Noise Prediction for Channels with Side Information at the Transmitter

Uri Erez and Ram Zamir, *Senior Member, IEEE*

***Abstract*—The computation of channel capacity with side information at the transmitter side (but not at the receiver side) requires, in general, extension of the input alphabet to a space of "strategies," and is often hard. We consider the special case of a discrete memoryless modulo-additive noise channel $Y = X + Z_S$, where the encoder observes *causally* the random state $S \in \mathcal{S}$ that governs the distribution of the noise Z_S. We show that the capacity of this channel is given by**

$$C = \log |\mathcal{X}| - \min_{t:\mathcal{S}\to\mathcal{X}} H(Z_S - t(S)).$$

This capacity is realized by a state-independent code, followed by a shift by the "noise prediction" $t_{\min}(S)$ that minimizes the entropy of $Z_S - t(S)$. If the set of conditional noise distributions $\{p(z \mid s),\, s \in \mathcal{S}\}$ is such that the optimum predictor $t_{\min}(\cdot)$ is independent of the state weights, then C is also the capacity for a *noncausal* encoder, that observes the entire state sequence in advance. Furthermore, for this case we also derive a simple formula for the capacity when the state process has memory.

***Index Terms*—Optimum transmitter, prediction with minimum error entropy, side information, time-varying channels.**

I. INTRODUCTION

The model of an underlying state process that governs the channel behavior fits well many communication links. Examples include wireless communication (fading channel), telephone lines (time-varying filter channel), partial time/band jamming, magnetic memory with defective cells, and more. To analyze such channels it is often useful, and sometimes makes sense, to assume that the channel states are available as Side Information (SI) to the encoder, to the decoder, or to both [16], [11], [2], [8].

One "good" reason for the latter two assumptions is that they simplify the calculation of channel capacity, and the design of optimum encoding/decoding strategies. A common example for that is a *slow* fading Gaussian channel, in which case the receiver can adapt to the fading level and perform "soft" decision [2], [15].

Before turning to the case of side information at the *encoder side only*, let us consider the channel model illustrated in Fig. 1. We use $\mathcal{X}$, $\mathcal{Y}$, and $\mathcal{S}$ to denote the input, output, and state alphabets of the channel, respectively. Given the states $s_1, s_2, \cdots$, the channel is memoryless with transition distribution $p(y \mid x, s)$, i.e.,

$$p\left(y_1^n \mid x_1^n, s_1^n\right) = \prod_{i=1}^{n} p(y_i \mid x_i, s_i)$$

where y_1^n denotes the sequence $y_1 \cdots y_n$; furthermore, the stationary state process $S_1, S_2, \cdots$ is statistically independent of the message to be sent, and it completely captures the memory, if there is any, in the channel, i.e.,

$$p\left(s_{n+1} \mid x_1^n, y_1^n, w, s_1^n\right) = p\left(s_{n+1} \mid s_1^n\right).$$

These assumptions model well a "pure noisy state," e.g., as in the interference, fading, or jamming channels, but they exclude an "input-dependent state," like in the intersymbol interference (ISI) channel [6] (in the latter case the state sequence depends on the message w).

The respective states of switches *A* and *B* in Fig. 1 give rise to the following distinct cases[1]:

- **Case I—SI not available**: The channel is regarded as an ordinary channel, given by the mixture of the component channels, e.g., for a memoryless state sequence it is a DMC with transition probability

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

  The capacity is denoted C_{NOSI}.
- **Case II—SI available at receiver only (*A* closed, *B* open)**: The state is regarded as an additional channel output, and the capacity is given by (see, e.g., [16])

  $$C_{\mathrm{SI@REC}} = \max_{p(x)} I(X; S, Y) = \max_{p(x)} I(X; Y \mid S). \tag{1}$$
- **Case III—SI available at transmitter and receiver alike (*A*, *B* closed)**: The capacity is given by the weighted sum of the individual capacities (see, e.g., [16])

  $$\begin{aligned} C_{\mathrm{SI@BOTH}} &= \sum_{s\in\mathcal{S}} p(s) \max_{p(x \mid s)} I(X; Y \mid S = s) \\ &\triangleq \sum_{s\in\mathcal{S}} p(s) C_s \end{aligned} \tag{2}$$

  where C_s is the capacity at state s. (Note that in Cases II and III the capacity depends only on the marginal distribution of the states.)
- **Case IV—SI available at transmitter only (*A* open, *B* closed)**: The capacity is denoted $C_{\mathrm{SI@TR}}$; we recall its formula in the next section.

The situation where the transmitter has access to side information which is not available to the decoder (Case IV above) is less common in applications. Furthermore, its solution, in terms of capacity and coding, is rather involved.

This problem divides into two categories, according to whether the encoder observes the state process *causally* or *anticipates future states*. In the causal case, considered by Shannon [14], the encoder maps the message $w \in \{1, 2, \cdots 2^{nR}\}$ into $\mathcal{X}^n$ using functions

$$x_i = f_i\left(w, s_1^i\right), \qquad 1 \le i \le n \tag{3}$$

where $s_1^i = s_1, \cdots, s_i$ are the states up to time i. In the noncausal case, considered by Gelfand and Pinsker [7], the encoder observes the entire state sequence before generating the code sequence, thus

$$x_i = f_i\left(w, s_1^n\right), \qquad 1 \le i \le n. \tag{4}$$

In both cases, the receiver decodes the message w from the whole received vector as $\hat{w} = g(y_1^n)$.

Very few explicit solutions exist for the capacity of channels with side information at the encoder; see, e.g., [3], [2], [12], and [8]. This is due in part to the computational complexity of the solution. For the case of discrete memoryless channels with a causal encoder, Shannon [14] showed that $C_{\mathrm{SI@TR}}$ is given by the ordinary capacity of a discrete

---

Manuscript received May 18, 1998; revised October 20, 1999. The material in this paper was presented in part at the IEEE International Symposium on Information Theory, MIT, Cambridge, MA, August 1998.

The authors are with the Department of Electrical Enggineering–Systems, Tel-Aviv University, Ramat-Aviv 69978 Israel (e-mail: erez@eng.tau.ac.il; zamir@eng.tau.ac.il).

Communicated by I. Csiszár, Associate Editor for Shannon Theory.
Publisher Item Identifier S 0018-9448(00)04646-0.

[1] As shown in [2], a "mixture" of these cases, i.e., a configuration where the encoder and the decoder have access to (possibly different) noisy versions of the state process, is effectively equivalent to one of Cases I–IV.

