Error Exponents of Modulo-Additive Noise Channels with Side Information at the Transmitter

Uri Erez and Ram Zamir *
Dept. of Elect. Eng. - Systems, Tel Aviv University, ISRAEL
Submitted to the IEEE Tr. on Infor. Theory. Revised Aug 2000.

Abstract

Consider the optimum strategy for using channel state ("side") information in transmission over a modulo-additive noise channel, with state dependent noise, where the receiver does not have access to the side information. Recent work showed that capacity-wise, the optimum transmitter shifts each code letter by a "prediction" of the noise sample based on the side information. We show that this structure achieves also the random-coding error exponent, and therefore is optimum at some range of rates below capacity. Specifically, the optimum transmitter-predictor minimizes the Rényi entropy of the prediction error; the Rényi order depends on the rate, and goes to one (corresponding to Shannon entropy) for rates close to capacity. In contrast, it is shown that this "prediction strategy" may not be optimal at *low* transmission rates.

Key words: Time varying channels, side information, Rényi entropy, prediction, error exponent.

I. Introduction

The somewhat uncommon scenario of a time varying channel with side information at the transmitter did not attract much attention in the communication literature. In his 1958 paper [21], Shannon showed that the capacity of this channel is given

*This work was supported in part by the TAU research fund. This work was presented in part at the IT workshop in Metsovo, June 1999.

by the ordinary capacity of a derived channel with an extended input alphabet. A number of researchers extended Shannon's result to channels with memory [16], and to channels with non-causal side information [12], and treated some special cases [2, 15]. To the best of our knowledge, optimum transmission with side information at rates *below capacity*, and the corresponding *error exponents*, were treated in the literature only for the case of non casual side information [14].

This paper considers the class of discrete modulo-additive noise channels, with (time-varying) noise state information at the transmitter. The channel output at time i is given by

$$Y_i = X_i + Z_i \quad i = 1...n \tag{1}$$

where X_i is the channel input at time i, $X, Y, Z \in \mathcal{X} = \{0, ..., a-1\}$, and $+$ denotes addition modulo a. The analysis applies also if $\mathcal{X}$ is the interval $[0, a)$ in $\mathcal{R}$ and addition is performed modulo-a. The transmitter, which maps the message $W = 1...2^{nR}$ to the codeword $X_1, ..., X_n$, has access to the channel "state" sequence $S_1...S_n$. It is assumed that S_i is correlated with the "noise" Z_i, and is statistically independent of the message W; furthermore, the joint distribution of the message, states, inputs and noise samples can be written as

$$\begin{aligned} p(w, s_1...s_n, x_1...x_n, z_1...z_n) = \\ p(w)\ p(s_1...s_n)\ p(x_1...x_n|w, s_1...s_n)\ \prod_{i=1}^{n} p(z_i|s_i), \end{aligned} \tag{2}$$

i.e., $(W, X_1, \ldots, X_n, S_1, \ldots, S_{i-1}, S_{i+1}, \ldots, S_n)$ is conditionally independent of Z_i given S_i. In the memoryless channel case we assume further that $p(s_1, ..., s_n) = \prod p(s_i)$. The case where the message W *determines* the state sequence ("intersymbol interference channel") is different; see [8].

We consider two types of transmitters, a causal transmitter for which $x_i = f_i(w, s_1, ..., s_i)$, and a non-causal transmitter $x_i = f_i(w, s_1, ..., s_n)$. In either case the receiver does not have access to $s_1, ..., s_n$, and decodes w as $\hat{w} = g(y_1, ..., y_n)$.

Note that an average cost condition ("power constraint") on the transmitter output changes the nature of the problem, and will not be treated here. See, e.g., [1, 10].

Shannon's solution for the capacity in the memoryless causal case is given in terms of the ordinary capacity of a "derived" channel, with $a^{|\mathcal{S}|}$ inputs called "strategies". An important feature of Shannon's optimum transmitter is that it is "instantaneous" with respect to the states, i.e., at each moment it suffices to use the current state (rather than the entire state history) in order to achieve capacity.

The finite-state additive-noise channel model above has some unique properties: (i) the same symmetry of the transition distribution holds with respect to all states; (ii) the input alphabet and the output alphabet are of the same size; and (iii) memory, if it exists, enters only through the state process. Since capacity (as well as random coding error exponent) is achieved by assigning positive probabilities to no more input letters than the size of the output alphabet, the second property above implies that most of the letters of Shannon's derived channel are not needed. These special properties, as well as the general nature of channels with side information, motivate us to raise the following questions:

- Which a out of the $a^{|\mathcal{S}|}$ strategies are needed to achieve capacity? Which are needed for optimum transmission at rates below capacity?
- Does the structure of the optimum transmitter obey a simple / intuitive law? Particularly, can an "instantaneous" transmitter (like the one proposed by Shannon for achieving capacity) achieve the optimum error-exponent?
- Can we find simple solutions for channels with memory and for non-causal side information?
- Which side of the communication channel makes a better use of the side-information - the transmitter or the receiver?

In a recent paper [9] we answered these questions with regard to capacity. We showed that an "instantaneous prediction encoder" of the form

$$x_i = f_i(w) - t^*(s_i) \qquad i = 1 \ldots n \tag{3}$$

achieves the capacity of the channel (1) with side information at the transmitter, where n is the block length, w is the message to be transmitted, $\mathbf{f}(w) = (f_1(w), \ldots, f_n(w))$ is a "side information independent" code, and the function $t^*(s)$, the "noise predictor", minimizes the Shannon entropy of $Z - t(S)$.

When working at rates below capacity, one is usually interested in $P_e^{opt}(n, R)$, the average error probability of the optimal code for a given block length n and rate R. In general, the codes appropriate for different rates are different. As we shall see, for the additive noise channel with side information at the transmitter, the instantaneous prediction structure (3) achieves the exponent of $P_e^{opt}(n, R)$ for rates above the critical rate; for this encoding structure the effect is that the optimal noise predictor varies with the rate, and coincides with the capacity achieving predictor for $R = C$. At lower transmission rates, however, the instantaneous prediction transmitter is not necessarily optimal; we provide an example of a channel for which at a certain rate the largest error exponent cannot be achieved by a code which has the structure of (3).

Our main result is stated and proved in the next Section. Section III extends the results to non-causal side information and to channels with memory. Section IV discusses transmission at low rates.

II. Main Result

Recall the basic definition of a causal encoder with side-information. At time instant i, the encoder transmits $x_i = f_i(w, s_1^i)$, where s_1^i denotes $s_1, \ldots, s_i$. This setting can be viewed as a regular channel with extended input alphabet

$$\mathcal{T}^i = \{t : \mathcal{S}_1^i \longrightarrow \mathcal{X}\}$$

and output alphabet $\mathcal{Y}$, where the encoding function f_i is replaced by mapping w to a $t_i \in \mathcal{T}^i$ such that $t_i(s_1^i) = f_i(w, s_1^i)$. For a block of length n we thus regard the channel as having input alphabet

$$\mathcal{T}^1 \times \mathcal{T}^2 \times \ldots \times \mathcal{T}^n \triangleq \mathcal{A}^{(n)} \tag{4}$$

and output alphabet $\mathcal{Y}^n$. Each input $\mathbf{t} \in \mathcal{A}^{(n)}$ is called a "strategy vector", and the transition probability $P(\mathbf{y}|\mathbf{t})$ from $\mathcal{A}^{(n)}$ to $\mathcal{Y}^n$ is defined by taking the expectation over S^n with respect to the underlying channel $p(y|x,s)$ [21, 9], i.e.,

$$P(\mathbf{y}|\mathbf{t}) = \sum_{s^n} p(s^n) p(y^n|x^n = \mathbf{t}(s^n), s^n).$$

Thus, even if the underlying channel is memoryless, the resulting channel in general is not.

