CCIT Report #580 March 2006

Information Rates Subjected to State Masking

Neri Merhav and Shlomo Shamai (Shitz)
EE Dept., Technion – I.I.T., Haifa 32000, Israel
Email: [merhav,sshlomo]@ee.technion.ac.il

***Abstract*— We consider the problem of rate–R channel coding with causal/non–causal side information at the transmitter, under an additional requirement of minimizing the amount of information that can be learned from the channel output about the state sequence, which is defined in terms of the equivocation E (i.e., the mutual information between the state sequence and the channel output sequence). A single–letter characterization is provided for the achievable region of pairs $\{(R, E)\}$. Explicit results for the Gaussian case (Costa's dirty–paper channel) are derived in full detail.**

I. Introduction

The problem of information transfer via state–dependent channels is classical (see [9] for a partial review). One of the most interesting models is the case where the channel states are available at the transmitter either causally or non–causally. This framework has been fully characterized for i.i.d. states in famous studies by Shannon [14] and by Gel'fand and Pinsker (G–P) [7], repectively. These models, and in particular the G–P setting, have gained much interest in the last few years, mainly due to the wide scope application areas, such as watermarking, [3], [10], [12], [15], [11], multi–input–multi–output (MIMO) broadcast channels, [1], [2], network [8] and cooperative networks, [6], just to name a few applications.

One of the most interesting and well known examples the the G–P channel is the Gaussian setting where the states impact the channel additively. The surprising result by Costa [4] demonstrates that no loss in capacity is suffered no matter how strong that independent interfering state sequence is. Evidently, the many applications and the challenge here motivated much work in terms of actual coding strategies that come close to the optimum. These coding strategies (see, e.g., [21] and references therein), build on the insight of random binning which is the central mechanism in showing achievability in this problem [7], and can, in fact, be interpreted as practical binning strategies. In the Gaussian channel, nick–named ''dirty-paper" [4], efficient techniques based on modern codes were recently reported as well (see [16] and references therein). Source–channel coding aspects in the framework of state–dependent channel of this type are also considered [13], and the source–channel separation principle has been shown valid in various scenarios, in which the model itself is intimately related to the Wyner–Ziv (W–Z) source coding problem with side information at the decoder [20], and the G–P channel [7].

While in models addressed in [13], the source and channel states are assumed independent, this is not always the case. In some applications, the channel–state process is not inherently channel–related (like in fading), but may rather be an information–bearing signal on its own. The MIMO broadcast channel serves as a typical example, where a state sequence for one user is just the information–carrying sequence for another, and all produced at the same transmitter who addresses both users simultaneously [1]. In fact, these are exactly the cases where the justification to the non–causality is self–evident, as the transmitter controls the state sequence. The state sequence is often modelled as i.i.d. whether it is a specific codeword of a good codebook operating on a memoryless channel, which essentially mimics an i.i.d., or it is i.i.d., and it represents raw data, as say a systematic part of the information [13]. Furthermore, the state sequence may model also analogue information which is conveyed over the same channel with an overlayed digital part. This sort of applications gave rise to an interesting problem addressed by Sutivong *et. al.* [17], [18], where the role of the transmitter is two–fold: to transmit independent reliable information on the one hand, and to boost the quality of the state estimator at the receiver, which adopts a prescribed distortion measure, on the other. A coding scheme has been suggested in [18], which combines W–Z coding, based on the side information about the state available at the receiver side, and G–P coding which is used to convey the independent reliable rate, as well as the W–Z coded information. In the Gaussian case, it has been verified that this achievable tradeoff is in fact optimal [19]. In this specific case, a simple technique where the transmitter optimally power–shares between pure information transmission via the Costa strategy and simple state amplification achieves the optimal tradeoff.

In this paper, we focus on another aspect of the problem. The state sequence is referred to as undesired information that leaks to the receiver. It indeed could model a leakage in the system of, say, secret analogue (sampled) information, or stand for a codeword which is not intended to that receiver and is therefore to be concealed from the receiver side. Thus, the goal of the transmitter now is to try and mask this undesired information as much as possible on the one hand, and to transmit reliable independent data rate on the other. The amount of information that the receiver retrieves about the state sequence is measured by the blockwise mutual information (equivocation), as is customary in measuring the security of the cipher systems, in the literature of the Shannon theory. This measure guarantees that even if there is coding involved, only a small value of the associated mutual information limits the reliable information that the non–intended receiver can retrieve about the state sequence.

We characterize the tradeoff between the reliably transmit-

ted rate and the masking ability of the state information, and that is in both the G–P and Shannon models, namely, where the state sequence is either available non–causally or causally respectively. We characterize, explicitly and completely, the tradeoff for the additive Gaussian example, and notice that also in this setting, an element of state cancellation (de–amplification) is optimal. In some cases, excess reliable rate can be transmitted at no cost to the masking ability.