0018–9448/00$10.00 © 2000 IEEE

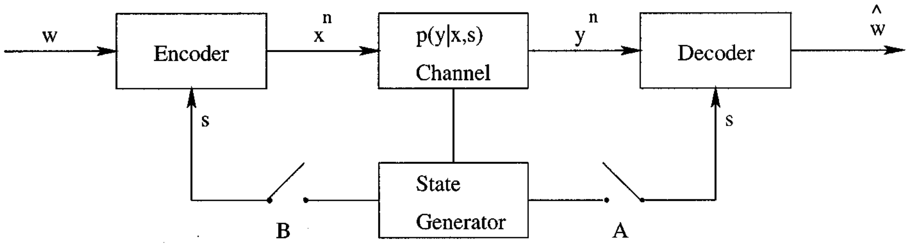

Fig. 1. Channel configuration.

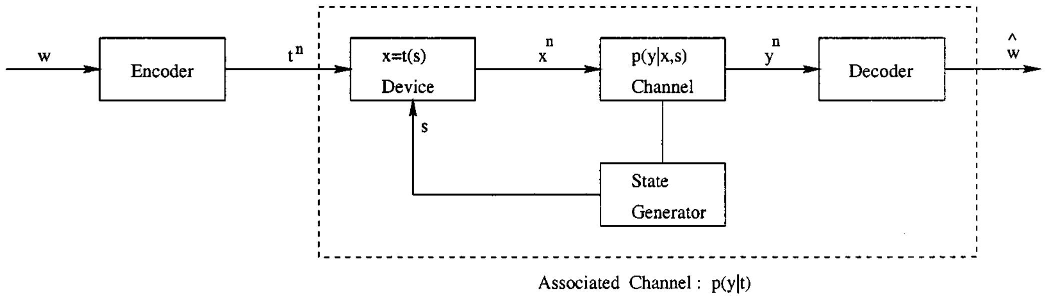

Fig. 2. The associated discrete memoryless channel.

memoryless channel (DMC) with an *extended input alphabet* of size $|\mathcal{X}|^{|\mathcal{S}|}$, commonly interpreted as a space of "strategies. "Even more involved are the general solutions for a state process with memory [9], [10], and for noncausal side information [7].

In this correspondence we find a simple and operational solution for the capacity of discrete channels of the form $Y = X + Z$, where the additive noise Z is correlated with a random variable S, which is available as side information to the encoder. Here $Z \in \mathcal{X}$, and the $+$ sign denotes modulo-$|\mathcal{X}|$ addition. The variable S may be thought of as the "channel state," or as a "noisy version" of the channel noise Z. No "power" constraint is imposed on the transmitter.

For a memoryless state process available causally to the encoder, we show that

$$C_{\mathrm{SI@TR}} = \log|\mathcal{X}| - \min_{t:\mathcal{S}\to\mathcal{X}} H(Z - t(S))$$

where $H(\cdot)$ denotes entropy. Furthermore, the optimal encoder is composed of a state-independent code, followed by a shift by the "minimum-error-entropy noise prediction" $t_{\min}(S)$ that achieves the minimum above. Prediction with minimum error entropy was introduced by Elias [4], in the context of predictive source coding. In general, the optimum predictor $t_{\min}(\cdot)$ depends on the set of conditional noise distributions $\{p(z \mid s), s \in \mathcal{S}\}$, and on the state distribution ("state weights") $p(s)$. We identify the property of State Weight Independent Prediction (SWIP), which means that $\{p(z \mid s), s \in \mathcal{S}\}$ is such that the optimizing function $t_{\min}(\cdot)$ does not depend on $p(s)$. For example, if $p(z \mid s)$ is unimodal and symmetric for each S then $p(z \mid s)$ is SWIP. If the pair (S, Z) satisfies the SWIP property, then *noncausal side information does not increase capacity*, i.e., the anticipatory encoder coincides capacity-wise and structure-wise with the causal encoder. Furthermore, in this case the same predictor is also optimal when there is memory in the state sequence.

In the next section we present our main result regarding the capacity of a memoryless channel with causal state information at the encoder. In Section III we examine the properties of the minimum error entropy prediction function $t_{\min}(\cdot)$, and discuss the SWIP property. For channels satisfying the SWIP property, we extend in Section IV the capacity formula for noncausal state information at the encoder and for channels with a stationary state process.

II. CAUSAL SIDE INFORMATION AT THE ENCODER

Consider encoding with causal side information at the encoder as defined in (3). For the general memoryless channel $p(y \mid x, s)$ shown in Fig. 1, Shannon [14] showed that $C_{\mathrm{SI@TR}}$ is equal to the regular capacity of the associated DMC illustrated in Fig. 2. The input alphabet of the associated channel, denoted $\mathcal{T}$, is the set of all possible mappings

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

which we refer to as *strategies* or *strategy functions*. We may describe each strategy $t(s) \in \mathcal{T}$ by the vector $(x^1, x^2, \cdots, x^{|\mathcal{S}|})$, i.e., $t(s) = x^s$ for $s = 1, \cdots, |\mathcal{S}|$. Therefore, $|\mathcal{T}| = |\mathcal{X}|^{|\mathcal{S}|}$. The output y of the associated channel is related to the input t according to the transition probability

$$p(y \mid t) \triangleq \sum_{s} p(s) p(y \mid x = t(s), s) \tag{5}$$

and also

$$p\left(y_1^n \mid t_1^n\right) = \prod_{i=1}^{n} p(y_i \mid t_i). \tag{6}$$

Thus the capacity with side information at the transmitter is given by [14]

$$C_{\mathrm{SI@TR}} = \max_{p(t)} I(T; Y) \tag{7}$$

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

Note that at most $|\mathcal{Y}|$ of the strategies need be given positive probability in order to achieve capacity [6, Ch. 4]. Therefore, if $\mathcal{X} = \mathcal{Y}$ then at most $|\mathcal{X}|$ of the $|\mathcal{X}|^{|\mathcal{S}|}$ strategies need be given positive probability. However, in the general case one does not know in advance which of the strategies are to be used to achieve capacity. We next treat a class of channels, namely, the class of symmetric, or modulo-additive noise channels, for which it is easy to identify this set of "active" strategies.

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

$$Y = X + Z \tag{8}$$

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

$$p(y \,|\, x, s) = p_{Z_s}(y - x).$$

Thus the transition probability of the associated channel defined in (5) is given by

$$p(y \,|\, t) = \sum_s p(s) p_{Z_s}(y - t(s)) = \Pr(Z + t(S) = y). \tag{9}$$

Theorem 1 (Memoryless Causal Case): The capacity of the discrete memoryless additive noise channel $Y = X + Z$ defined above, with causal side information S at the encoder, is given by

$$C_{\text{SI@TR}} = \log|\mathcal{X}| - H_{\min} \tag{10}$$

where

$$H_{\min} \triangleq \min_{t \in \mathcal{T}} H(Z - t(S)). \tag{11}$$

Note that (9) implies that

$$H_{\min} = \min_{t \in \mathcal{T}} H(Y \,|\, T = t) = \min_{t \in \mathcal{T}} H(Z + t(S)) \tag{12}$$

because the minimization of the entropy of $Z + t(S)$ and of $Z - t(S)$ is the same.