Let $Q_n(\mathbf{t})$ be an arbitrary probability assignment on the possible strategy vectors of length n, i.e. on $\{(t_1, t_2, \ldots, t_n)\} = \mathcal{A}^{(n)}$. For a given rate R, the Gallager random coding bound on the optimum error probability $P_e^{opt}(n, R)$ when transmitting at rate R through the channel $P(\mathbf{y}|\mathbf{t})$ (or equivalently, through the channel $p(y|x,s)$ with side information s), is given by, [11],

$$P_e^{opt}(n, R, Q_n) \le e^{nR\rho} \sum_{\mathbf{y}} \Big(\sum_{\mathbf{t}} Q_n(\mathbf{t}) P(\mathbf{y}|\mathbf{t})^{\frac{1}{1+\rho}} \Big)^{1+\rho} \triangleq e^{-nE_r^n(Q_n,\rho,R)} \tag{5}$$

for $0 \le \rho \le 1$, where $E_r^n(Q_n, \rho, R)$ is the *block* form of the Gallager random coding error exponent. For a given rate, the tightest bound is obtained by maximizing $E_r^n(Q_n, \rho, R)$ over Q_n and ρ. Define

$$E_r^n(\rho, R) = \max_{Q_n} E_r^n(Q_n, \rho, R). \tag{6}$$

It follows from the discussion above that for a general strategy vector $\mathbf{t}$ the transition distribution $P(\mathbf{y}|\mathbf{t})$ in (5) has memory. Nonetheless, as Shannon [21] showed with respect to the capacity, we show in Theorem 1 below that if the underlying channel is memoryless with additive noise, $E_r^n(\rho, R)$ is achieved by *instantaneous* strategies only, i.e., strategies belonging to $\mathcal{T}^1$. This implies that the maximizing distribution Q_n is *i.i.d.*, and that the error exponent $E_r^n(\rho, R)$ does not depend on the block length n. We derive an explicit expression for $E_r^n(\rho, R)$ in terms of the Rényi entropy.

For any α, $\alpha > 0, \alpha \neq 1$, the Rényi entropy of a distribution $p(x)$ is defined by [3]

$$H_\alpha(X) \triangleq \frac{\alpha}{1-\alpha} \log \|p(x)\|_\alpha \quad \alpha > 0\ , \quad \alpha \neq 1 \tag{7}$$

where

$$\|p(x)\|_\alpha \triangleq (\sum_x p(x)^\alpha)^{\frac{1}{\alpha}}$$

and $\log(\cdot)$ denotes natural logarithm. The Rényi entropy of order $\alpha = 1$, defined by taking the limit of (7) as $\alpha \longrightarrow 1$, coincides with the Shannon entropy, i.e. $H_1(X) = H(X)$.

Theorem 1 *(***Memoryless causal case***) The Gallager random coding error exponent (6), for the discrete memoryless additive noise channel (1) with* causal *side information S at the transmitter, is given by*

$$E_r^n(\rho, R) = \rho\Big[\log|\mathcal{X}| - \min_{t:\mathcal{S}\to\mathcal{X}} H_{\bar{\rho}}\Big(Z - t(S)\Big) - R\Big], \tag{8}$$

independently of n, where $0 \le \rho \le 1$ and $\bar{\rho} = \frac{1}{1+\rho}$.

For the proof we need a property of convex functions which we call "irrelevancy". The following lemma states this property and Appendix A proves it.

Lemma 1 ("Irrelevancy") *Let U be a r.v. defined over the finite alphabet $\mathcal{U}$, and let $F(\cdot)$ be a convex $\cap$ function defined upon the probability distributions on $\mathcal{X}$. If U, S, Z form a Markov chain $U \leftrightarrow S \leftrightarrow Z$ then for any function $g : \mathcal{S} \times \mathcal{U} \to \mathcal{X}$*

$$F(Z - g(S,U)) \ge F(Z - t^*(S)) \tag{9}$$

where

$$t^*(\cdot) \triangleq \arg\min_{t:\mathcal{S}\to\mathcal{X}} F(Z - t(S)) \tag{10}$$

and $F(X)$ denotes $F(P_X)$. In particular, if the pair (Z, S) is independent of U then $F(Z - g(S,U)) \ge F(Z - t^*(S))$. Also, if Z is independent of (S,U) then $F(Z - g(S,U)) \ge F(Z)$.

Note that the lemma applies for Rényi entropies for any ρ (i.e., taking $F(\cdot)$ to be $H_\rho(\cdot)$) since $\sum_x p(x)^\alpha$ is convex $\cap$ in p.

Proof of Theorem 1: Partition the set of strategy vectors $\mathcal{A}^{(n)}$ into classes of constant difference. That is, two n-sequences $[t_1(s_1), \ldots, t_n(s_1^n)]$ and $[\bar{t}_1(s_1), \ldots, \bar{t}_n(s_1^n)]$ in $\mathcal{A}^{(n)}$ are members of the same class iff

$$t_i(s_1^i) - \bar{t}_i(s_1^i) = c_i \quad \forall i = 1 \ldots n \tag{11}$$

for some constants $c_1, \ldots, c_n \in \mathcal{X}$ that do not depend on the states. Thus each class contains $|\mathcal{X}|^n$ members. Due to the additivity of the channel, we have that

$$P_{\mathbf{y}|\mathbf{t}}(\mathbf{y}|\mathbf{t}+\mathbf{c}) = P_{\mathbf{y}|\mathbf{t}}(\mathbf{y}-\mathbf{c}|\mathbf{t}). \tag{12}$$

We can thus associate with each class an "effective noise" as follows. Pick an arbitrary $\mathbf{t}$ to represent the class, and define the effective noise $\mathbf{Z_t}$ to have the distribution

$$p_{\mathbf{Z_t}}(\mathbf{z}) = P_{\mathbf{y}|\mathbf{t}}(\mathbf{z}|\mathbf{t}) \quad \text{i.e.,} \quad \mathbf{Z_t} = \mathbf{Z} + \mathbf{t}(\mathbf{S}). \tag{13}$$

Picking a different representative for the class would only shift the noise by some constant vector $\mathbf{c}$. It follows that a distribution $\tilde{Q}_n(\mathbf{t})$ which is uniform within the class containing $\tilde{\mathbf{t}}$ and zero elsewhere induces a uniform distribution on $\mathcal{Y}^n$. Hence substituting this $\tilde{Q}_n$ in (5) gives (see Appendix D)

$$E_r^n(\tilde{Q}_n, \rho, R) = \rho\Big[\log|\mathcal{X}| - \frac{1}{n}H_{\bar{\rho}}(\tilde{\mathbf{Z}}) - R\Big], \tag{14}$$

where $\tilde{\mathbf{Z}}$ is the effective noise associated with the class of $\tilde{\mathbf{t}}$, and $H_{\bar{\rho}}(\cdot)$ denotes *joint* Rényi entropy of order $\bar{\rho}$. Theorem 5.6.5 of [11] gives a necessary and sufficient condition for an input distribution Q_n to maximize the Gallager error exponent for a given ρ. Applying this theorem to $\tilde{Q}_n$ above (note that Gallager's "$\alpha_j(\tilde{Q}_n)$" is independent of j in our case) we obtain that it achieves $E_r^n(\rho, R)$ iff the following inequality is satisfied for *all* strategy vectors $\mathbf{t} \in \mathcal{A}^{(n)}$

$$\sum_{\mathbf{y}} P(\mathbf{y}|\mathbf{t})^{\bar{\rho}} \geq \sum_{\mathbf{y}} P(\mathbf{y}|\tilde{\mathbf{t}})^{\bar{\rho}} = \sum_{\mathbf{z}} p_{\tilde{\mathbf{Z}}}(\mathbf{z})^{\bar{\rho}}. \tag{15}$$