II. Notation and Problem Formulation

Throughout this paper, scalar RVs will be denoted by capital letters, their sample values will be denoted by the respective lower case letters, and their alphabets will be denoted by the respective calligraphic letters. A similar convention will apply to random vectors and their sample values, which will be denoted with same symbols superscripted by the dimension. Thus, for example, X^n will denote a random n-vector $(X_1, ..., X_n)$, and $x^n = (x_1, ..., x_n)$ is a specific vector value in $\mathcal{X}^n$, the n-th Cartesian power of $\mathcal{X}$. The notations x_i^j and X_i^j, where i and j are integers and $i \leq j$, will designate segments $(x_i, \ldots, x_j)$ and $(X_i, \ldots, X_j)$, respectively, where for $i = 1$, the subscript will be omitted (as above). For $i > j$, x_i^j (or X_i^j) will be understood as the null string. Sequences without specifying indices are denoted by $\{\cdot\}$. Sources and channels will be denoted generically by the letter P or Q. Information theoretic quantities like entropies, and mutual informations will be denoted following the usual conventions of the information theory literature, e.g., $H(X^n)$, $I(S^n; Y^n)$, etc. Differential entropy will be denoted by h, e.g., $h(S^n)$.

Consider the Gel'fand–Pinsker (G–P) discrete memoryless channel (DMC) G–P channel

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

where $\{x_i\}$ are the transmitted symbols, taking on values in a finite alphabet $\mathcal{X}$, $\{s_i\}$ are the corresponding channel states, taking values in a finite state set $\mathcal{S}$, and drawn from a discrete memoryless source (DMS),

$$Q(s^n) = \prod_{i=1}^{n} Q(s_i),$$

and $\{y_i\}$ are the corresponding channel outputs, taking on values in a finite output alphabet $\mathcal{Y}$. The channel input signal is subjected to a limitation

$$\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{E}\{\phi(X_i)\} \leq \Gamma, \tag{1}$$

where $\phi : \mathcal{X} \to \mathbb{R}^+$ is the transmission cost function and $\Gamma > 0$ is a given constant. Let $w \in \mathcal{W} = \{0, 1, \ldots, 2^{nR}-1\}$ denote (the index of) an nR–bit digital message, R being the coding rate, to be conveyed via the channel. The random variable W that designates the message is uniformly distributed across $\mathcal{W}$ independently of S^n. We assume that the encoder (also referred to as the transmitter) is non–causally aware of the state sequence s^n, and it transmits an input vector x^n, which is a (possibly stochastic) function of w and s^n. A rate–R encoder for n-blocks is therefore characterized by a conditional probability distribution $P(x^n|s^n, w)$, which maintains the channel input limitation (1), w.r.t. the randomness of S^n and W as well as the possible randomness of the transmitter itself. The corresponding decoder maps the channel output y^n to $\hat{w} \in \mathcal{W}$, and the probability of error P_e is defined as $\Pr\{\hat{W} \neq W\}$.

We are interested in the interplay between reliable coding at rate R, which would like to keep as large as possible, and an equivocation level, $I(S^n; Y^n)/n$, which we would like to make as small as possile. For a given $\Gamma > 0$, a pair (R, E) is called *achievable* if for every $\epsilon > 0$ and sufficiently large n, there exist a rate–R encoder–decoder for n-blocks such the following conditions are simultaneously satisfied:

1) $\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{E}\{\phi(X_i)\} \leq \Gamma$
2) $P_e \leq \epsilon$
3) $\frac{1}{n} I(S^n; Y^n) \leq E + \epsilon$.

The achievable region $\mathcal{A}$ is the set of all achievable pairs $\{(R, E)\}$.

Our main goal is to provide a single–letter characterization of $\mathcal{A}$ as well as some insights on good coding schemes. We also show how our coding theorem should be modified to the case where the transmitter has causal, rather than non–causal, access to the side informaton.

III. The Zero–Rate Case

For the sake of simplicity of the exposition, we begin with zero–rate case, i.e., $R = 0$, and then our only goal is to minimize $I(S^n; Y^n)/n$ subject to (1).

Let $\mathcal{F}(\Gamma)$ denote the minimum of $I(S^n; Y^n)/n$ over all channels $\{P(x^n|s^n)\}$ that satisfy (1). Define also the single–letter function $F(\Gamma) = \min I(S; Y)$, where the minimum is over all $\{P(x|s)\}$ s.t. $E\phi(X) \leq \Gamma$. Our first theorem is the following:

Theorem 1: $\mathcal{F}(\Gamma) = F(\Gamma)$.

Proof. As for the direct part, apply the DMC

$$P^*(x^n|s^n) = \prod_{i=1}^{n} P^*(x_i|s_i),$$

where the single–letter channel $P^*(x|s)$ achieves $F(\Gamma)$. Since the induced channel $P(y^n|s^n)$ will be a DMC as well, and $I(S_i, Y_i) = F(\Gamma)$ for all i, then so will be $I(S^n; Y^n)/n = \frac{1}{n}\sum_{i=1}^{n} I(S_i; Y_i)$.