Proof: We first show the converse part, i.e.,

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

Since

$$H(Y) \le \log|\mathcal{X}| \text{ and } H(Y \,|\, T) \ge \min_{t \in \mathcal{T}} H(Y \,|\, T = t) = H_{\min} \tag{13}$$

we have

$$I(T; Y) = H(Y) - H(Y \,|\, T) \le \log|\mathcal{X}| - H_{\min}$$

for any distribution on $\mathcal{T}$, and the converse follows from (7).

We next show the direct part, i.e.,

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

Let t^* denote a strategy, i.e., a mapping from $\mathcal{S}$ to $\mathcal{X}$, for which $H(Y \,|\, T = t^*) = H_{\min}$. Define the following *class of strategies*:

$$\mathcal{T}^* \triangleq \{t_j\} \quad \text{where } t_j(s) = t^*(s) + j, \qquad j = 1 \cdots |\mathcal{X}|. \tag{14}$$

From (9) we see that

$$p(y \,|\, t_j) = \Pr(Z + t_j(S) = y) = \Pr(Z + t^*(S) = y - j) \tag{15}$$

i.e., $p(y \,|\, t_j)$ is the transition probability $p(y \,|\, t^*)$ shifted (modulo-$|\mathcal{X}|$) by j. This clearly implies $H(Y \,|\, t_j) = H(Y \,|\, t^*) = H_{\min}$ for all j. Furthermore, choosing T distributed uniformly within $\mathcal{T}^*$ (and zero on strategies not in $\mathcal{T}^*$) induces a *uniform* distribution on $\mathcal{Y}$. Thus for such T we have equality in both inequalities in (13), and the direct part follows. □

Note that by restricting the input alphabet to the set $\mathcal{T}^*$, the resulting channel from $\mathcal{T}^*$ to $\mathcal{Y}$ can be viewed as an additive noise channel, whose alphabet is $\mathcal{X}$, and whose noise is distributed as

$$\tilde{Z} \triangleq Z + t^*(S) \triangleq Z - t_{\min}(S) \tag{16}$$

where

$$t_{\min}(\cdot) \triangleq \arg \min_{t: \mathcal{S} \to \mathcal{X}} H(Z - t(S)). \tag{17}$$

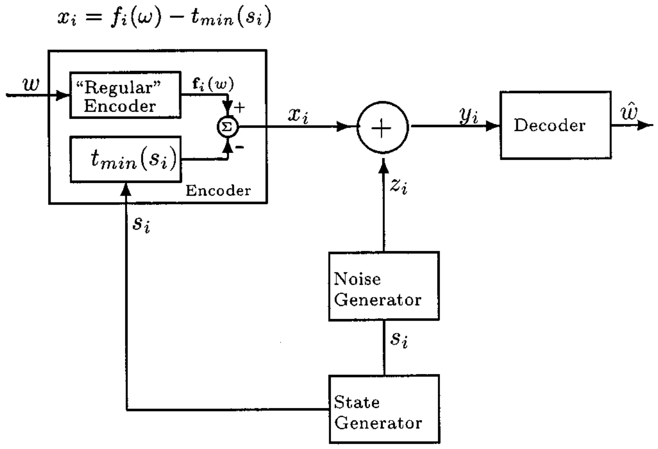

Fig. 3. Instantaneous prediction encoding scheme.

Thus we have simplified the associated channel into a $|\mathcal{X}|$-input/$|\mathcal{X}|$-output channel. Note that actually any single strategy function t, not necessarily t^*, generates a class of $|\mathcal{X}|$ strategies as in (14), which induces an additive noise channel as in (16).

Theorem 1 implies that the optimal transmitter with side information has a simple modular structure, consisting of an ordinary (i.e., state-independent) Shannon code for a symmetric channel, followed by a shift by $t_{\min}(s)$, as shown in Fig. 3. Specifically, suppose a code $\mathcal{C} = \{(c_{w1}, \cdots, c_{wn})\}_{w=1}^{2^{nR}}$ is drawn at random, using a uniform distribution over $\mathcal{X}$, with rate $R < C_{\text{SI@TR}}$. Given a message $1 \le w \le 2^{nR}$, and a channel state s_i at time i, the transmitter outputs

$$x_i = c_{wi} - t_{\min}(s_i), \qquad i = 1 \cdots n. \tag{18}$$

The receiver decodes w from $Y_1 \cdots Y_n$ as if the channel were an ordinary DMC with additive noise $\tilde{Z}$. It then follows from Theorem 1 and the forward channel coding theorem [6] that w is decoded reliably with high probability. In practice, any "good" code for a symmetric channel (with a decoder optimized to the noise $\tilde{Z}$) can replace the random code in the scheme above. We call the configuration of (18) "the instantaneous-predictor encoder"; the relation to prediction will become clearer in the next section.

This structure of the optimum transmitter leads to the following interpretation. Since the receiver does not know S, it sees a channel with an effective noise $\tilde{Z} = Z - t_{\min}(S)$. Shifting Z by $t_{\min}(S)$ thus makes the effective noise the least harmful for the uninformed receiver. Note also that since

$$I(S; \tilde{Z}) = H(\tilde{Z}) - H(\tilde{Z} \,|\, S) = H(\tilde{Z}) - H(Z \,|\, S)$$

minimizing $H(Z - t(S)) = H(\tilde{Z})$ is equivalent to minimizing the information carried by the effective noise about the channel state.