Taking the logarithm and dividing by $1-\bar{\rho}$, we see that inequality (15) is equivalent to

$$H_{\bar{\rho}}(\mathbf{Z_t}) \geq H_{\bar{\rho}}(\tilde{\mathbf{Z}}) \tag{16}$$

where $\mathbf{Z_t}$ and $\tilde{\mathbf{Z}}$ are the effective noises induced via (13) by $\mathbf{t}$ and $\tilde{\mathbf{t}}$, respectively. Thus the problem reduces to finding the vector of shift functions $[t_1(s_1), t_2(s_1^2), \ldots, t_n(s_1^n)]$ resulting in effective noise vector $\mathbf{Z_t}$, having minimum joint Rényi entropy of order $\bar{\rho}$, and choosing $\tilde{Q}_n$ to be *uniform* over the class associated with $\mathbf{Z_t}$. To prove (8), it remains to be shown that this optimal shift vector function, $\mathbf{t}^*$, decomposes into the instantaneous time invariant form $[t^*(s_1), t^*(s_2), \ldots, t^*(s_n)]$. For any $0 < \alpha < 1$, minimizing $H_\alpha(\tilde{\mathbf{Z}})$ is equivalent to minimizing

$$\begin{aligned} A_\alpha(\tilde{\mathbf{Z}}) &\triangleq \sum_{\tilde{\mathbf{z}}} [p(\tilde{\mathbf{z}})]^\alpha \\ &= \sum_{\tilde{z}_1 \ldots \tilde{z}_n} p(\tilde{z}_1)^\alpha p(\tilde{z}_2|\tilde{z}_1)^\alpha \ldots p(\tilde{z}_n|\tilde{z}_1^{n-1})^\alpha \\ &= \sum_{\tilde{z}_1, \tilde{z}_2, \ldots, \tilde{z}_{n-1}} \left[p(\tilde{z}_1)^\alpha p(\tilde{z}_2|\tilde{z}_1)^\alpha \ldots p(\tilde{z}_{n-1}|\tilde{z}_1^{n-2})^\alpha \sum_{\tilde{z}_n} p(\tilde{z}_n|\tilde{z}_1^{n-1})^\alpha \right] \end{aligned} \quad (17)$$

where for ease of notation we suppress the subscript in the distribution $p_{\tilde{Z}}(z)$ and the associated conditional distributions. We wish to minimize the last term in (17) which we rewrite as

$$\sum_{\tilde{z}_n} p(\tilde{z}_n|\tilde{z}_1^{n-1})^\alpha = \sum_z \Pr\Big(Z_n + t_n(S_1^n) = z \Big| \tilde{Z}_1^{n-1} = \tilde{z}_1^{n-1}\Big)^\alpha. \quad (18)$$

Since the pair (Z_n, S_n) is statistically independent of (Z_1^{n-1}, S_1^{n-1}), we have

$$\min_{t:\mathcal{S}^n \to \mathcal{X}} \sum_z \Pr\Big(Z_n + t(S_1^n) = z \Big| \tilde{Z}_1^{n-1} = \tilde{z}_1^{n-1}\Big)^\alpha \quad (19)$$

$$= \min_{t:\mathcal{S}^1 \to \mathcal{X}} \sum_z \Pr\Big(Z_n + t(S_n) = z \Big| \tilde{Z}_1^{n-1} = \tilde{z}_1^{n-1}\Big)^\alpha \quad (20)$$

$$= \min_{t:\mathcal{S}^1 \to \mathcal{X}} \sum_z \Pr\Big(Z_n + t(S_n) = z\Big)^\alpha \quad (21)$$

where (20) follows from Lemma 1, letting Z_n,S_n and S_1^{n-1} play the roles of Z,S and U in the lemma, since the above independence relation implies that $S_1^{n-1} \leftrightarrow S_n \leftrightarrow Z_n$ form a Markov chain when conditioned on $\tilde{Z}_1^{n-1} = \tilde{z}_1^{n-1}$; and (21) follows from independence. Since replacing the "+" sign by a "−" would not change the result of the minimization, we define

$$Z_\rho^* = Z - t_\rho^*(S) \quad ; \quad t_\rho^*(\cdot) = \arg \min_{t:\mathcal{S} \to \mathcal{X}} H_{\bar{\rho}}(Z - t(S)) \quad (22)$$

as the minimum Rényi entropy effective noise, and the "optimal predictor" of order $\bar{\rho}$, respectively. Then, (19)-(21) imply that setting $t_n = t_\rho^*$ can only reduce the entropy of $H_{\bar{\rho}}(\widetilde{\mathbf{Z}})$. Having done so, due to the causality of the side information, the above argument can be applied to t_{n-1}, and by induction, the entire minimization reduces to that of the single letter minimization of $H_{\bar{\rho}}(Z - t(S))$. Therefore the minimizing strategy vector in (16) is $[t^*_\rho(s_1), \dots, t^*_\rho(s_n)]$, the minimum Rényi entropy is

$$H_{\bar{\rho}}(\mathbf{Z}^*) = nH_{\bar{\rho}}(Z - t^*_\rho(S)), \tag{23}$$

and the minimum over $\tilde{Z}$ of the error exponent (14) becomes the one given in (8). $\square$

Remarks:

1) *Tightness of the bound*: By (5), (8) and (22) we have

$$P_e^{opt}(n, R) \leq \min_{0 \leq \rho \leq 1} \exp\Big\{-n\rho\Big(\log|\mathcal{X}| - H_{\bar{\rho}}(Z^*_\rho) - R\Big)\Big\}. \tag{24}$$

Furthermore, by releasing the constraint on ρ, the random coding (upper) bound becomes the sphere-packing lower bound, [11], and we obtain

$$P_e^{opt}(n, R) \geq \min_{\rho > 0} \exp\Big\{-n\rho\Big(\log|\mathcal{X}| - H_{\bar{\rho}}(Z^*_\rho) - R + o(1)\Big)\Big\} \tag{25}$$

where $o(1) \longrightarrow 0$ as $n \longrightarrow \infty$. Therefore, we have an explicit expression for the exponent of $P_e^{opt}(n, R)$ for rates above the critical rate (the rate above which $E_r(R, \rho)$ is maximized by $\rho < 1$) [11].

2) *The optimum transmitter:* The proof above shows that the optimizing $Q_n(\mathbf{t})$ in (5) is positive only for strategies $\mathbf{t}$ which are constant shifts of $[t^*_\rho(s_1), \dots, t^*_\rho(s_n)]$. Thus, for rates above the critical rate the "instantaneous prediction" encoding structure

$$x_i = f_i(w) - t^*(s_i; R) \tag{26}$$

achieves the exponent of $P_e^{opt}(n, R)$, where $t^*(s_i; R)$ is the optimal "noise predictor" t^*_ρ associated with the ρ that achieves the minimum in (24), and $\mathbf{f}(w)$ is a "good" rate R code for a modulo-additive noise channel *without* side information (See Figure 1). We show in the sequel by example that for low rates this ceases to be true.