Turning now to the converse part, we first observe that $F(\Gamma)$ is convex. This is very easy to see in the very same manner as the classical informational rate–distortion is shown to be convex [5]: The mutual information $I(S^n; Y^n)$ is a convex functional of $\{P(y^n|s^n)\}$, which is a linear functional of $\{P(x^n|s^n)\}$, which in turn is subjected to the power

constraint, which is linear. Thus,

$$\begin{aligned} I(S^n;Y^n) &\geq \sum_{i=1}^{n} I(S_i;Y_i) \\ &\geq \sum_{i=1}^{n} F(\boldsymbol{E}\phi(X_i)) \\ &\geq nF\left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{E}\phi(X_i)\right) \\ &\geq nF(\Gamma), \end{aligned} \tag{2}$$

where the first inequality is by the memorylessness of S^n, the second is by definition of F, the third by convexity, and the fourth by monotonicity. This completes the proof of Theorem 1. •

It is interesting to observe that in the zero–rate case considered here, the optimum transmitter works in a single–letter (scalar) fashion, i.e., no long blocks are needed. This means that the solutions to the causal and non-causal problems coincide in the zero–rate case. It also means that the solution is strictly optimum and not only asymptotically so. As we shall see, this will no longer be true for positive rates. We next study the example of the Gaussian inteference channel in some detail.

Example. Consider the Gaussian inteference channel

$$Y = X + S + Z, \tag{3}$$

where S and Z are zero–mean independent Gaussian RV's with variances σ_s^2 and σ_z^2, respectively. We would like to characterize the optimum conditional distribution $P^*(x|s)$. Since

$$I(S;Y) = h(S) - h(S|Y), \tag{4}$$

and $h(S)$ is given, minimization of $I(S;Y)$ is equivalent to maximization of $h(S|Y)$. Now, for a given $\sigma_x^2 \leq \Gamma$, and $\rho = \boldsymbol{E}(XS)/(\sigma_x\sigma_s)$, we have:

$$\begin{aligned} h(S|Y) &= h(S - \boldsymbol{E}(S|Y)|Y) \\ &\leq h(S - \boldsymbol{E}(S|Y)) \\ &\leq \frac{1}{2}\log\left[2\pi e\cdot\boldsymbol{E}(S-\boldsymbol{E}(S|Y))^2\right] \\ &\leq \frac{1}{2}\log\left[2\pi e\cdot\min_a \boldsymbol{E}(S-aY)^2\right] \\ &= \frac{1}{2}\log\left[2\pi e(\sigma_s^2 - \sigma_{\hat{s}}^2)\right] \end{aligned} \tag{5}$$

with

$$\sigma_{\hat{s}}^2 = \frac{(\sigma_s^2 + \rho\sigma_x\sigma_s)^2}{\sigma_s^2 + 2\rho\sigma_x\sigma_s + \sigma_x^2 + \sigma_z^2}, \tag{6}$$

which is the variance of the optimum linear estimator of S based on Y. The last inequality is due to the fact the MSE of the optimum linear estimator of S is never smaller than the MSE of the optimum estimator, which is the conditional mean. As is easily see, all inequalities become equalities if (X,S) are jointly Gaussian. It remains then to minimize $\sigma_{\hat{s}}^2$ w.r.t. (σ_x^2,ρ) over the rectangle $[0,\Gamma]\times[-1,1]$. First observe that whenever $\Gamma \geq \sigma_s^2$, the solution is trivially $X = -S$. Assume then that $\Gamma < \sigma_s^2$. Minimizing $\sigma_{\hat{s}}^2$ is equivalent to maximizing $G \stackrel{\Delta}{=} 1/\sigma_{\hat{s}}^2$, which is given by

$$\begin{aligned} G &= \frac{\sigma_s^2 + 2\rho\sigma_x\sigma_s + \sigma_x^2 + \sigma_z^2}{(\sigma_s^2 + \rho\sigma_x\sigma_s)^2} \\ &= \frac{2(\sigma_s^2 + \rho\sigma_x\sigma_s) + \sigma_x^2 + \sigma_z^2 - \sigma_s^2}{(\sigma_s^2 + \rho\sigma_x\sigma_s)^2} \\ &= \frac{2}{\sigma_s^2 + \rho\sigma_x\sigma_s} + \frac{\sigma_x^2 + \sigma_z^2 - \sigma_s^2}{(\sigma_s^2 + \rho\sigma_x\sigma_s)^2} \\ &\stackrel{\Delta}{=} \frac{2}{t} + \frac{\sigma_x^2 + \sigma_z^2 - \sigma_s^2}{t^2}, \end{aligned} \tag{7}$$

where for a given σ_x^2, we have the freedom to maximize A over t in the range $[\sigma_s(\sigma_s - \sigma_x), \sigma_s(\sigma_s + \sigma_x)]$. First, observe that $\sigma_x^2 = \Gamma$ is always the optimum choice – this choice both maximizes the numerator of the second term, and broadens the range of allowable values of u as much as possible. Let us set then $\sigma_x^2 = \Gamma$. Moving on to the maximization w.r.t. t, the derivative $\partial G/\partial t$, is given by

$$\frac{\partial G}{\partial t} = -\frac{2(t + \sigma_x^2 + \sigma_z^2 - \sigma_s^2)}{t^3}, \tag{8}$$

which vanishes at

$$t = \sigma_s^2 - \sigma_x^2 - \sigma_z^2. \tag{9}$$

This means that

$$\rho = -\frac{\Gamma + \sigma_z^2}{\sqrt{\Gamma}\sigma_s}, \tag{10}$$

which can be the case only if

$$\sigma_s \geq \sqrt{\Gamma} + \frac{\sigma_z^2}{\sqrt{\Gamma}}, \tag{11}$$

otherwise, $\rho = -1$. To summarize then, the solution divides into three cases, according to the intensity of the interference, S:

