation given above.

Before presenting a few examples, we return to Cases I-IV (of availability of SI) in the context of symmetric channels. Since choosing a uniform

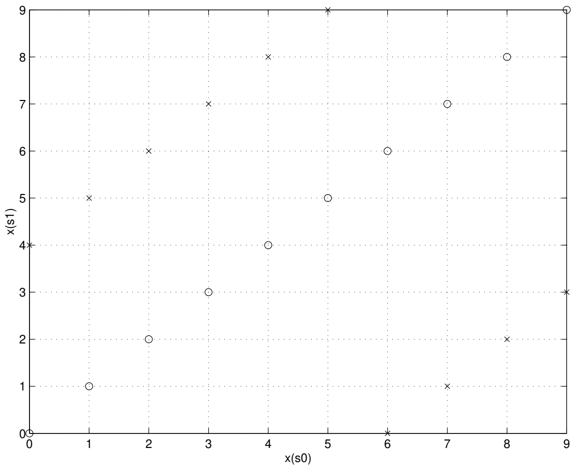

Figure 3: o : Strategies used when no SI is available. x : Strategies used when SI is available at the transmitter: $|\mathcal{S}| = 2 \quad \mathcal{X} = \{0, 1 \ldots 9\} \quad \alpha_0 = 0$, $\alpha_1 = 4$

distribution for X achieves capacity for any additive noise channel, we have for Cases II and III

$$C_{SI@REC} = C_{SI@BOTH} = \log|\mathcal{X}| - \sum_s p(s)H(Z_s) = \log|\mathcal{X}| - H(Z|S) \quad (13)$$

and for Case I

$$C_{NO\ SI} = \log|\mathcal{X}| - H(Z). \quad (14)$$

Comparing (12), (13) and (14)) we obtain

$$\begin{aligned} C_{SI@BOTH} - C_{SI@REC} &= 0 \\ C_{SI@BOTH} - C_{SI@TR} &= I(S;\widetilde{Z}) \\ C_{SI@BOTH} - C_{NOSI} &= I(S;Z). \end{aligned} \quad (15)$$

Also the following chain of inequalities holds:

$$C_{NOSI} \leq C_{SI@TR} \leq C_{SI@REC} = C_{SI@BOTH} \quad (16)$$

Note that by the convexity of the entropy

$$H(\widetilde{Z}) \geq \sum_s p(s)H(Z_s - \alpha_s) = \sum_s p(s)H(Z_s) = H(Z|S) \quad (17)$$

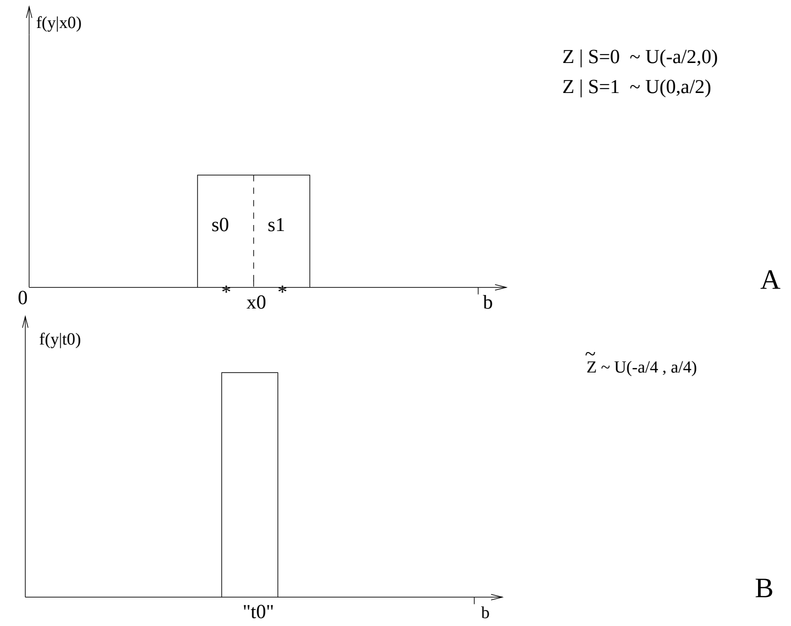

Figure 4: Example III: "t_0" means we can envisage using strategy t_0 as sending $x = x_0$ in a channel with "effective noise" $\tilde{Z}$

with equality iff the distribution of $Z_s - \alpha_s$ is independent of s. Furthermore, $H(\widetilde{Z}|S) = H(Z + t(S)|S) = H(Z|S)$. Thus we have equality in the second inequality in (16), i.e., $C_{SI@REC} = C_{SI@TR}$, iff the distributions of Z_s $s \in \mathcal{S}$ differ in a shift only, in which case $\widetilde{Z}$ is statistically independent of S. We have equality in the first inequality in (16), i.e. $C_{SI@TR} = C_{NOSI}$, iff $H(\widetilde{Z}) = H(Z)$, i.e., iff the optimal shifts,$\{\alpha_s\}$, are the set of zero shifts (or any set of constant shifts).