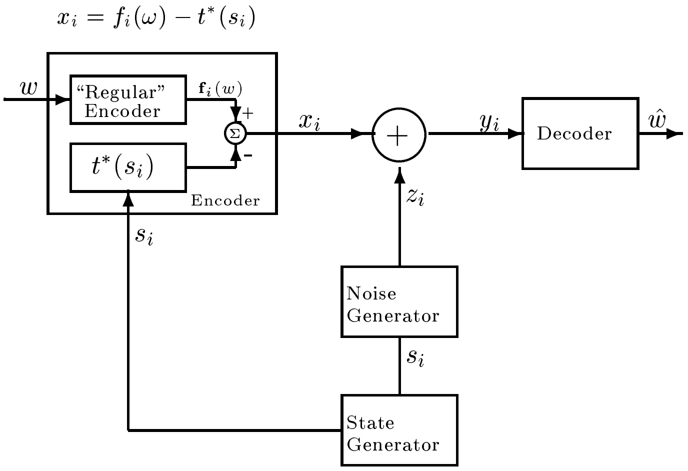

Figure 1: Instantaneous prediction encoding scheme

3) *Limiting cases for* ρ: Note that Z_0^* and t_0^* coincide with the effective noise and the predictor achieving capacity in [9], since the Rényi entropy of order one reduces to the Shannon entropy. On the other extreme, when ρ tends to infinity ($\bar{\rho} \to 0$) we have that (see, e.g., [3])

$$\lim_{\bar{\rho}\to 0} H_{\bar{\rho}}(X) = \log[\text{Support}P(X)]. \tag{27}$$

Thus if $\lim_{\rho\to\infty} \log[\text{Support}P(Z - t_\rho^*(S))] < \log|\mathcal{X}|$, then the *exponent* of the sphere packing bound (25) diverges to infinity at

$$R_\infty = \log|\mathcal{X}| - \lim_{\rho\longrightarrow\infty} \log[\text{Support}P(Z - t_\rho^*(S))] \tag{28}$$

and (28) is an upper bound to the zero error capacity of the channel [4].

4) *Dependence of the optimal predictor on the rate*: It is known that the error exponent is more sensitive to the tail of the noise distribution for rates small compared to the channel capacity. As discussed in the previous remark, and by the monotonicity of the Rényi entropy $H_\alpha(\cdot)$ in α, the optimal predictor will tend to *squeeze* the support of the effective noise more and more as the rate decreases. To illustrate this, consider the channel depicted in Figure 2. The alphabet is the interval $[0, 16) \subset \mathcal{R}$ and the channel has additive noise modulo-16. Note that all our results hold if the alphabet is a real interval $[0, a)$ and addition is performed modulo-a. The noise distribution is given by

$$\begin{aligned} p(Z = z|s_0) &= d \quad 0 < z < a \\ p(Z = z|s_0) &= c \quad a < z < b \end{aligned} \tag{29}$$

and $p(Z=z|s_1)$ is the mirror image of (29), where we take $a=1, b=8, c=2/112$, and $d=42/112$. Figure 2 depicts two effective noise channels (corresponding to two different predictors). By examining the behavior of the Rényi entropy of the effective noise as a function of the relative displacement of the two component noises, it can be verified that these two predictors are the only optimal ones for any $\rho \geq 0$. Predictor (b) is optimal up to some rate from which predictor (a) is optimal up to capacity. Figure 3 depicts $E_r(Q_i, R) = \max_{0\leq\rho\leq 1} E_r(Q_i, \rho, R), \quad i=1,2$, for these two predictors (predictor i corresponds to a uniform distribution $Q_i(t)$ within the corresponding class) from the critical rate of the channel (which is the greater of the critical rates of the two) up to capacity. Thus $E_r(R) = \max_\rho E(\rho, R)$ is the upper envelope of the two curves. Note that in accordance with Theorem 1 the superscript n of these error exponents was supressed.

5) *Feedback cannot improve the sphere packing bound of a symmetric channel*: Formally, noiseless feedback in a modulo-additive noise channel, which amounts to defining $S_n = Z_{n-1}$, does not fall into our model in (2), since S_{n+1} and Z_n are not conditionally independent given S_n. Nevertheless, we might note that the above derivation can be applied to a memoryless symmetric (modulo-additive noise) channel with noiseless feedback. Specifically, in view of Lemma 1, it is evident from (18), since Z_n is independent of $(Z_{n-1}, Z_{n-2}, \ldots)$, that prediction is useless. Therefore, feedback cannot improve the error exponent of block codes beyond the sphere packing bound, as is well known (see e.g., [4, sec. 2.5, prob. 33]).

6) *Side information at the receiver, transmitter and both*: When no Side Information (SI) is available, the error exponent, $E_r^{NOSI}(\rho, R)$, is clearly given by $\rho\Big[\log|\mathcal{X}| - H_{\tilde{\rho}}(Z) - R\Big]$. With SI available at transmitter alone we obtained the error exponent in (8). For the case where SI is available at receiver only, since a uniform distribution achieves $E_r(\rho, R)$ for each state s, we have (see e.g., [19] for the case $\rho = 1$)

$$e^{-E_r^{SI@REC}(\rho,R)} = \sum_s p(s) e^{-\rho\Big[\log|\mathcal{X}| - H_{\tilde{\rho}}(Z_s) - R\Big]}, \tag{30}$$

where $Z_s \sim p(z|s)$. From the definition of Rényi entropy (7) we have that $e^{\rho H_{\tilde{\rho}}(Z_s)} =$