- *Weak interference:* If $\sigma_s^2/\Gamma \leq 1$, then take $X = -S$, and then $F(\Gamma) = 0$.
- *Moderate interference:* If

$$1 < \frac{\sigma_s^2}{\Gamma} \leq \left(1 + \frac{\sigma_z^2}{\Gamma}\right)^2, \tag{12}$$

then

$$X = -\sqrt{\frac{\Gamma}{\sigma_s^2}}\cdot S, \tag{13}$$

and then

$$F(\Gamma) = \frac{1}{2}\log\left[1 + \left(\frac{\sigma_s}{\sigma_z} - \frac{\sqrt{\Gamma}}{\sigma_z}\right)^2\right]. \tag{14}$$

- *Strong interference:* If

$$\frac{\sigma_s^2}{\Gamma} > \left(1 + \frac{\sigma_z^2}{\Gamma}\right)^2, \tag{15}$$

then

$$X = -S \cdot \left(\frac{\Gamma}{\sigma_s^2} + \frac{\sigma_z^2}{\sigma_s^2} \right) + V, \tag{16}$$

where V is a zero–mean Gaussian RV, independent of S, with variance

$$\Gamma \left[1 - \left(\frac{\sqrt{\Gamma}}{\sigma_s} + \frac{\sigma_z}{\sqrt{\Gamma}} \cdot \frac{\sigma_z}{\sigma_s} \right)^2 \right], \tag{17}$$

and in this case,

$$F(\Gamma) = \frac{1}{2} \log \left(1 + \frac{A}{B} \right), \tag{18}$$

where

$$A = \sigma_s^2 \left[1 - \frac{\Gamma}{\sigma_s^2} \left(1 + \frac{\sigma_z^2}{\Gamma} \right) \right]^2 \tag{19}$$

and

$$B = \sigma_z^2 + \Gamma \left[1 - \frac{\sqrt{\Gamma}}{\sigma_s} \left(1 + \frac{\sigma_z^2}{\Gamma} \right) \right]^2. \tag{20}$$

IV. The Positive Rate Case

Turning now to the more general positive rate case, our main result is the following:

Theorem 2: $\mathcal{A}$ 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) $U \to (X, S) \to Y$ is a Markov chain.
2) $\boldsymbol{E}\phi(X) \le \Gamma$.
3) $R \le I(U;Y) - I(U;S)$.
4) $E \ge I(S;U,Y)$.

Proof. We begin with the converse part. The channel–coding part is exactly as in [7], except that here, we present it slightly differently in order to establish the fact the same random variable U that meets the rate requirement, also meets the equivocation requirement and the power constraint. For $i = 1, \ldots, n$, let $U_i = (W, Y^{i-1}, S_{i+1}^n)$. Define a RV T, uniformly distributed over $\{1, 2, \ldots, n\}$ (independently of the other RV's), and let $U \stackrel{\Delta}{=} (U_T, T)$. We also define the RV's $Y = Y_T$, $S = S_T$, and

$$\delta(\epsilon) = \epsilon \log \epsilon - (1-\epsilon)\log(1-\epsilon) + \epsilon R,$$

for $\epsilon \in [0, 1]$. Now,

$$\begin{aligned}
R - \delta(\epsilon) &\le R - \delta(P_e) \\
&\le \frac{1}{n} \sum_{i=1}^n [I(U_i;Y_i) - I(U_i;S_i)] \\
&= I(U_T;Y_T|T) - I(U_T;S_T|T) \\
&= I(U_T,T;Y_T) - I(T;Y_T) - \\
&\quad I(U_T,T;S_T) + I(T;S_T) \\
&\le I(U_T,T;Y_T) - I(U_T,T;S_T) + I(T;S_T) \\
&= I(U_T,T;Y_T) - I(U_T,T;S_T) \\
&= I(U;Y) - I(U;S),
\end{aligned} \tag{21}$$

where the first inequality is by the requirement that $P_e \le \epsilon$, the second is as in [7, Proposition 3, Lemma 4], and in the second to the last equality we have used the fact that $S = S_T$ is independent of T (due to stationarity). As for the equivocation, we have the following:

$$\begin{aligned}
I(S^n;Y^n) &= I(S^n;Y^n,W) - I(S^n;W|Y^n) \\
&\ge H(S^n) - H(S^n|Y^n,W) - H(W|Y^n) \\
&\ge \sum_{i=1}^n [H(S_i) - H(S_i|S_{i+1}^n, Y^n, W)] - n\delta(P_e) \\
&\ge \sum_{i=1}^n [H(S_i) - H(S_i|Y_i, S_{i+1}^n, Y^{i-1}, W)] \\
&\quad -n\delta(P_e) \\
&\ge \sum_{i=1}^n [H(S_i) - H(S_i|Y_i, U_i)] - n\delta(\epsilon) \\
&= n[H(S_T|T) - H(S_T|Y_T, U_T, T)] - n\delta(\epsilon) \\
&= n[H(S) - H(S|Y,U) - \delta(\epsilon)] \\
&= n[I(S;Y,U) - \delta(\epsilon)],
\end{aligned} \tag{22}$$

where the second inequality is by Fano's inequality, and where we have used again the fact that S_T is independent of T. The channel input constraint is maintained by definition of X_T. Finally, note that due to the stationarity of the memoryless channel $P(y|x,s)$, the Markov relation $U \to (X,S) \to Y$ is maintained (and is not violated by the presence of the RV T).