We turn to consider a few examples. We first note that the results of Section 2 hold equally well when the alphabet $\mathcal{X}$ is a continuous interval $[0, a] \in \mathcal{R}^1$, with addition performed modulo a.

Example I: A Gaussian noise channel $Z_\sigma \sim N(0, \sigma^2)$ $\;a \gg \sigma$. The channel state is the noise variance, who has some arbitrary distribution. This is an example of a case in which $C_{SI@TR} = C_{NO\ SI}$. It follows from the fact that Z is an ordered average of the set of distributions corresponding to $\{Z_\sigma\}$.

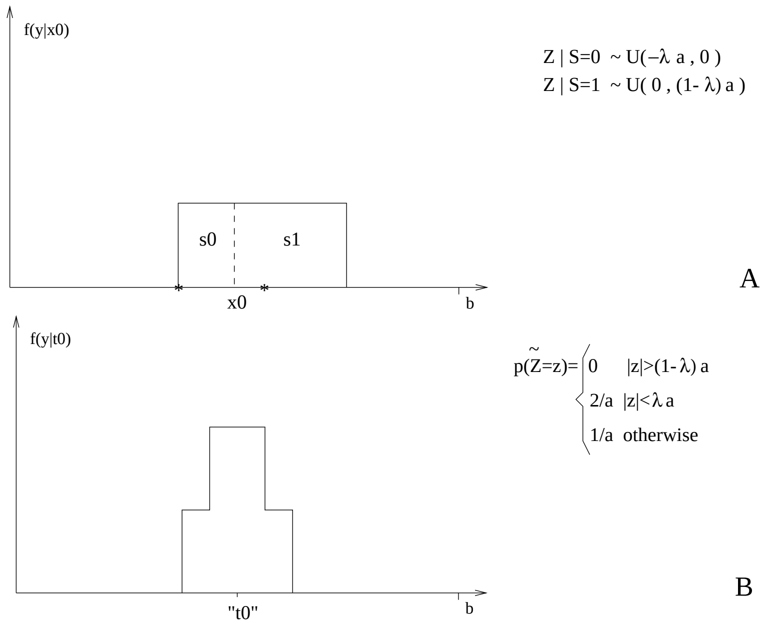

Figure 5: Example IV

Example II BSC with crossover probability θ_s. Here

$$C_{SI@REC} = 1 - \sum_s p(s) H_b(\theta_s)$$

where $H_b(\cdot)$ denotes the binary entropy function. In the case where the SI consists of two states, θ_1 θ_2, we have:

$$C_{SI@TR} = \begin{cases} C_{NOSI}, & \text{if } \theta_1, \theta_2 \leq \frac{1}{2} \\ 1 - H_b(p(s=1)\theta_1 + p(s=2)\bar{\theta_2}), & \text{if } \theta_1 \leq \frac{1}{2} \leq \theta_2 \end{cases} \tag{18}$$

The $\tilde{Z}$ thus obtained is an ordered average. The extension of the result to the general case (more than two states) is obvious: We obtain an ordered average by ”flipping” each θ_i that is greater than $\frac{1}{2}$.

Example III: Consider the channel shown in Fig.4 (a) We have :

$$C_{SI@TR} = C_{SI@REC} = \log |\mathcal{X}| - H(Z|S = s_1) = \log b - log\frac{a}{2} = \log \frac{b}{a} + 1$$

The corresponding channel with noise $\tilde{Z}$ is shown in Fig.4 (b).

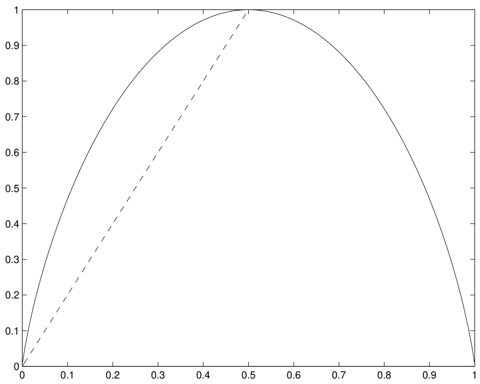

Figure 6: $C_{SI@REC} - C_{SI@TR} = H_b(\lambda) - 2\lambda$.