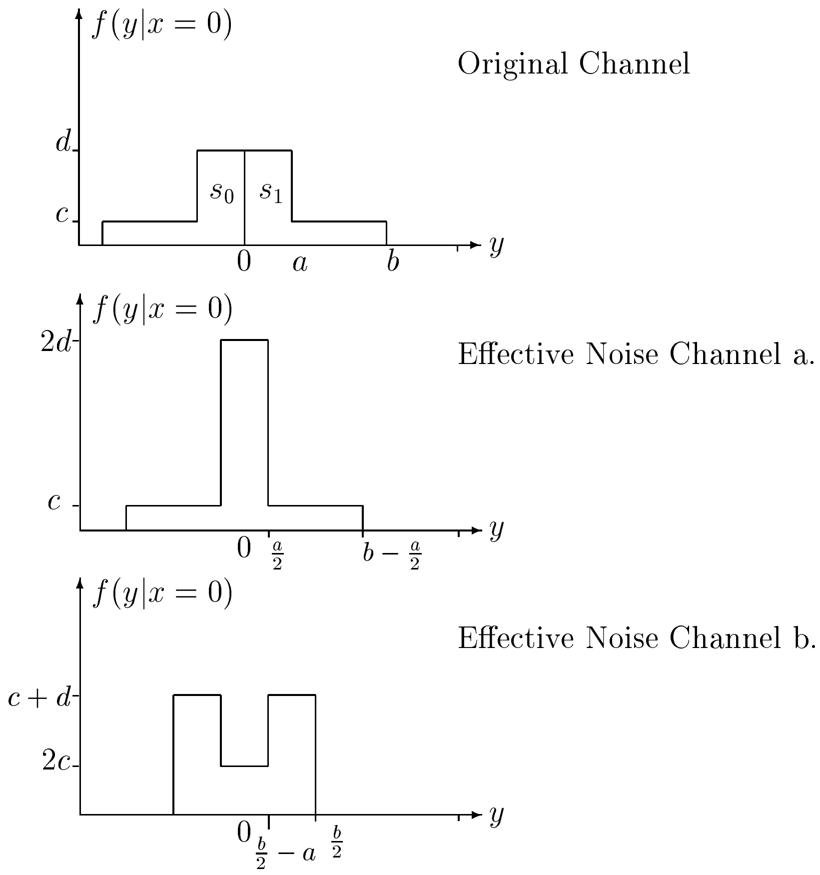

Figure 2: Example

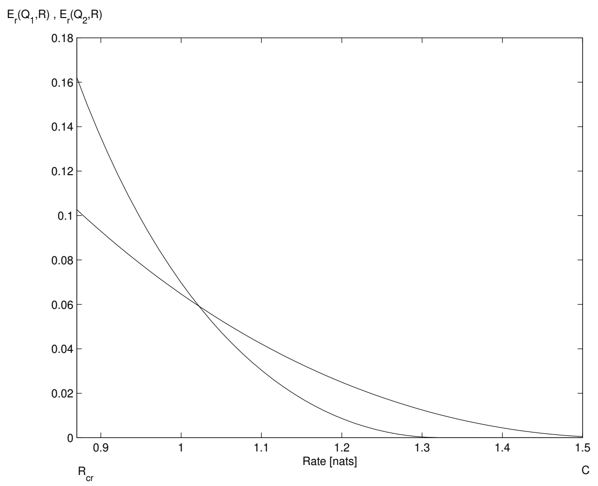

Figure 3: Error exponent for the example of Figure 2

$||p_{Z_s}||_{\bar{\rho}}$, thus we can summarize the various exponents as

$$E_r^{NOSI}(\rho, R) = \rho\Big[\log|\mathcal{X}| - R - \frac{1}{\rho}\log||p_Z||_{\bar{\rho}}\Big] \quad (31)$$

$$E_r^{SI@TR}(\rho, R) = \rho\Big[\log|\mathcal{X}| - R - \frac{1}{\rho}\log||p_{Z^*}||_{\bar{\rho}}\Big] \quad (32)$$

$$E_r^{SI@REC}(\rho, R) = \rho\Big[\log|\mathcal{X}| - R - \frac{1}{\rho}\log\sum_s p(s)||p_{Z_s}||_{\bar{\rho}}\Big] \quad (33)$$

where $Z^* = Z - t_\rho^*(S)$. It can be shown that the latter is also the random coding exponent when the SI is available at both receiver and transmitter, i.e. that we have $E_r^{SI@BOTH} = E_r^{SI@REC}$. As shown in Appendix C, these exponents satisfy the relation

$$E_r^{SI@BOTH} = E_r^{SI@REC} \geq E_r^{SI@TR} \geq E_r^{NOSI}. \quad (34)$$

7) *Continuous amplitude channels:* The above derivation can be extended to *non* modulo- additive noise channels, with continuous amplitude (i.e., $\mathcal{X} = \mathcal{R}$) and peak input constraint, in the limit of high signal to noise ratio. Specifically, if the input X must satisfy the constraint $|X| \leq A/2$, and if $E(Z^2|S = s) \ll A^2 \ \forall s$, then the error exponent (6) of the channel can be approximated by the error exponent of a modulo-A channel [18]. This follows since the optimum input distribution tends to a uniform over $(-A/2, A/2)$ in the limit as $A \to \infty$, and by the continuity of the error exponent in the channel distribution. Thus, under some technical conditions on the "smoothness" of the noise distribution, (31) - (33) give good approximations for the corresponding error exponents, provided that $\log|\mathcal{X}|$ is replaced by $\log(A)$. We emphasize again that with an average cost ("power") constraint the problem is inherently different.

III. Non-Causal Side Information and State Process with Memory

In this section we generalize our results to state processes with memory as well as to "non causal" side information, as defined in the Introduction.

For a general noise distribution $p(z|s)$, the crucial step in the proof of Theorem 1, i.e. (19)-(21), does not carry over if the state process $S_1, S_2, \dots$ has memory and/or the encoder is non-causal. To overcome this obstacle, we confine our treatment to a special class of noise distributions, the so called State Weight Independent Prediction (SWIP) noises, which allows us to treat both problems together and to obtain simple expressions for the error exponent for both cases. For SWIP noise we establish that having the entire state sequence in advance does not yield a greater error exponent than that obtained by a causal encoder, having the "instantaneous prediction" form of (26).

Definition 1 ("SWIP Noise") *The conditional distribution $p(z|s)$ is "ρ-SWIP" if the optimal predictor t^*_ρ does not depend on the state weights, i.e. on $p(s)$.*

In particular, if $p(z|s)$ (as a function of z) is either unimodal and symmetric for each state s, or monotonically increasing/decreasing for all the states, then it is "ρ-SWIP" for all $\rho \geq 0$. In both cases the assertion follows from the observation that choosing $t^*(s)$ so as to align the maxima together, i.e., so that $\arg\max_x p_{Z_s}(x - t^*(s))$ is the same for all s, minimizes the corresponding Rényi entropy (see Appendix B). Note that a general noise $p(z|s)$ is not necessarily ρ-SWIP; see the treatment in [9] for the case $\rho = 0$.

Theorem 2 (SWIP Noise with Memory: Causal and Non-Causal Case) *For ρ-SWIP noise and stationary state process, the instantaneous shift function*

$$t^n(s^n) = (t^*_\rho(s_1), \dots, t^*_\rho(s_n)),$$

where $t^*_\rho(\cdot)$ is defined in (22), achieves the random coding error exponent for both causal and non causal side information at the encoder. Thus

$$E^n_r(\rho, R) = \rho\Big[\log|\mathcal{X}| - \frac{1}{n} H_{\bar{\rho}}\Big(Z_1 - t^*_\rho(S_1), \dots, Z_n - t^*_\rho(S_n)\Big) - R\Big], \tag{35}$$

which can be achieved by the instantaneous-prediction encoder of (26).

Note that the Gallager random coding bound (5) is not asymptotic, therefore the error exponent (35) provides a bound for each n. It might be possible to compute the asymptotic form of (35) as $n \to \infty$ using, e.g., the technique of [22].

Proof: Analogously to our treatment in Section II we now regard the channel as having a block of length n as its input alphabet. In effect, the input alphabet is $[\mathcal{T}^1]^n \triangleq \mathcal{A}^{(n)}$ and the output alphabet is $\mathcal{Y}^n$. The whole derivation leading to (16) can be carried over. We thus wish to find the strategy vector $[t_1(s_1^n), t_2(s_1^n), \ldots, t_n(s_1^n)]$ inducing effective noise, $\mathbf{Z_t} = [Z_1 + t_1(S_1^n), \ldots, Z_n + t_n(S_1^n)]$, having the minimum possible Rényi entropy of order $\bar{\rho}$. Minimizing $H_{\bar{\rho}}(\tilde{\mathbf{Z}})$ is equivalent to minimizing $A_{\bar{\rho}}(\tilde{\mathbf{Z}})$ of (17). At this point our derivation diverges from that of Theorem 1, because (Z_n, S_n) is no longer independent of (S_1^{n-1}, Z_1^{n-1}), and we require the noise to be SWIP. Specifically, for any $1 \le i \le n$ isolate the i-th term in $A_{\bar{\rho}}(\tilde{\mathbf{Z}})$ using Bayes' law

$$A_{\bar{\rho}}(\tilde{\mathbf{Z}}) = \sum_{\tilde{z}_1, \ldots, \tilde{z}_{i-1}, \tilde{z}_{i+1}, \ldots, \tilde{z}_n} p(\tilde{z}_1, \ldots, \tilde{z}_{i-1}, \tilde{z}_{i+1}, \ldots, \tilde{z}_n)^{\bar{\rho}} \sum_{\tilde{z}_i} p(\tilde{z}_i | \tilde{z}_1^{i-1}, \tilde{z}_{i+1}^n)^{\bar{\rho}} \quad (36)$$