Regarding the direct part, consider the ordinary construction of the G–P code using binning. Reliable decoding is proved exactly as in [7]. The power constraint is maintained by joint typicality considerations. As for the equivocation, first, we have the following:

$$\begin{aligned}
& I(S^n;Y^n) \le I(S^n;U^n,Y^n) \\
= \;& I(S^n;U^n) + I(S^n;Y^n|U^n) \\
= \;& I(S^n;U^n) + H(Y^n|U^n) - H(Y^n|S^n,U^n)
\end{aligned} \tag{23}$$

The first term is bounded as follows:

$$\begin{aligned}
I(S^n;U^n) &\le I(S^n;W,U^n) \\
&= I(S^n;U^n|W) \\
&\le H(U^n|W) \\
&\le n[I(U;S) + \epsilon],
\end{aligned} \tag{24}$$

where the equality is due to the independence between S^n and W, and the last inequality is due to the fact that the size of each bin is less than $2^{n[I(U;S)+\epsilon]}$. As for the second term on the right–most side of (23), we have:

$$H(Y^n|U^n) \le \sum_{i=1}^n H(Y_i|U_i) = nH(Y|U), \tag{25}$$

where we have used the fact that the empirical distribution of each codeword U^n is according to the desired choice of the distribution of U and that each Y_i is generated from U_i according to

$$P(y|u) = \sum_{s,x} P(s|u)P(x|u,s)P(y|x,s). \tag{26}$$

As for the third term on the right–most side of (23):

$$H(Y^n|S^n,U^n)=\sum_{i=1}^{n}H(Y_i|S_i,U_i)=nH(Y|S,U), \quad (27)$$

where the first equality is due to the memorylessness of the channel $P(y|u,s)$ (which is the cascade of the memoryless channel $P(x|u,s)$ and the memoryless channel $P(y|x,s)$), and the second equality is explained similarly as before. Thus, we obtain:

$$\begin{aligned} I(S^n;Y^n) &\leq n[I(U;S)+\epsilon]+nH(Y|U)-nH(Y|S,U) \\ &= n[I(S;Y,U)+\epsilon]. \end{aligned} \quad (28)$$

The power constraint is maintained by joint typicality considerations. This completes the proof of Theorem 2. •

A few comment are in order: First, as the auxilary RV U, includes the time variable T, the achievable region is convex. Second, the cardinality of the alphabet of U is by two letters larger than in ordinary G–P coding because of the additional equivocation and power constraints. Finally, note that here, unlike the pure G-P coding, the channel $P(x|u,s)$ is not necssarily deterministic: For example, in the Gaussian case with $R=0$ that was studied earlier, U is degenerate, but $P(x|u,s)=P(x|s)$ is non–deterministic in the case of very strong interference.

We next revisit the Gaussian example, this time, for positive rates. One of the interesting points in this example is that it turns out that the same RV U that maximizes the information rate, $I(U;Y)-I(U;S)$ (as in Costa's channel) turns out to minimize the equivocation $I(S;Y,U)$ and bring it to the level of $I(S;Y)$. In other words, U does not improve on the estimation of S once Y is observed (even in the single–letter level).

Example – Gaussian channel revisited. First, it should be noted, that similarly as in [4], here too, Theorem 2 extends to continuous alphabets by taking limits of $I(U;Y)-I(U;S)$ and $I(S;Y,U)$ over sequences of successively refined partitions of the alphabets $\mathcal{U}$, $\mathcal{S}$ and $\mathcal{Y}$. As before, the actual input power $\sigma_x^2 \stackrel{\Delta}{=} \boldsymbol{E}(X^2)$ will be assumed less than or equal to Γ. However, observe that in case of $R>0$, the best choice of σ_x^2 is always $\sigma_x^2=\Gamma$, because the part of the power of X that may not be needed to cancel S (when $\sigma_s^2<\Gamma$), is always fully utilized to convey digital information. Thus, σ_x^2 and Γ are two notations for the same entity, in this example.