To conclude this section, we return to Cases I–IV of the switches in Fig. 1, and compare the various capacities with and without side information of an additive noise channel. Since choosing a uniform distribution for X achieves capacity for any discrete additive noise channel, we have for Cases II and III

$$\begin{aligned} C_{\text{SI@REC}} = C_{\text{SI@BOTH}} &= \log|\mathcal{X}| - \sum_s p(s) H(Z_s) \\ &= \log|\mathcal{X}| - H(Z \,|\, S) \end{aligned} \tag{19}$$

and for Case I

$$C_{\text{NOSI}} = \log|\mathcal{X}| - H(Z). \tag{20}$$

We thus have the following chain of inequalities:

$$C_{\text{NOSI}} \le C_{\text{SI@TR}} \le C_{\text{SI@REC}} = C_{\text{SI@BOTH}} \tag{21}$$

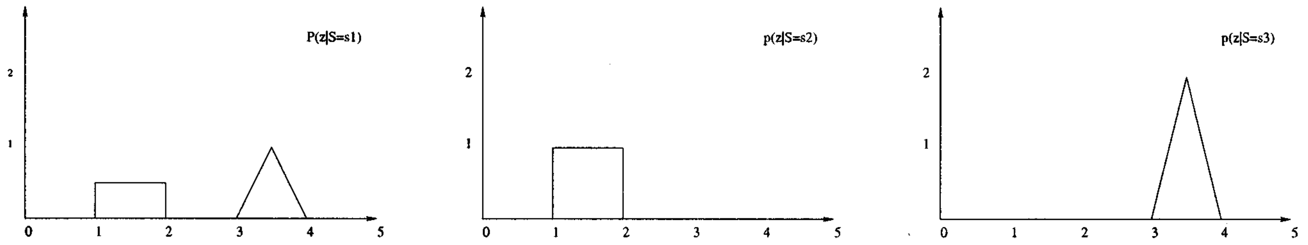

Fig. 4. Example of conditional noise distributions which do not satisfy the SWIP property.

where

$$C_{\mathrm{SI@BOTH}} - C_{\mathrm{SI@TR}} = I(S;\tilde{Z})$$
$$C_{\mathrm{SI@BOTH}} - C_{\mathrm{NOSI}} = I(S;Z). \qquad (22)$$

We have equality in the second inequality in (21), i.e., $C_{\mathrm{SI@REC}} = C_{\mathrm{SI@TR}}$, iff the distributions of Z_s, $s \in \mathcal{S}$ differ by a shift only, in which case the optimum $\tilde{Z}$ is statistically independent of S. We have equality in the first inequality in (21), i.e., $C_{\mathrm{SI@TR}} = C_{\mathrm{NOSI}}$, iff $H(\tilde{Z}) = H(Z)$, i.e., iff the optimal shifts $\{t_{\min}(s), s \in \mathcal{S}\}$ are the set of zero shifts.

III. PREDICTION WITH MINIMUM ERROR ENTROPY AND EXAMPLES

As discussed above, the function $t_{\min}(\cdot)$ minimizes the entropy of the difference $Z - t_{\min}(S)$; we view $t_{\min}(S)$ as the "prediction of Z from S with minimum error entropy." In this section we review some properties of minimum error entropy prediction that will be useful in the next section, in particular the SWIP property. We then illustrate these concepts with two examples.

Elias [4] arrives at this very solution in his treatment of entropy-coded predictive source coding. Specifically, for a discrete stationary ergodic source X_i, one wishes to find the prediction function $f(x_{i-1}, x_{i-2}, \cdots)$ such that the entropy of the error, $E_i \triangleq X_i - f(X_{i-1}, X_{i-2}, \cdots)$, is minimized. This will minimize the coding rate assuming that the process $E_1, E_2, \cdots$ is entropy-coded according to its *marginal* distribution (ignoring the possible residual memory in the sequence $E_1, E_2 \cdots$). The analogy to the problem we study is clear. The past source samples play the role of the current state, while the conditional random variable $X_i \mid x_{i-1}, x_{i-2}, \cdots$ plays the role of Z_s.

Prediction under the minimum error entropy criterion features some interesting properties, not found in the more common criteria for prediction and estimation, the minimum mean-squared error criterion $\min_g E[Z - g(S)]^2$ (continuous case), and the minimum error frequency criterion $\min_g \Pr[Z \neq g(S)]$ (discrete case). While for these criteria the optimum predictor $g(s)$ is a function only of the conditional noise distribution $p(z \mid s)$ for each s, the optimal shift $t(s)$ of the minimum error entropy predictor depends, in general, on the *entire* set of conditional noise distributions $\{p(z \mid s)\}$, and on the *state probabilities* $p(s)$ (the "state weights").

To illustrate this property, consider the set of conditional noise distributions shown in Fig. 4. Define the following weight distributions on the states:

$$\boldsymbol{w}_1 = \left(\frac{1}{2}, \frac{1}{2}, 0\right) \quad \boldsymbol{w}_2 = \left(\frac{1}{2}, 0, \frac{1}{2}\right) \quad \boldsymbol{w}_3 = \left(0, \frac{1}{2}, \frac{1}{2}\right)$$

We shall now find the corresponding optimal shift functions and show that no single strategy can be optimal for all three weights. Without loss of generality we set $t_{\min}(s_1) = 0$ for all strategies below. For $(p(s_1), p(s_2), p(s_3)) = \boldsymbol{w}_1$, the vector $(t_{\min}(s_1), t_{\min}(s_2), t_{\min}(s_3))$ is given by $(0, 0, *)$, where $*$ means that the corresponding letter may be arbitrarily chosen. Any other strategy will induce strictly higher entropy for $Z - t(S)$. For $\boldsymbol{w}_2$, the vector of $t_{\min}(\cdot)$ is given by $(0, *, 0)$. Finally, for $\boldsymbol{w}_3$, the vector of $t_{\min}(\cdot)$ is given by $(0, * - 2, *)$ (e.g., $(0, -2, 0)$ or $(0, 0, 2)$). It is easy to see that no single strategy is consistent with all three cases. Note that by the continuity of the entropy function $H(Z - t(S))$ as a function of the state weight vector, the above argument can easily be extended to examples of state weight vectors whose components are all nonzero.

Definition 1 (The State Weight Independent Prediction (SWIP) Property): Fix the set of conditional noise distributions $\{p(z \mid s), z \in \mathcal{X}, s \in \mathcal{S}\}$. If the same function $t = t_{\min}(\cdot)$ minimizes the entropy of $Z - t(S)$ for any state weights $\{p(s), s \in \mathcal{S}\}$, then we say that the noise satisfies the State Weight Independent Prediction Property, or in short, "the noise is SWIP."

We shall now demonstrate the existence of SWIP noises. For that we recall the concept of *ordered average*, introduced by Elias [4]. 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

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

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 or equal to 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.