and bound it for each value of the condition $\tilde{z}_1^{i-1}, \tilde{z}_{i+1}^n$ as follows:

$$\begin{aligned} \sum_{\tilde{z}_i} p(\tilde{z}_i | \tilde{z}_1^{i-1}, \tilde{z}_{i+1}^n)^{\bar{\rho}} &= \sum_z \Pr\Big(Z_i - t_i(S_1^n) = z \Big| \tilde{Z}_1^{i-1} = \tilde{z}_1^{i-1}, \tilde{Z}_{i+1}^n = \tilde{z}_{i+1}^n \Big)^{\bar{\rho}} \\ &\ge \sum_z \Pr\Big(Z_i - t_\rho^*(S_i) = z \Big| \tilde{Z}_1^{i-1} = \tilde{z}_1^{i-1}, \tilde{Z}_{i+1}^n = \tilde{z}_{i+1}^n \Big)^{\bar{\rho}} \quad (37) \end{aligned}$$

where the lower bound follows from two facts: (i) By Lemma 1, $Z_i \leftrightarrow S_i \leftrightarrow (S_1^{i-1}, S_{i+1}^n)$ (playing the roles of Z, S, U in the lemma) form a Markov chain conditionally on $(\tilde{Z}_1^{i-1}, \tilde{Z}_{i+1}^n)$, and so each t_i can be taken to be a function of s_i only. (ii) Due to the SWIP property, the condition $(\tilde{z}_1^{i-1}, \tilde{z}_{i+1}^n)$, even if it affects the distribution of S_i, it does not affect the optimizing predictor which hence is equal to the optimum *marginal* predictor t_ρ^* for all i. Note though, that the sum itself does depend on the condition (unlike in (20)-(21)), so the optimum $A_{\bar{\rho}}(\tilde{\mathbf{Z}})$ does not break into a product (unlike in Theorem 1). The above argument applies for t_i , $i = 1 \ldots n$, and we obtain

$$H_{\bar{\rho}}(\mathbf{Z} - \mathbf{t}(\mathbf{S})) \ge H_{\bar{\rho}}\Big(Z_1 - t_\rho^*(S_1), \ldots, Z_n - t_\rho^*(S_n) \Big) \quad (38)$$

for any function $\mathbf{t}$. As a consequence (35) follows. $\square$

We now use $E_r^{caus}(\rho, R)$ to denote the error exponent with *causal* SI at the transmitter, and $E_r^{ncaus}(\rho, R)$ for the error exponent with *non causal* SI. We have:

Corollary 1 (Memoryless Non-Causal Case) *For the memoryless additive noise channel, if the noise satisfies the SWIP property, then*

$$E_r^{ncaus}(\rho, R) = E_r^{caus}(\rho, R) = \rho\Big[\log|\mathcal{X}| - H_{\bar{\rho}}(Z - t_\rho^*(S)) - R\Big]. \tag{39}$$

It is worth noting that the derivation up to (16) above holds even without the SWIP assumption, and leads to a vector form of the error exponent

$$E_r^n(\rho, R) = \rho\Big[\log|\mathcal{X}| - \frac{1}{n}\min_{\mathbf{t}:\mathcal{S}^n \to \mathcal{X}^n} H_{\bar{\rho}}\big(\mathbf{Z} - \mathbf{t}(\mathbf{S})\big) - R\Big]. \tag{40}$$

Although this expression is less useful than (39), it still provides further insight into the involved (albeit, single letter) expressions of [14].

IV. Transmission at rates below the critical rate

So far we have shown that noise prediction achieves the error exponents for rates above the channel's critical rate. Given the simplicity of the scheme, one may wonder whether the instantaneous noise prediction scheme of (26) might be optimal for *any* rate of transmission. However, it is easily seen that the pair of strategies maximizing the Bhattacharyya distance may belong to different classes. This hints that at low rates the prediction scheme may cease to be optimal. The following is an example of a channel having a *zero error capacity* (i.e., an infinite error exponent below some rate) which cannot be achieved by prediction.

Let the alphabet size be $|\mathcal{X}| = 16$ and the number of states $|\mathcal{S}| = 2$. Figure 4A and 4B show the support of the channel output distribution for $x = 0$ and states $s1$ and $s2$, respectively (stars: $p(y|x = 0, s1)$ circles: $p(y|x = 0, s2)$). For $x = 1, 2, \ldots, 15$ this picture is appropriately "rotated" according to modulo-16 addition.

Figure 4C depicts the effective output distribution associated with two strategies,

$$t_1(s) = \begin{cases} 0 & s = s_1 \\ 1 & s = s_2 \end{cases} \quad \text{(regular line)}$$

and

$$t_2(s) = \begin{cases} 8 & s = s_1 \\ 0 & s = s_2 \end{cases} \quad \text{(boldface)}$$

Clearly, these two strategies do not belong to the same equivalence class as defined in (11), and thus they do not correspond to a single predictor. Also, they are disjoint, i.e. for any y

$$p(y|t_1) \neq 0 \Leftrightarrow p(y|t_2) = 0,$$

so they allow zero-error transmission of 1 bit. On the other hand, it can readily be verified that none of the effective noise channels, corresponding to an equivalence class of a *single* strategy, has a zero-error capacity. Therefore, for rates below 1 bit, the prediction scheme does not achieve the best error exponent and is therefore suboptimal.

V. Discussion

We have investigated the error exponent for modulo-additive noise channels with side information at the transmitter and obtained simple expressions for rates above the critical rate. As far as we know, this is the first treatment of error exponents of time varying channels with causal side information at the transmitter. It should be noted that our treatment assumed transmission with *fixed* block codes only. It is known that for the case of channels with feedback there is an essential difference between the performance of block codes and variable rate codes [5],[17],[20]. This distinction, as well as the effect of an input constraint, remain to be studied for channels with side information at the transmitter.

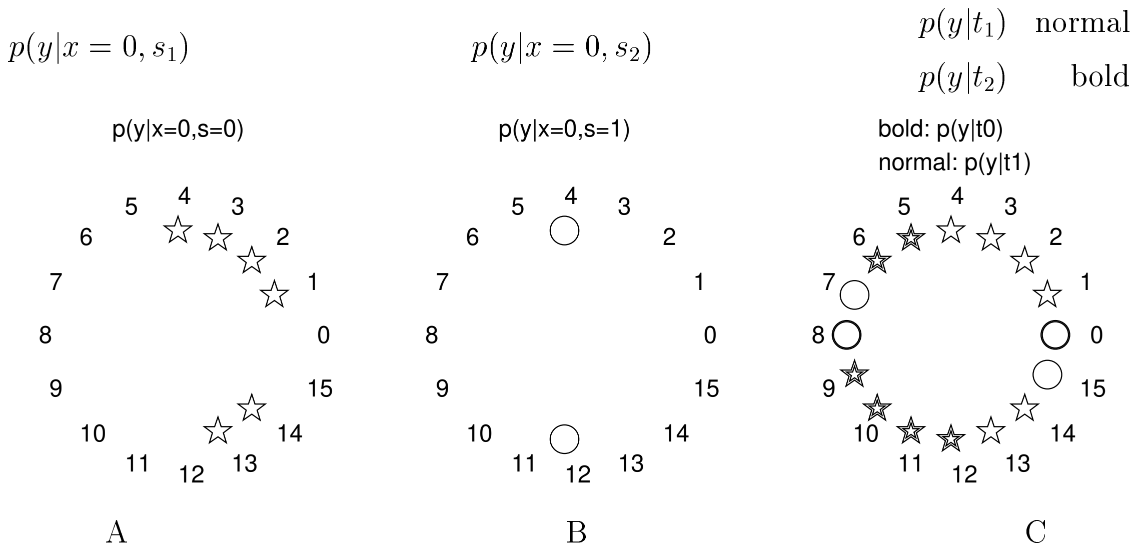

Figure 4: Non predictive strategies at low rate.