Proposition 1: Let $Y=X+S+Z$, where S is zero–mean with variance σ_s^2, $Z\sim\mathcal{N}(0,\sigma_z^2)$ is independent of X and S, $\boldsymbol{E}(X^2)=\sigma_x^2$, and $\boldsymbol{E}(XS)=\rho\sigma_s\sigma_x$. Further, let U be an RV that satisfies the Markov relation $U\rightarrow(X,S)\rightarrow Y$. Then,

$$I(U;Y)-I(U;S)\leq\frac{1}{2}\log\left[1+\frac{\sigma_x^2(1-\rho^2)}{\sigma_z^2}\right]. \quad (29)$$

Proof. Let $\tilde{X}=X-aS$, where aS stands for the best linear estimator of X given S, that is, $a=\rho\sigma_x/\sigma_s$. Thus, Y can be represented as

$$Y=\tilde{X}+(1+a)S+Z, \quad (30)$$

where $\tilde{X}$ is uncorrelated with S, and $\boldsymbol{E}(\tilde{X}^2)=\sigma_x^2(1-\rho^2)$. Now,

$$\begin{aligned} & I(U;Y)-I(U;S) \\ \leq\ & I(U;Y,S)-I(U;S) \\ =\ & I(U;Y|S) \\ \leq\ & I(X,S;Y|S) \\ =\ & I(\tilde{X},S;Y|S) \\ =\ & I(\tilde{X};\tilde{X}+Z|S) \\ =\ & h(\tilde{X}+Z|S)-h(\tilde{X}+Z|S,\tilde{X}) \\ \leq\ & h(\tilde{X}+Z)-h(\tilde{X}+Z|\tilde{X}) \\ \leq\ & \frac{1}{2}\log\left[2\pi e\left(\sigma_x^2(1-\rho^2)+\sigma_z^2\right)\right]-\frac{1}{2}\log(2\pi e\sigma_z^2) \\ =\ & \frac{1}{2}\log\left[1+\frac{\sigma_x^2(1-\rho^2)}{\sigma_z^2}\right], \end{aligned} \quad (31)$$

where the second inequality is due to the Markov relation $U\rightarrow(X,S)\rightarrow Y$ and the data processing theorem, the following equality is due to the fact that the transformation from (X,S) to $(\tilde{X},S)$ is one–to–one, and the following inequality is due to the fact the conditioning reduces entropy and the fact that $\tilde{X}+Z$ is independent of S given $\tilde{X}$ (since Z is independent of both $\tilde{X}$ and S). This completes the proof.•

Proposition 2: Let $Y=X+S+Z$, where $S\sim\mathcal{N}(0,\sigma_s^2)$, $Z\sim\mathcal{N}(0,\sigma_z^2)$ is independent of X and S, $\boldsymbol{E}(X^2)=\sigma_x^2$, and $\boldsymbol{E}(XS)=\rho\sigma_s\sigma_x$. Further, let U be an RV that satisfies the Markov relation $U\rightarrow(X,S)\rightarrow Y$. Then,

$$I(S;Y,U)\geq\frac{1}{2}\log\frac{\sigma_s^2}{\sigma_s^2-\sigma_{\hat{s}}^2}, \quad (32)$$

where $\sigma_{\hat{s}}^2$ is as in eq. (6).

Proof. $I(S;Y,U)\geq I(S;Y)$ and the rest is like in the Gaussian example for $R=0$.•

Corollary 1: Let

$$R<\frac{1}{2}\log\left(1+\frac{\sigma_x^2}{\sigma_z^2}\right),$$

$$\varrho(R)=\sqrt{1-(2^{2R}-1)\frac{\sigma_z^2}{\sigma_x^2}},$$

and let $E(\varrho)$, $\varrho\geq 0$, denote the minimum of

$$\frac{1}{2}\log\frac{\sigma_s^2}{\sigma_s^2-\sigma_{\hat{s}}^2}$$

with $\sigma_{\hat{s}}^2$ as in eq. (6), as a function of ρ across the interval $[-\varrho,+\varrho]$. Then, for the channel $Y=X+S+Z$, and a given coding rate R for reliable communication, the minimum achievable per–symbol masking mutual information is lower bounded by $E(\varrho(R))$.

Comment: Referring to the discussion after eq. (7), the interval of t where optimum is sought, shrinks to $[\sigma_s(\sigma_s-\varrho\sigma_x),\sigma_s(\sigma_s+\varrho\sigma_x)]$.

Proposition 3: $E(\varrho(R))$ is achievable.

Proof. Given R, let ρ be the achiever of $E(\varrho(R))$. Now, apply dirty–paper coding to the (modified) Costa channel

$$Y = \tilde{X} + (1+a)S + Z,$$

where $\tilde{X}$ is, as was shown above, Gaussian and independent of S, and where $U = \tilde{X} + c(1+a)S$, with

$$c = \frac{\sigma_x^2(1-\rho^2)}{\sigma_x^2(1-\rho^2)+\sigma_z^2}.$$

This means that

$$U = \tilde{X} + c(1+a)S = X - aS + c(1+a)S = X + bS,$$

where

$$b = c(1+a) - a = \frac{\sigma_x^2(1-\rho^2) - \rho\sigma_z^2\sigma_x/\sigma_s}{\sigma_x^2(1-\rho^2)+\sigma_z^2}.$$

Since the power of $\tilde{X}$ is $\sigma_x^2(1-\rho^2)$, any coding rate up to

$$\frac{1}{2}\log\left[1 + \frac{\sigma_x^2(1-\rho^2)}{\sigma_z^2}\right]$$

is achievable as in [4].