Following Elias, we identify two simple cases of interest in which the optimal shifts are independent of $p(s)$, i.e., cases of SWIP noises

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 $x_0(s)$

$$p_{Z_s}(x) = 0, \qquad \text{for } x < x_0(s)$$
$$p_{Z_s}(x) \geq p_{Z_s}(y), \qquad \text{for } x_0(s) \leq x < y.$$

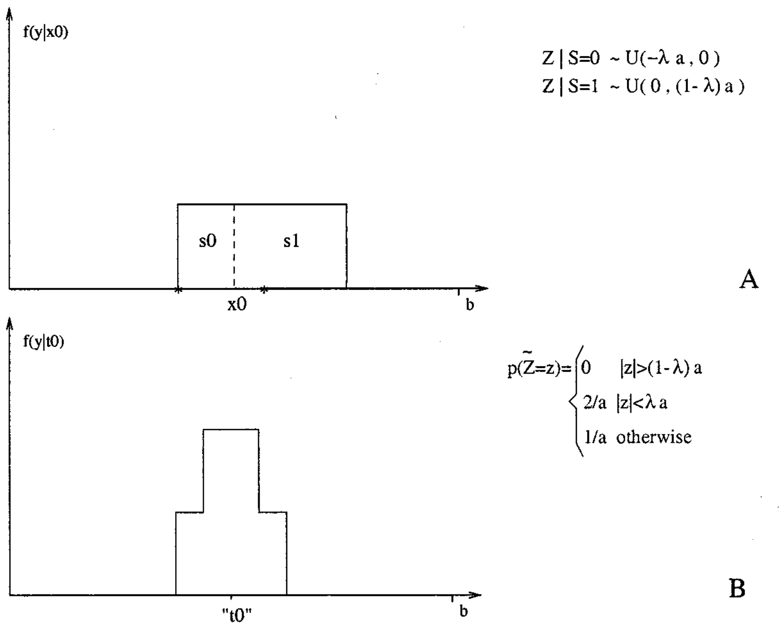

Fig. 5. Example II.

In both cases the assertion follows from the observation that choosing $t_{\min}(s)$ so as to align the maxima together yields an ordered average irrespective of $p(s)$.

The following two lemmas characterize some important properties of minimum error entropy prediction which we use for proving our main results in the next section.

Lemma 1 (Irrelevancy): Let U be a random variable defined over the finite alphabet $\mathcal{U}$ such that U, S, Z form a Markov chain $U \leftrightarrow S \leftrightarrow Z$. Then, for any function $g : \mathcal{U} \times \mathcal{S} \to \mathcal{X}$

$$H(Z - g(S, U)) \geq H(Z - t_{\min}(S))$$

where $t_{\min}$ is defined in (17). Thus U is irrelevant for minimum error entropy prediction of Z.

The surprisingly nonstraightforward proof is given in the Appendix.

Lemma 2 (Conditional Irrelevancy for SWIP Noise): Let U, V be random variables defined over the finite alphabets $\mathcal{U}, \mathcal{V}$, respectively, such that U, V, S, Z form a Markov chain $(U, V) \leftrightarrow S \leftrightarrow Z$. If the noise Z satisfies the SWIP property above, then for any function $g : \mathcal{U} \times \mathcal{S} \to \mathcal{X}$

$$H(Z - g(S, U) \,|\, V) \geq H(Z - t_{\min}(S) \,|\, V) \tag{23}$$

where $t_{\min}$ is defined in (17). Thus for SWIP noise U is conditionally irrelevant for minimum error entropy prediction of Z.

Proof: By Lemma 1 we have for any $v \in \mathcal{V}$

$$H(Z - g(S, U) \,|\, V = v) \geq H(Z - t_{\min}(S) \,|\, V = v)$$

and since the noise is SWIP the function $t_{\min}$ is independent of $p(s \,|\, v)$ and therefore independent of the value of v. Thus (23) follows by taking expectation over V. □

Note that (23) does not hold in general for non-SWIP noises, i.e., for noises for which $t_{\min} = t_{\min}(s, p(s))$ *does* depend on the state weights. For such noises, if the state weights $p(s \,|\, v)$ depend on the value of the condition V, and if U depends on V, then the function $g(s, u)$ can take advantage of these dependencies in order to reduce the conditional entropy of $Z - g(U, S)$ given V beyond that achieved by a function of S alone.

A. Examples

We turn to consider a few examples of finite-state additive noise channels, their associated minimum error entropy predictors, and their capacities. We first note that the results of Theorem 1 hold equally well when the alphabet $\mathcal{X}$ is a continuous interval $[0, b] \in \mathcal{R}^1$, with addition performed modulo b, and regular entropy replaced by differential entropy.

Example I: Binary-Symmetric Channel (BSC) with crossover probability θ_s. Here

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

while

$$C_{\mathrm{NOSI}} = 1 - h\left(\sum_s p(s)\theta_s\right).$$

In the case where the SI consists of two states, θ_1 and θ_2, we have

$$C_{\mathrm{SI@TR}} = \begin{cases} C_{\mathrm{NOSI}}, & \text{if } \theta_1, \theta_2 \leq \dfrac{1}{2} \\ 1 - h[p(s=1)\theta_1 + p(s=2)\bar{\theta_2}], & \text{if } \theta_1 \leq \dfrac{1}{2} \leq \theta_2 \end{cases} \tag{24}$$

where $\bar{\theta_2} = 1 - \theta_2$. This follows since binary noise falls into Case b) of SWIP noise above, so Elias' ordered average condition applies. The extension of the result to more than two states is obvious; capacity is achieved by "flipping" each θ_i that is greater than $\frac{1}{2}$.

Example II: A channel with $\mathcal{X} = [0, b]$ and transition distribution as shown in Fig. 5(*A*). We have

$$\begin{aligned} H(Z \,|\, S) &= p(s_1)H(Z \,|\, S = s_1) + p(s_0)H(Z \,|\, S = s_0) \\ &= \lambda \log \lambda a + (1 - \lambda) \log (1 - \lambda) a = \log a - h(\lambda) \end{aligned}$$

while

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

where to obtain $\tilde{Z}$ we optimized the shifts according to Elias' conditions for symmetric unimodal distributions (Case a) above). Therefore,

$$C_{\mathrm{SI@REC}} = \log\frac{b}{a} + h(\lambda) \ \text{ while } \ C_{\mathrm{SI@TR}} = \log\frac{b}{a} + 2\lambda.$$

The corresponding distribution of the noise $\tilde{Z}$ is depicted in Fig. 5(*B*).

IV. NONCAUSAL SIDE INFORMATION AND STATE PROCESS WITH MEMORY

In this section we generalize our results in two directions, first to the case where the state process has memory and secondly to a noncausal encoder as discussed in Section I. Jelinek treated channels having memory with causal side information at the transmitter in [9], [10]. However, the expressions he arrives at are given in terms of complex double limits. This remains the case even when they are applied to additive noise channels, thus giving little new insight. Gelfand and Pinsker [7] studied the problem of coding for a discrete memoryless channel with channel states available noncausally to the encoder. Their solution, as Shannon's solution for a causal encoder, involves maximization over an extended alphabet, and moreover the optimization is done over conditional input probabilities (given the state). We shall confine our treatment to SWIP noise which will allow us to treat both problems together and to obtain simple expressions for the capacities for both cases.