Example IV: Consider the channel shown in Fig.5 (a)

$$H(Z|S) = p(s_1)H(Z|S=s_1)+p(s_0)H(Z|S=s_0) = \lambda \log \lambda a+(1-\lambda)\log(1-\lambda)a = \log a - H_b(\lambda)$$

$$H(\tilde{Z}) = -\int_0^{\lambda a} \frac{2}{a}\log\frac{2}{a} - \int_0^{(1-2\lambda)a} \frac{1}{a}\log\frac{1}{a} = -2\lambda\log\frac{2}{a} - (1-2\lambda)\log\frac{1}{a} = \log a - 2\lambda$$

Therefore :

$$C_{SI@REC} = \log\frac{b}{a} + H_b(\lambda) \qquad C_{SI@TR} = \log\frac{b}{a} + 2\lambda$$

The difference between the two (i.e. $H_b(\lambda) - 2\lambda$) is depicted in Fig. 6. The corresponding channel with noise $\tilde{Z}$ is shown in Fig.5(b).

4 Capacity for Markov State Processes

Let the channel of Section 1 be given with the modification that the state process forms an irreducible aperiodic Markov chain of order one. These conditions ensure the existence of a well defined capacity, independent of the initial state S_0. We now review the results obtained for this class of channels by Jelinek [3]. We have

$$p(y_1^n, s_1^n | x_1^n, s_0) = \prod_{i=1}^{n} p(s_i|s_{i-1}) \prod_{i=1}^{n} p(y_i|x_i, s_i) \ .$$

We define the following associated channel of order m. Let $\mathcal{T}$ denote the set of all $|\mathcal{X}|^{|\mathcal{S}|^m}$ functions, called strategies

$$t : \mathcal{S}^m \longrightarrow \mathcal{X} \ .$$

Using these strategy letters we construct as in the previous section, via a transducer as in Fig.2, an associated channel of order m. The input alphabet is letters $t \in \mathcal{T}$ and the output is in $\mathcal{Y}$. In addition to t, the transducer views the current and the last $m-1$ letters of the state sequence. Upon receiving, at time instant i, the input $t_i \in \mathcal{T}$ the transducer sends as output

$$x_i = t_i(s_i, s_{i-1}, \ldots, s_{i-m+1}).$$

We thus construct the associated channel (with no side information) K^m, with transition probability law

$$p(y^n|t^n) = \sum_{s^n_{-m+2}} p^*(s^n_{-m+2}) p\Big(y^n \Big| x^n = t^n(s^n_{-m+2}), s^n_{-m+2}\Big)$$

$$= \sum_{s^n_{-m+2}} p^*(s_{-m+2}) \prod_{j=-m+3}^{n} p(s_j|s_{j-1}) \prod_{i=1}^{n} p\Big(y_i \Big| x_i = t_i(s^i_{i-m+1}), s_i\Big)$$

where $p^*(\cdot)$ denotes the stationary distribution of the state sequence. We define

$$C_{mn} = \max_{p(t^n)} \frac{1}{n} I(T^n; Y^n)$$

where $p(t^n)$ is the distribution of the strategy n-sequence T^n. Thus $\lim_{n\to\infty} C_{mn}$ is the capacity of the channel K^m. Jelinek [3] showed that the capacity of the original channel with SI at the transmitter is equal to the limit of the capacity of K^m as $m \to \infty$, i.e.,

$$C_{SI@TR} = \lim_{m\to\infty} \lim_{n\to\infty} C_{mn} \ . \tag{19}$$

We now specialize to the additive noise case. The channel is now given by

$$Y_i = X_i + Z_i \quad \text{where} \quad \Pr(z_1^n, s_1^n | x_1^n, s_0) = \prod_{i=1}^{n} p(s_i|s_{i-1}) \Pr(z_i|s_i). \tag{20}$$

Note that the additive noise satisfies a hidden Markov model (relative to the state sequence).