Regarding the equivocation, we now show that with this choice of U, we have $I(S;U,Y) = I(S;Y)$, i.e., in the presence of Y, the observation of U, defined as in Costa [4], does not improve the MSE of linear estimation of S, and so the lower bound to the equivocation is met. In other words, the above choice of U simultaneously maximizes $I(U;Y) - I(U;S)$ and minimizes $I(S;Y,U)$. To show this, consider the minimum mean square error associated with optimum (linear) estimation of S given $Y = \tilde{X}+(1+a)S+Z$ and $U = \tilde{X}+bS$, i.e., $\boldsymbol{E}(S-\alpha Y-\beta U)^2$. We have to show that for the optimum coefficients (α^*, β^*), we have $\beta^* = 0$. Now, by solving the linear equations associated with (α^*, β^*), it is readily seen that β^* is given by a ratio of two expressions whose numerator is given by

$$\boldsymbol{E}(Y^2)\cdot\boldsymbol{E}(SU) - \boldsymbol{E}(UY)\cdot\boldsymbol{E}(SY). \tag{33}$$

Thus, proving that $\beta^* = 0$ is equivalent to proving that

$$\boldsymbol{E}(Y^2)\cdot\boldsymbol{E}(SU) = \boldsymbol{E}(UY)\cdot\boldsymbol{E}(SY). \tag{34}$$

Now, the left–hand side of the last equation is given by

$$\boldsymbol{E}(Y^2)\cdot\boldsymbol{E}(SU) = [\boldsymbol{E}(\tilde{X}^2) + (1+a)^2\sigma_s^2 + \sigma_z^2]\cdot b\sigma_s^2 \tag{35}$$

whereas the right–hand side is given by

$$\boldsymbol{E}(UY)\cdot\boldsymbol{E}(SY) = [E(\tilde{X}^2) + b(1+a)\sigma_s^2]\cdot(1+a)\sigma_s^2. \tag{36}$$

By using the above defined expressions of $\boldsymbol{E}(\tilde{X}^2)$, a, b, and c, the equality between the two expressions is readily verified. This completes the proof.•

Note that as long as the achiever ρ of $E(1)$ has absolute value striclty less than unity (which is the case of strong interference, cf. the Gaussian example at rate $R = 0$), then it is possible to transmit at a positive rate, without any loss in equivocation. In other words, the random variable V, in the earlier Gaussian example pertaining to $R = 0$, could be used for dirty–paper coding a la Costa, at rate up to

$$R = \frac{1}{2}\log\frac{B}{\sigma_z^2},$$

where B is defined as in eq. (20). A similar comment applies to weak interference, where the remainder power, not used to cancel S, can be harnessed to convey information at any rate up to

$$R = \frac{1}{2}\log\left(1 + \frac{\Gamma - \sigma_s^2}{\sigma_z^2}\right).$$

Another observation is that while in general, the channel $P(x|s,u)$ might be stochastic (in contrast to the ordinary G-P problem), in the Gaussian case, it remains always deterministic when $R > 0$ ($U = X + bS$ is equivalent to $X = U - bS$). Recall that for the case $R = 0$, it is not necessarily true.

V. Causal Side Information

In analogy to the Shannon model of causal side information [14], the question of trading off coding rate and equivocation is applicable also when the channel input is a *causal* (stochastic) function of s^n and the message w, i.e.,

$$P(x^n|s^n, w) = \prod_{i=1}^{n} P(x_i|s^i, x^{i-1}, w). \tag{37}$$

We argue that the characterization of the achievable region of pairs $\{(R,E)\}$ is the same as before, with the additional constraint that U is independent of S, and so, $I(U;S) = 0$ in the rate inequality, and $I(S;Y,U) = I(S;Y|U)$ in the equivocation inequality.

As for the converse part, we first note that the auxilary RV $U_i = (W, Y^{i-1}, S_{i+1}^n)$ is independent of S_i whenever the the encoder is as in eq. (37). In particular,

$$\begin{aligned} P(u_i, s_i) &= P(w, y^{i-1}, s_{i+1}^n, s_i) \\ &= P(w)\sum_{x^{i-1},s^{i-1}}\prod_{j=1}^{i-1}[P(s_j)P(x_j|x^{j-1},s^j,w)\times \\ &\quad P(y_j|x_j,s_j)]P(s_{i+1}^n)P(s_i) \\ &= P(u_i)P(s_i). \end{aligned} \tag{38}$$

In other words, given T, the RV's $U = (U_T, T)$ and $S = S_T$ are independent, i.e.,

$$P(u,s|t) = P(u|t)P(s|t) = P(u|t)P(s),$$

where the second equality follows again from the fact that S ans T are independent. Thus,

$$\begin{aligned} P(u,s) &= \frac{1}{n}\sum_{t=1}^{n} P(u,s|t) \\ &= \left[\frac{1}{n}\sum_{t=1}^{n} P(u|t)\right]P(s) = P(u)P(s). \end{aligned} \tag{39}$$

This means that $I(U;S) = 0$ in the rate inequality, and $I(S;Y,U) = I(S;Y|U)$ in the equivocation inequality.