Acknowledgments

We thank Tamas Linder, Fady Alajaji and Neri Merhav for useful comments.

Appendix

A. Proof of Lemma 1

Define $\widetilde{Z} = Z - g(S,U)$, and denote $p_{Z_s}(a) = p_{Z|S}(a|s)$. Therefore

$$p_{\widetilde{Z}}(z) = \sum_{s,u} p(s,u) p_{Z_s}(z + g(s,u)).$$

Let s_0 be any state. We have

$$\begin{aligned}
p_{\widetilde{Z}}(z) &= \Big[\sum_{s \neq s_0} \sum_u p(s,u) p_{Z_s}(z + g(s,u))\Big] + \Big[p(s_0) \sum_u p(u|s_0) p_{Z_{s_0}}(z + g(s_0,u))\Big] \\
&= \sum_u p(u|s_0) \{ \sum_{s \neq s_0} \sum_{u'} p(s,u') p_{Z_s}(z + g(s,u')) + p(s_0) p_{z_{s_0}}(z + g(s_0,u)) \} \\
&= \sum_u p(u|s_0) p^{(u)}(z). \qquad (41)
\end{aligned}$$

where

$$p^{(u)}(z) \triangleq \{\sum_{s\neq s_0}\sum_{u'} p(s,u')p_{Z_s}(z+g(s,u'))\} + p(s_0)p_{z_{s_0}}(z+g(s_0,u)).$$

By the downward convexity $\cap$ of $F(\cdot)$ we have

$$F(\sum_u p(u|s_0)p^{(u)}(z)) \geq F(p^{(u_0)}(z)) \tag{42}$$

where $u_0 \triangleq \arg\min_u F(p^{(u)}(z))$. Define

$$g^*(s,u) = \begin{cases} g(s,u_0) & s = s_0 \\ g(s,u) & s \neq s_0 \end{cases} \quad \text{and} \quad \tilde{Z}^* = Z - g^*(U,S). \tag{43}$$

Therefore $p_{\tilde{Z}^*}(z) = p^{(u_0)}(z)$ and by (42) we have

$$F(\tilde{Z}) \geq F(\tilde{Z}^*).$$

Note that we thus eliminate the dependence on u in $g^*(s,u)$, for $s = s_0$. This process of "elimination" can be continued to the other states, thereby producing $g^{**}(s)$ (which does not depend on u) satisfying

$$F(Z - g(S,U)) \geq F(Z - g^{**}(S)) \geq F(Z - t^*(S)).$$

B. Ordered Average: Sufficient Conditions for ρ-SWIP Noise.

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:

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

For each i, let π_i^* be a permutation of $\{1,2\ldots m\}$ such that the resulting distribution $\{p_{i,\pi_i^*(j)}\}_{j=1}^m$ is ordered according to decreasing probability, i.e.

$$p_{i,\pi_i^*(j)} \geq p_{i,\pi_i^*(k)} \quad \text{for all} \quad k > j.$$

The "ordered average", which was introduced by Elias [7, 6] is defined by

$$p_j^{ord} = \sum_{i=1}^{n} w_i p_{i,\pi_i^*(j)}.$$

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. Let $\phi : [0,1] \longrightarrow [a,b] \quad a, b \in \mathcal{R}$ be convex $\cap$ and let $F(p_1, p_2, \ldots, p_m)$ be a function from the m-dimensional probability simplex into $\mathcal{R}$ defined by

$$F(p_1, p_2, \ldots, p_m) = \sum_{i=1}^{m} \phi(p_i). \tag{44}$$

Then we have:

Lemma 2 *The "ordered average", p_j^{ord}, minimizes $F(\sum_{i=1}^n w_i p_{i,\pi_i(j)})$ w.r.t all possible permutations $\pi_1, \ldots, \pi_n$.*

Taking $\phi(x) = x^\alpha$, $0 \leq \alpha \leq 1$, the lemma applies to the function $F(p_1, p_2, \ldots, p_m) = \sum_{i=1}^m p_i^\alpha$ and since the exponential function is monotonically increasing, the lemma also applies to the Rényi entropy. The lemma also applies to the Shannon entropy by taking $\phi(x) = -x \log x$. The latter is stated in Elias [6] and is proved in [7].

We now show that if $p(z|s)$ (as a function of z) is either unimodal and symmetric for each state s, or monotonically increasing/decreasing for all the states, then it is "ρ-SWIP" for all $\rho \geq 0$ and find $t^*(s)$ for this case. To this end, we view the effective $Z - t(S)$ as a mixture of the conditional noises $p(z|s_i)$, each of which is shifted by $t(s_i)$. Now for a general $p(z|s)$ such shifts cannot produce the optimal permutations that yield an ordered average. Nonetheless, for the above two cases choosing $t^*(s)$ so as to align the maxima together does yield an *ordered average*, irrespective of $p(s)$. Thus, by Lemma 2 this $t^*(s)$ minimizes the corresponding Rényi entropy of any order ρ and any $p(s)$. Therefore the noises are ρ-SWIP for any $\rho \geq 0$.

Proof of Lemma 2: Let $\pi_1, \ldots, \pi_n$ be any set of permutations and denote the resulting averaged distribution $\tilde{p}_j = \sum_{i=1}^n w_i p_{i,\pi_i(j)}$. Suppose these permutations do not define an ordered average (i.e. for at least one row i $\{p_{i,\pi_i(j)}\}_{j=1}^n$ is not ordered). We now show that $F(p_1^{ord}, p_2^{ord}, \ldots, p_m^{ord}) \leq F(\tilde{p}_1, \tilde{p}_2, \ldots, \tilde{p}_m)$, i.e., that this ordering yields a value of $F(\cdot)$ no smaller than that obtained by the ordered average. Consider the first two columns ($j = 1, j = 2$). Form a new ordering of the i-th row by interchanging $\tilde{p}_{i,1}$ with $\tilde{p}_{i,2}$ if $\tilde{p}_{i,2} > \tilde{p}_{i,1}$. Perform this for each row i. We thus form a new distribution $\bar{p}_{i,j}$ defined by

$$\bar{p}_{i,j} = \tilde{p}_{i,\bar{\pi}_i(j)}$$

where $\bar{\pi}_i$ is defined by the equations

$$\begin{aligned} \pi^{r,t}(j) &= j \quad j \neq r, t \\ \pi^{r,t}(r) &= t \\ \pi^{r,t}(t) &= r \end{aligned}$$

and

$$\bar{\pi}_i(j) = \begin{cases} \pi^{1,2}(j) & if \quad \tilde{p}_{i,2} > \tilde{p}_{i,1} \\ \\ j & if \quad \tilde{p}_{i,2} \leq \tilde{p}_{i,1} \end{cases} .$$

Thus the first column of $\bar{p}_{i,j}$ contains the greater of the entries in its row of the first two columns of $\bar{p}_{i,j}$. Let $\bar{p}_j = \sum_{i=1}^n w_i \bar{p}_{i,j}$. We want to show that $\bar{p}_j$ yields a value of $F(\cdot)$ that is at most as great as that corresponding to $\tilde{p}_j$. Now

$$F(\bar{p}_1, \bar{p}_2, \ldots, \bar{p}_m) - F(\tilde{p}_1, \tilde{p}_2, \ldots, \tilde{p}_m) = \phi(\bar{p}_1) + \phi(\bar{p}_2) - \Big(\phi(\tilde{p}_1) + \phi(\tilde{p}_2)\Big),$$

so that it suffices to ascertain that

$$\phi(\bar{p}_1) + \phi(\bar{p}_2) \leq \phi(\tilde{p}_1) + \phi(\tilde{p}_2).$$