For our purpose we can proceed directly from the definition of the encoding function (4). Consider an encoder which observes the entire state sequence $\boldsymbol{S} = S_1, \cdots, S_n$ prior to encoding. The state process is a general stationary process, such that the noise sample Z_i is conditionally independent of $(X_1 \cdots X_n, S_1 \cdots S_{i-1}, S_{i+1} \cdots S_n)$ given S_i. The encoding and decoding functions now assume the form

$$\boldsymbol{x} = \boldsymbol{f}(w, \boldsymbol{s}) \quad \text{and} \quad \hat{w} = g(\boldsymbol{y}) \tag{25}$$

where boldface denotes n-vectors, i.e.,

$$\begin{aligned}\boldsymbol{x} &= x_1, \cdots, x_n\\ \boldsymbol{y} &= y_1, \cdots, y_n\end{aligned}$$

and

$$\boldsymbol{f}(\cdot) = f_1(\cdot), \cdots, f_n(\cdot).$$

We first give a (*non* "single-letter") upper bound on the capacity of general additive noise channels.

Lemma 3: Assume the discrete additive noise channel defined in (8). The rate R of any code with noncausal side information $\boldsymbol{S}$ at the encoder (as defined in (25)) and error probability $\Pr(\hat{W} \neq W) \le \epsilon$, satisfies

$$R \le \log|\mathcal{X}| - H_{\min}^{(n)} + \epsilon'$$

where $\epsilon' \to 0$ as $\epsilon \to 0$ uniformly in the code length n, and

$$H_{\min}^{(n)} \triangleq \frac{1}{n}\min_{\boldsymbol{t}:\mathcal{S}^n \to \mathcal{X}^n} H(\boldsymbol{Z} - \boldsymbol{t}(\boldsymbol{S})). \tag{26}$$

Proof: By Fano's inequality [6], we have

$$H(W \,|\, \boldsymbol{Y}) \le h(\epsilon) + \epsilon n R \le h(\epsilon) + \epsilon n \log|\mathcal{X}| \triangleq n\epsilon'.$$

Clearly, $\epsilon' \to 0$ as $\epsilon \to 0$ uniformly in n. On the other hand,

$$\begin{aligned}H(\boldsymbol{Y} \,|\, W) &= H(\boldsymbol{f}(W, \boldsymbol{S}) + \boldsymbol{Z} \,|\, W)\\ &\overset{\text{a)}}{=} \sum_w p(w) H(\boldsymbol{f}(w, \boldsymbol{S}) + \boldsymbol{Z})\\ &\ge \min_w H(\boldsymbol{f}(w, \boldsymbol{S}) + \boldsymbol{Z})\\ &\ge nH_{\min}^{(n)}\end{aligned} \tag{27}$$

where in a) we used the fact that W is statistically independent of $(\boldsymbol{Z}, \boldsymbol{S})$. Combining the above and the fact that $H(\boldsymbol{Y}) \le n\log|\mathcal{X}|$, we have

$$\begin{aligned}nR &= H(W)\\ &= H(\boldsymbol{Y}) - H(\boldsymbol{Y} \,|\, W) + H(W \,|\, \boldsymbol{Y})\\ &\le n\log|\mathcal{X}| - nH_{\min}^{(n)} + n\epsilon'\end{aligned} \tag{28}$$

which proves the lemma. □

Lemma 3 implies that the capacity of this channel is upper-bounded by

$$\log|\mathcal{X}| - \inf_n H_{\min}^{(n)},$$

a quantity which may be hard to compute in general. Nevertheless, we now show that for SWIP noise the instantaneous prediction scheme of Section II is still optimal.

Lemma 4: For SWIP noise we have for every n

$$H_{\min}^{(n)} = \frac{1}{n}H(Z_1 - t_{\min}(S_1), \cdots, Z_n - t_{\min}(S_n)) \tag{29}$$

where $t_{\min}$ achieves $H_{\min}^{(1)}$ in (26).

Proof: let $\tilde{\boldsymbol{Z}} = \boldsymbol{Z} - \boldsymbol{t}(\boldsymbol{S})$ for some function $\boldsymbol{t}: \mathcal{S}^n \to \mathcal{X}^n$. We have

$$\begin{aligned}H\left(\tilde{Z}_1^n\right) &= H(\tilde{Z}_1) + H(\tilde{Z}_2 \,|\, \tilde{Z}_1) + \cdots + H\left(\tilde{Z}_n \,\middle|\, \tilde{Z}_1^{n-1}\right) &(30)\\ &= H(\tilde{Z}_1) + H(\tilde{Z}_2 \,|\, \tilde{Z}_1) + \cdots + H\left(Z_n + t_n\left(S_1^n\right) \,|\, \tilde{Z}_1^{n-1}\right). &(31)\end{aligned}$$

Since $Z_n \leftrightarrow S_n \leftrightarrow (\tilde{Z}_1^{n-1}, S_1^{n-1})$ form a Markov chain, we have by Lemma 2 and the SWIP property

$$H\left(Z_n - t_n\left(S_1^n\right) \middle| \tilde{Z}_1^{n-1}\right) \ge H\left(Z_n - t_{\min}(S_n) \middle| \tilde{Z}_1^{n-1}\right)$$

i.e., assigning $t_n(s_1^n) = t_{\min}(s_n)$ can only reduce $H(\tilde{Z}_1^n)$. Since the ordering of the set $\{\tilde{Z}_1, \cdots, \tilde{Z}_n\}$ can be arbitrarily changed, the above argument holds for all the $\{t_i\}$ and we obtain

$$H(\boldsymbol{Z} - \boldsymbol{t}(\boldsymbol{S})) \ge H(Z_1 - t_{\min}(S_1), \cdots, Z_n - t_{\min}(S_n)). \tag{32}$$

Since (32) is true for any function $\boldsymbol{t}$, the optimum prediction function and the minimum noise entropy are

$$\boldsymbol{t}_{\min}(\boldsymbol{s}) = (t_{\min}(s_1), t_{\min}(s_2) \cdots t_{\min}(s_n))$$

and

$$H_{\min}^{(n)} = \frac{1}{n}H(Z_1 - t_{\min}(S_1), \cdots, Z_n - t_{\min}(S_n))$$

respectively, and the proof follows. □

As a consequence we have the following results.

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

$$t^n(s^n) = (t_{\min}(s_1), \cdots, t_{\min}(s_n))$$

where $t_{\min}$ achieves $H_{\min}^{(1)}$ in (26), is optimal for both causal and noncausal side information at the encoder. Thus

$$\begin{aligned}C_{\mathrm{SI@TR}}^{\mathrm{ncaus}} &= C_{\mathrm{SI@TR}}^{\mathrm{caus}} \\ &= \log|\mathcal{X}| - \lim_{n=\to\infty} \frac{1}{n} H(Z_1 - t_{\min}(S_1), \cdots, Z_n - t_{\min}(S_n))\end{aligned} \tag{33}$$

i.e., the optimum (causal or noncausal) encoder reduces to the instantaneous-prediction encoder of (18).