As for the direct part, let U and X be random variables, where U independent of S, the Markov relation $U \rightarrow (X,S) \rightarrow Y$ is met, and the power constraint

$$\sum_{x,u,s} P(s)P(u)P(x|u,s)\phi(x) \leq \Gamma,$$

the rate constraint, $R < I(U;Y)$, and the equivocation constraint, $E \geq I(S;Y|U)$, are all met. Randomly select 2^{nR} independent codewords $\{u^n(1),\ldots,u^n(2^{nR})\}$ with uniform distribution within the type class corresponding to P_U. Given a message w and a state sequence s^n, x^n is generated by the product channel (37), where

$$P(x_i|x^{i-1},s^i,w) = P(x_i|s_i,u_i(w)). \tag{40}$$

First, observe that this induces a memoryless channel from U^n to Y^n, given by $P(y^n|u^n) = \prod_{i=1}^n P(y_i|u_i)$, thus, $U^n(W)$ is communicated reliably for $R < I(U;Y)$, by the ordinary coding theorem for DMC's. Regarding the equivocation, consider again the inequality (10). Now, $I(S^n,U^n) = I(S^n;U^n(W)) = 0$, as S^n and W are independent. The second term, $H(Y^n|U^n)$ is upper bounded by $\sum_{i=1}^n H(Y_i|U_i)$, as before, and

$$H(Y^n|U^n,S^n) = \sum_{i=1}^n H(Y_i|U_i,S_i) = nH(Y|U,S),$$

since $P(y^n|u^n,s^n)$ is a DMC and the joint statistics of U and S are according to $P(u,s) = P(u)P(s)$ (again, due to the independence between S^n and W).

REFERENCES

[1] G. Caire and S. Shamai (Shitz), "On the achievable throughput of a multi–antenna Gaussian broadcast channel," *IEEE Trans. Inform. Theory*, vol. 49, no. 7, pp. 1691-1706, July 2003.

[2] G. Caire, S. Shamai (Shitz), Y. Steinberg and H. Weingarten, "Information theoretic overview of MIMO broadcast channels, (invited paper), Chapter 18 in Space-Time Wireless Systems, From Array Processing to MIMO Communications, H. Bolcskei, D. Gebert, C. Papadias and A. J. van der Veen (Editors), Cambridge Press, London 2006.

[3] A. S. Cohen and A. Lapidoth, "The Gaussian watermarking game," IEEE Trans. Inform. Theory, vol. 48, no. 6, pp. 1639-1667, June 2002.

[4] M. H. M. Costa, "Writing on dirty paper," *IEEE Trans. Inform. Theory*, vol. IT-29, no. 5, pp. 439–441, May 1983.

[5] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, John Wiley & Sons, 1991.

[6] A. Host–Madsen, "On the capacity of cooperative diversity in slow fading channels,"*Proc. 40th Annual Allerton Conference on Communication, Control and Computing*, Allerton House, Monticello, Illinois, USA, October 2–4, 2002.

[7] S. I. Gel'fand and M. S. Pinsker, "Coding for channel with random parameters," *Problems of Information and Control*, vol. 9, no. 1, pp. 19-31, 1980.

[8] S. A. Jafar, G. J. Foschini and A. Goldsmith, "PhantomNet: exploring optimal multicellular multiple antenna systems," *EURASIP J. in Applied Signal Processing*, vol. 5, pp. 591-604, 2004.

[9] A. Lapidoth, P. Narayan, "Reliable communication under channel uncertainty," *IEEE Trans. Inform. Theory*, vol. 44, no. 6, pp. 2148-2177, October 1998.

[10] A. Maor and N. Merhav, "On joint information embedding and lossy compression in the presence of a stationary memoryless attack channel," *IEEE Trans. Inform. Theory*, vol. 51, no. 9, pp. 3166–3175, September 2005.

[11] N. Merhav, "On joint coding for watermarking and encryption," *IEEE Trans. Inform. Theory*, vol. 52, no. 1, pp. 190–205, January 2006.

[12] P. Moulin and J. A. O'Sullivan, "Information–theoretic analysis of information hiding," *IEEE Trans. Inform. Theory*, vol. 49, pp. 563-593, March 2003.

[13] 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, November 2003.

[14] C. E. Shannon, "Channels with side information at the transmitter," *IBM J. Res. Develop.*, vol. 2, pp. 289–293, October 1958.

[15] 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, March 2004.

[16] Y. Sun, A. Liveris, V. Stankovic, and Z. Xiong, "Near–capacity dirty–paper code designs based on TCQ and IRA codes," *Proc. 2005 Int. Symp. Inform. Theory (ISIT 2005)*, Adelaide, Australia, 4–9 September, 2005

[17] A. Sutivong, T. M. Cover, and M. Chiang, "Tradeoff between message and state information rates," *Proc. IEEE Int. Symp. Inform. Theory (ISIT 2001)*, Washington, DC, p. 303, June 2001.

[18] A. Sutivong, T. M. Cover, M. Chiang, and Y.-H. Kim, "Rate vs. distortion tradeoff for channels with state information," *Proc. ISIT 2002*, Lausanne, Switzerland, June–July, p. 226, 2002.

[19] 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, April 2005.

[20] 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, January 1976.

[21] 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, June 2002.