We shall now partition the set of n-sequences of strategies $\mathcal{T}^n = \{t^n\}$ into classes of constant difference. Given an n-sequence of strategies $t^n = t^n(s^n) = [t_1(s_{2-m}^1), \ldots, t_n(s_{n-m+1}^n)]$, define the class of strategies

$$\mathcal{T}_{t^n}^n = \{(t_1 + c_1, \ldots, t_n + c_n) : \; c_1, \ldots, c_n \in \mathcal{X}\},$$

where as before $t_i + c$ means adding a constant c to the the strategy function $t_i(s_{i-m+1}^i)$. Let $f^n, g^n \in \mathcal{T}^n$. It is easily seen that

a) If $g^n \in \mathcal{T}_{f^n}^n$ then $\mathcal{T}_{f^n}^n = \mathcal{T}_{g^n}^n$

b) If $g^n \notin \mathcal{T}_{f^n}^n$ then $\mathcal{T}_{f^n}^n \cap \mathcal{T}_{g^n}^n = \emptyset$

c) $\mathcal{T}_{f^n}^n = \mathcal{T}_{g^n}^n$ iff $f^n - g^n = c^n$, c^n being a vector of constants (i.e., not a function of s^n).

Now, $|\mathcal{T}^n| = (|\mathcal{X}|^{|\mathcal{S}|^m})^n = |\mathcal{X}|^{(n|\mathcal{S}|^m)}$ and there are $|\mathcal{X}|^n$ members in each class. Similarly to (8), by the additivity of the noise we have for any $c^n \in \mathcal{X}^n$

$$p(y^n|T^n = t^n + c^n) = p(y^n - c^n|T^n = t^n).$$

Hence the conditional entropy $H(Y^n|t^n)$ is constant within each class. Let us denote

$$H_{min}^{mn} = \frac{1}{n} \min_{t^n \in \mathcal{T}^n} H(Y^n|t^n) \qquad \text{and} \qquad \mathcal{T}^{n*} = \arg\min H(Y^n|t^n).$$

Lemma 1 *In the additive noise case*

$$C_{mn} = log|\mathcal{X}| - H_{min}^{mn}.$$

Proof: The proof is similar to the proof of Theorem 1, with the obvious modifications. ■

Let t^n_{min} be any member of $\mathcal{T}^{n*}$. Like in the memoryless case, finding C_{mn} amounts to finding the optimal shifts minimizing the (joint) entropy of the 'effective noise' $\widetilde{Z}^{mn} \stackrel{\Delta}{=} Z^n + t^n_{min}(S^n)$.

As a consequence of Lemma 1 and (19) we have the following theorem:

Theorem 2 *For the additive noise channel (20) define H_{min} by*

$$H_{min} \stackrel{\Delta}{=} \lim_{m\to\infty} \lim_{n\to\infty} H^{mn}_{min}.$$

Then

$$C_{SI@TR} = \log|\mathcal{X}| - H_{min}.$$

Discussion

As is well known, the capacity of a stationary channel can be achieved by a stationary input distribution []. Clearly, for any m the derived channel K^m above is stationary (assuming of course the stationary distribution over the $m-1$ initial states). Thus, it is tempting to suggest that for every m, as n becomes large, the optimal strategy sequence t^n_{min} is composed of repeating n times the *same time invariant function* $t(s^m)$, i.e.,

$$t^n_{min} = \Big(t(s_1 \ldots s_{2-m}), t(s_2 \ldots s_{3-m}), \ldots, t(s_n \ldots s_{n-m+1})\Big).$$

Unfortunately, however, this simplifying conclusion may not be correct, since even if $t_i(s^i)$ *varies* with time i, the process $t_1(S^1), t_2(S^2), \ldots$ may be "stationarized", e.g. by means of random initialization.

An even "longer shot" would be to ask whether the optimal strategy function can be memoryless, i.e.,

$$t^n_{min} = \Big(t(s_1), \ldots, t(s_n)\Big).$$

There is a reason to believe that this is indeed the case if the optimum shifting function relative to $\{p(y|x,s)\}$ is a state-weight independent strategy (see Section 3). But this is also an open question at the moment.

References

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

[2] P. Elias. Predictive coding. *IRE Trans. Inform. Theory*, pages 16–33, MARCH 1955.

[3] F. Jelinek. Indecomposible channels with side information at the transmitter. *Inform. Control*, 8:36–55, 1965.

[4] F. Jelinek. Determination of capacity achieving input probabilities for a class of finite state channels with side information. *Inform. Control*, 9:101–129, 1966.

[5] M. Salehi. Capacity and coding for memories with real-time noisy defect information at encoder and decoder. *IEE Proceedings I*, 139:113–117, 1992.

[6] S. Shamai and A. Lapidoth. Private communication. 1997.

[7] C.E. Shannon. Channels with side information at the transmitter. *IBM J*, 2:289–293, 1958.

[8] J. Wolfowitz. *Coding theorems of information theory.* Springer Verlag, 1964.