Note that unlike the general expression of Jelinek [9], for SWIP noise the optimization leading to the optimal strategy ($t_{\min}$) is done with respect to single-letter quantities.

Proof: Lemma 3 implies that

$$C_{\mathrm{SI@TR}}^{\mathrm{ncaus}} \le \log|\mathcal{X}| - \inf_n H_{\min}^{(n)}.$$

Furthermore, by Lemma 4

$$H_{\min}^{(n)} \ge (1/n) H(Z_1 - t_{\min}(S_1), \cdots, Z_n - t_{\min}(S_n))$$

which is further lower bounded by taking the limit as n goes to infinity (since $Z_i - t_{\min}(S_i)$, $i = 1, 2, \cdots$ is a stationary process). Thus the noncausal capacity is *upper*-bounded by (33). Achievability follows by noting that the effective noise channel, resulting from the instantaneous encoding scheme, is a stationary additive noise channel having (33) as its capacity; see, e.g., [1]. □

Corollary 1 (Memoryless Noncausal Case): For the memoryless additive noise channel (8), if the noise satisfies the SWIP property, then

$$C_{\mathrm{SI@TR}}^{\mathrm{ncaus}} = C_{\mathrm{SI@TR}}^{\mathrm{caus}} = \log|\mathcal{X}| - H(Z - t_{\min}(S)).$$

V. DISCUSSION

We identified the role of *instantaneous* noise prediction as the main ingredient in optimum transmission with side information over discrete additive-noise channels. This allows us to derive formulas for the capacity of memoryless channels with causal side information, and for SWIP noise, for the capacity with noncausal side information and for channels with memory.

In a future paper [5] we extend these concepts to rates below capacity and obtain expressions for the error exponent for additive noise channels with side information at the transmitter.

Our results do not extend to channels where the input is subject to some average input constraint. One reason for that is that under such a constraint each shift $t(s)$ of the input may have a different "cost, " and this changes the optimization of $t(s)$. Another reason can be seen in the case of a power-constrained Gaussian fading channel, with fading level known (say, causally) to the transmitter. Here, for a zero-mean noise our theory suggests that a fading independent code with zero shift is optimal. However, due to the average power constraint, allocating different power to code symbols at different fading levels improves the system performance [13].

APPENDIX
PROOF OF LEMMA 1

Suppose $\mathcal{S} = \{s_1, \cdots, s_k\}$. Define

$$g_i(s,u) = \begin{cases} g(s_i, u), & \text{if } s = s_i \\ 0, & \text{if } s \ne s_i \end{cases}$$

so that

$$g(s,u) = \sum_{i=1}^{k} g_i(s,u).$$

Let us introduce the auxiliary random variables $\{U_1, \cdots, U_k\}$, which are mutually independent and independent of (S, Z), and are distributed as

$$p_{u_i}(u) \triangleq p_{U\,|\,S}(u\,|\,s_i).$$

With these definitions we have

$$\begin{aligned}&H(Z + g(S,U)) \\ &\overset{\text{a)}}{=} H\left(Z + \sum_{i=1}^{k} g_i(S,U)\right) \\ &\overset{\text{b)}}{=} H\left(Z + \sum_{i=1}^{k} g_i(S,U_i)\right) \\ &\overset{\text{c)}}{\ge} H\left(Z + \sum_{i=1}^{k} g_i(S,U_i)\,\middle|\, U_1 \cdots U_k\right) \\ &\overset{\text{d)}}{\ge} \min_{u_1 \cdots u_k} H\left(Z + \sum_{i=1}^{k} g_i(S,u_i)\,\middle|\, U_1 = u_1 \cdots U_k = u_k\right) \\ &\overset{\text{e)}}{=} \min_{u_1 \cdots u_k} H\left(Z + \sum_{i=1}^{k} g_i(S,u_i)\right) \\ &\overset{\text{f)}}{\ge} \min_{t:\mathcal{S}\to\mathcal{X}} H(Z + t(S)) \\ &\overset{\text{g)}}{=} H(Z + t_{\min}(S))\end{aligned}$$

where b) follows from the Markov relation $U - S - Z$; c) since conditioning reduces the entropy; and e) since (Z, S) is independent of $(U_1, \cdots, U_k)$.

ACKNOWLEDGMENT

The authors wish to thank Shlomo Shamai for his valuable comments, and for pointing their attention to the noncausal side information case and the work by Gelfand, Pinsker, and Costa. They also wish to thank Meir Feder for bringing to their attention the work of Elias on minimum error entropy prediction.

REFERENCES

[1] F. Alajaji, "Feedback does not increase the capacity of discrete channels with additive noise," *IEEE Trans. Inform. Theory*, vol. 41, pp. 546–549, Mar. 1995.

[2] G. Caire and S. Shamai, "On the capacity of some channels with channel state information," *IEEE Trans. Inform. Theory*, vol. 45, pp. 2007–2019, Sept. 1999.

[3] M. H. M. Costa, "Writing on dirty paper," *IEEE Trans. Inform. Theory*, vol. IT-29, pp. 439–441, May 1983.

[4] P. Elias, "Predictive coding," *IRE Trans. Inform. Theory*, vol. IT-1, pp. 16–33, Mar. 1955.

[5] U. Erez and R. Zamir, "Error exponents of modulo additive noise channels with side information at the transmitter," *IEEE Trans. Inform. Theory*, May 2000, revised, submitted for publication.

[6] R. G. Gallager, *Information Theory and Reliable Communication*. New York: Wiley, 1968.

[7] S. I. Gelfand and M. S. Pinsker, "Coding for channel with random parameters," *Probl. Pered. Inform. (Probl. Inform. Transm.)*, vol. 9, no. 1, pp. 19–31, 1980.

[8] C. Heegard and A. El Gamal, "On the capacity of computer memory with defects," *IEEE Trans. Inform. Theory*, vol. IT-29, pp. 731–739, Sept. 1983.

[9] F. Jelinek, "Indecomposable channels with side information at the transmitter," *Inform. Contr.*, vol. 8, pp. 36–55, 1965.

[10] ——, "Determination of capacity achieving input probabilities for a class of finite state channels with side information," *Inform. Contr.*, vol. 9, pp. 101–129, 1966.

[11] R. J. McEliece and W. E. Stark, "Channels with block interference," *IEEE Trans. Inform. Theory*, vol. IT-30, pp. 44–53, Jan. 1984.