But this is clearly true since the function $\phi(p) + \phi(1-p)$ is convex $\cap$ and symmetric around $p = \frac{1}{2}$. Next, repeat this process to column $j = 1$ with all the other columns $j = 3, \ldots, m$ successively. We thus obtain an ordering of the original distribution, with the first column containing entries that are the greatest in their respective rows. We next perform this procedure upon the second column with all the columns to its right and so forth. This process can be continued until we obtain an ordered average. In each step the resulting average cannot increase the value of $F(\cdot)$. Thus our claim is proved. $\square$

C. Proof of (34)

To compare the error exponents in (34), note that for $\alpha < 1$ the L_α "norm", $||\cdot||_\alpha$, is convex $\cap$, i.e. for any two vectors of positive numbers $\mathbf{u}, \mathbf{v}$ and $0 < \lambda < 1$

$$||\lambda \mathbf{u} + (1-\lambda)\mathbf{v}||_\alpha \geq \lambda ||\mathbf{u}||_\alpha + (1-\lambda)||\mathbf{v}||_\alpha \tag{45}$$

This assertion follows from the Minkowski inequality for $\alpha < 1$ [13] and from the homogeneity of the norm. Therefore

$$\sum_s p(s)||p_{Z_s}||_{\bar{\rho}} \leq ||\sum_s p(s)p_{Z_s}||_{\bar{\rho}} = ||p_Z||_{\bar{\rho}}, \tag{46}$$

and clearly $||p_{Z^*}||_{\bar{\rho}}$ should lie somewhere in between. Combining this with (33),(31) and (32), we establish that

$$E_r^{SI@BOTH} = E_r^{SI@REC} \geq E_r^{SI@TR} \geq E_r^{NOSI}. \tag{47}$$

D. Error exponent in terms of Rényi entropy (Eq. (14))

We now derive equation (14). Starting with (5) we have

$$\begin{aligned}
P_e^{opt}(n, R, \tilde{Q}_n) &\leq e^{nR\rho} \sum_{\mathbf{y}} \Big(\sum_{\mathbf{t}} Q_n(\mathbf{t}) P(\mathbf{y}|\mathbf{t})^{\frac{1}{1+\rho}} \Big)^{1+\rho} & (48)\\
&= e^{nR\rho} \sum_{\mathbf{y}} \Big(\sum_{\mathbf{c} \in \mathcal{X}^{\mathbf{n}}} \frac{1}{|\mathcal{X}|^n} P(\mathbf{y} - \mathbf{c}|\tilde{\mathbf{t}})^{\frac{1}{1+\rho}} \Big)^{1+\rho} & (49)\\
&= e^{nR\rho} \sum_{\mathbf{y}} |\mathcal{X}|^{-n(1+\rho)} \Big(\sum_{\mathbf{c}} P_{Z_{\tilde{\mathbf{t}}}}(\mathbf{c})^{\frac{1}{1+\rho}} \Big)^{1+\rho} & (50)\\
&= e^{nR\rho} |\mathcal{X}|^n |\mathcal{X}|^{-n(1+\rho)} \Big(\sum_{\mathbf{c}} P_{Z_{\tilde{\mathbf{t}}}}(\mathbf{c})^{\frac{1}{1+\rho}} \Big)^{1+\rho} & (51)\\
&= e^{nR\rho} |\mathcal{X}|^{-n\rho} \Big(\sum_{\mathbf{c}} P_{Z_{\tilde{\mathbf{t}}}}(\mathbf{c})^{\frac{1}{1+\rho}} \Big)^{1+\rho} & (52)
\end{aligned}$$

where (49) follows from (12) and since since $\tilde{Q}_n$ equals $\frac{1}{|\mathcal{X}|^n}$ for each member of the class; (50) follows by change of variable of summation, $\mathbf{c}' = \mathbf{y} - \mathbf{c}$; (51) follows since the summation over $\mathbf{y}$ is independent of $\mathbf{y}$. Using the definition of the Rényi entropy (7), by taking the logarithm of (52) and dividing by $-\frac{1}{n}$ we obtain (14).

References

[1] F. Alajaji and N.Whalen. The capacity-cost function of discrete additive noise channels with and without feedback. *IEEE Trans. Information Theory*, IT-46:1131–1140, May 2000.

[2] M.H.M. Costa. Writing on dirty paper. *IEEE Trans. Information Theory*, IT-29:439–441, May 1983.

[3] I. Csiszar. Generalized Cutoff Rates and Rényi's Information Measures. *IEEE Trans. Information Theory*, IT-41, No. 1, January 1995.

[4] I. Csiszar and J. Korner. *Information Theory - Coding Theorems for Discrete Memoryless Systems*. Academic Press, New York, 1981.

[5] E.A.Haroutunian. A lower bound of the probability of error for channels with feedback (in russian). *Problems of Information Transmission*, 13:36–44, 1977.

[6] P. Elias. Predictive coding. *IRE Trnsc. on Info. Theory*, pages 16–33, March 1955.

[7] P. Elias. *Predictive Coding*. PhD thesis, Harvard, May 1950.

[8] E.M.Gadibulin. Bounds on the probability of error for certain channels with memory. *Problems of Inform. Trans.*, 5, No. 1:33–38, 1969.

[9] U. Erez and R. Zamir. Noise prediction for channel coding with side-information at the transmitter. *IEEE Trans. Information Theory*, 46:1610–1617, July 2000.

[10] F.M.J.Willems. On gaussian channels with side information at the transmitter. In *Proc. of the Ninth Symposium on Information Theory in the Benelux*, Enschede, The Netherlands, 1988.

[11] R. G. Gallager. *Information Theory and Reliable Communication*. Wiley, New York, N.Y., 1968.

[12] S.I. Gelfand and M. S. Pinsker. Coding for channel with random parameters. *Problemy Pered. Inform. (Problems of Inform. Trans.)*, 9, No. 1:19–31, 1980.

[13] G.Pólya G.H.Hardy, J.E.Littlewood. *INEQUALITIES*. Cambridge University Press, Cambridge, second edition, 1959.

[14] E.A. Haroutunian and M.E.Haroutinian. E-capacity upper bound for a channel with random parameter. *Problems of Control and Information Theory*, 17:99–105, 1988.

[15] C. Heegard and A. El Gamal. On the capacity of computer memory with defects. *IEEE Trans. Information Theory*, IT-29:731–739, Sept. 1983.

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

[17] G. D. Forney Jr. Exponential error bounds for erasure, list, and decision feedback schemes. *IEEE Trans. Information Theory*, IT-14:206–220, 1968.

[18] T. Linder. Private communication.

[19] R.J. McEliece and W.E. Stark. Channels with block interference. *IEEE Trans. Information Theory*, 30:44–53, 1984.

[20] M.V.Burnashev. Information transmission over a discrete channel with feedback. Random transmission time (in russian). *Problems of Information Transmission*, 12:10–30, 1976.

[21] C. E. Shannon. Channels with side information at the transmitter. *IBM Journal Research and Development*, 2:289–293, Oct. 1958.

[22] Z.Rached, F.Alajaji and L.L. Campbell. Rényi's entropy rate for sources with memory. *Proc. CISS'99, pp. 613-618, March 1999, Baltimore, MD.*