[12] M. Salehi, "Capacity and coding for memories with real-time noisy defect information at encoder and decoder," *Proc. Inst. Elect. Eng.*, vol. 139, pp. 113–117, Apr. 1992.

[13] S. Shamai, private communication, 1997.

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

[15] H. Viswanathan, "Capacity of Markov channels with receiver CSI and delayed feedback," *IEEE Trans. Inform. Theory*, vol. 45, pp. 761–771, Mar. 1999.

[16] J. Wolfowitz, *Coding Theorems of Information Theory*. New York: Springer-Verlag, 1964.

On the Training Distortion of Vector Quantizers

Tamás Linder, *Member, IEEE*

Abstract—**The in-training-set performance of a vector quantizer as a function of its training set size is investigated. For squared error distortion and independent training data, worst case type upper bounds are derived on the minimum training distortion achieved by an empirically optimal quantizer. These bounds show that the training distortion can underestimate the minimum distortion of a truly optimal quantizer by as much as a constant times $n^{-1/2}$, where n is the size of the training data. Earlier results provide lower bounds of the same order.**

Index Terms—**Empirical design, training distortion, vector quantization, worst case bounds.**

I. INTRODUCTION

Vector quantizer design is usually based on a collection of example vectors, called the training set or training data. In general, the objective of a design algorithm (such as the popular generalized Lloyd algorithm [1]) is to find an empirically optimal quantizer, that is, a quantizer of a given codebook size whose distortion in quantizing the training data is minimum. The underlying principle of empirical design is that good performance inside the training set will imply good performance on other data produced by the source if the training set size is sufficiently large to represent well the source statistics. But training vectors may be costly to obtain and the computational cost of design may become prohibitive for large training sets. Therefore, it is of interest to quantify how the performance of the designed vector quantizer improves as the size of the training set increases.

Assume that the quantizer dimension and the codebook size are fixed. For any quantizer Q_n trained on n vectors, let $D_n(Q_n)$ denote the training distortion of Q_n (its average distortion inside the training set) and let $D(Q_n)$ denote the test distortion of Q_n (its distortion in coding independent test data). Note that both $D_n(Q_n)$ and $D(Q_n)$ are functions of the training set and therefore are random quantities. The quantity $D(Q_n)$ is the "true" distortion of the designed quantizer; it is the performance figure one wants to be as close as possible to $D(Q^*)$, the distortion of a truly optimal quantizer Q^*. A design procedure is called consistent if the test distortion $D(Q_n)$ of the resulting quantizer Q_n converges (in some sense) to its lower bound $D(Q^*)$ as $n \to \infty$. Of particular interest are the empirically optimal quantizers Q_n^* minimizing the training distortion: $D_n(Q_n^*) = \min_{Q_n} D_n(Q_n)$. The consistency of empirically optimal quantizers was first investigated by Pollard [2], [3] for the case of mean-squared quantizer distortion. His results show, among other things, that for a stationary and ergodic training sequence, the test distortion $D(Q_n^*)$ of an empirically optimal quantizer converges to $D(Q^*)$ with probability one as $n \to \infty$.

Pollard's results imply that the performance of empirically optimal quantizers will approach the optimum performance as the training set size increases without bound. On the other hand, to determine the training set size sufficient for achieving a preassigned level of performance, one needs to study the dependence of $D(Q_n^*)$ on finite n. Assume that the training set consists of n independent sample vectors drawn from the source distribution and let $E[D(Q_n^*)]$ denote the expected value (taken over the training sequence) of the mean-squared test distortion $D(Q_n^*)$. In [4] it was shown that for all source distributions supported by a given bounded region, the test distortion of the empirically optimal quantizer satisfies $E[D(Q_n^*)] - D(Q^*) \le cn^{-1/2}$ for some positive constant c. This upper bound was shown to have the right order in a minimax sense in [5], where it was demonstrated that for any quantizer design method, there exist "bad" source distributions for which the test distortion of the resulting quantizer Q_n is lower-bounded as $E[D(Q_n)] - D(Q^*) \ge c_1 n^{-1/2}$ for another positive constant c_1. The sample behavior of $D(Q_n^*) - D(Q^*)$ for a class of smooth source densities was studied by Chou [6], and upper bounds on $E[D(Q_n^*)] - D(Q^*)$ for dependent (mixing) training data were developed by Zeevi [7]. The dependence of the test distortion on the training set size was also empirically investigated by Cosman *et al.* [8] and Cohn *et al.* [9] in the context of image coding.

In this correspondence, the focus of attention is the less studied training distortion $D_n(Q_n^*)$. Since the value of $D_n(Q_n^*)$ is obtained as a by-product of the design procedure without requiring additional test data, it can be considered an inexpensive estimate of $D(Q_n^*)$ or $D(Q^*)$. For an empirically optimal quantizer minimizing $D_n(Q_n)$, one always has

$$E[D_n(Q_n^*)] \le D(Q^*) \le E[D(Q_n^*)].$$

The exact relationship between the training, test, and optimal distortions is only known in the special case of quantizers with codebook size $k = 1$. In this case, it is easy to see that the single unique codepoint of the empirically optimal quantizer is the arithmetic average of the n training vectors, and therefore,

$$E[D_n(Q_n^*)] = D(Q^*)\left(1 - \frac{1}{n}\right)$$

and

$$E[D(Q_n^*)] = D(Q^*)\left(1 + \frac{1}{n}\right)$$

for all source distributions with finite second moment.

The problem becomes nontrivial when quantizers with more than one codepoint are considered, and in general little is known about the size of the difference $D(Q^*) - E[D_n(Q_n^*)]$. In this respect, Abaya and Wise [10] proved that under general conditions the expected training distortion is a consistent estimate of the optimal distortion in the sense that $D(Q^*) - E[D_n(Q_n^*)] \to 0$ as $n \to \infty$. The size of the bias of $D_n(Q_n^*)$ in estimating $D(Q^*)$ was first investigated in a recent work by Kim and Bell [11] who showed that for squared error distortion

$$E[D_n(Q_n^*)] \le D(Q^*)\left(1 - \frac{1}{n}\right) \qquad (1)$$

Manuscript received July 14, 1999; revised February 1, 2000. This work was supported in part by Queen's University, Kingston, Ont., Canada, and by the Natural Sciences and Engineering Research Council (NSERC) of Canada.

The author is with the Department of Mathematics and Statistics, Queen's University, Kingston, Ont., Canada K7L 3N6 (e-mail: linder@mast.queensu.ca).

Communicated by P. A. Chou, Associate Editor for Source Coding.
Publisher Item Identifier S 0018-9448(00)04282-6.

0018–9448/00$10.00 © 2000 IEEE