$$+ (2^{1-\alpha} - 1) H_M^\alpha(P_1, P_2, \ldots, P_M) H_N^\alpha(Q_1, Q_2, \ldots, Q_N)$$

$[(P_1, P_2, \ldots, P_M) \in \mathcal{P}([M]), (Q_1, Q_2, \ldots, Q_N) \in \mathcal{P}([N]); M = 2, 3, \ldots; N = 2, 3, \ldots]$.

strong additive of degree α:

$$H_{MN}^\alpha(P_1Q_{11}, P_1Q_{12}, \ldots, P_1Q_{1N}, P_2Q_{21}, P_2Q_{22}, \ldots, P_2Q_{2N}, \ldots, P_MQ_{M1}, P_MQ_{M2}, \ldots, P_MQ_{MN}) = H_M^\alpha(P_1, P_2, \ldots, P_M) + \sum_{j=1}^{M} P_j^\alpha H_N^\alpha(Q_{j1}, Q_{j2}, \ldots, Q_{jN})$$

$[(P_1, P_2, \ldots, P_M) \in \mathcal{P}([M]), (Q_{j1}, Q_{j2}, \ldots, Q_{jN}) \in \mathcal{P}([N]); j = 1, 2, \ldots, M; M = 2, 3, \ldots; N = 2, 3, \ldots]$.

recursive of degree α:

$$H_N^\alpha(P_1, P_2, \ldots, P_N) = H_{N-1}^\alpha(P_1 + P_2, P_3, \ldots, P_N) + (P_1 + P_2)^\alpha H_2^\alpha\left(\frac{P_1}{P_1 + P_2}, \frac{P_2}{P_1 + P_2}\right)$$

$[(P_1, P_2, \ldots, P_N) \in \mathcal{P}([N]), N = 3, 4, \ldots$ with $P_1 + P_2 > 0]$.

(In consequence, entropies of type α also have the branching property.)

It is clear now that for binary alphabet the ID-entropy is exactly the entropy of type $\alpha = 2$.

However, prior to [13] there are hardly any applications or operational justifications of the entropy of type α.

Moreover, the q-ary case did not exist at all and therefore the name ID-entropy is well justified.

We feel that it must be said that in many papers (with several coauthors) Tsallis at least developed ideas to promote non-standard-equilibrium theory in Statistical Physics using generalized entropies S_α and generalized concepts of inner energy.

Our attention has been drawn also to the papers [5], [11], [12] with possibilities of connections to our work.

Recently, clear-cut progress was made by C. Heup in his forthcoming thesis with a generalization of ID-entropy motivated by L-identification.

REFERENCES

[1] S. Abe, "Axioms and uniqueness theorem for Tsallis entropy," *Phys. Lett.*, vol. A 271, no. 1–2, pp. 74–79, 2000.

[2] J. Aczél and Z. Daróczy, *On Measures of Information and their Characterizations, Mathematics in Science and Engineering*. New York/London: Academic, 1975, vol. 115.

[3] R. Ahlswede, "General theory of information transfer," *Discr. Appl. Math.*, updated Special issue "General Theory of Information Transfer and Combinatorics", to be published.

[4] R. Ahlswede, Identification Entropy, General Theory of Information Transfer and Combinatorics, Oct. 1, 2002–Aug. 31, 2004, Report on a Research Project at the ZIF (Center of Interdisciplinary Studies) in Bielefeld, edited by R. Ahlswede with the assistance of L. Bäumer and N. Cai, *Lecture Notes in Computer Sceince*, no. 4123, to be published..

[5] L. L. Campbell, "A coding theorem and Rényi's entropy," *Inf. Contr.*, vol. 8, pp. 423–429, 1965.

[6] Z. Daróczy, "Generalized information functions," *Inf. Contr.*, vol. 16, pp. 36–51, 1970.

[7] J. Havrda and F. Charvát, "Quantification method of classification processes, concept of structural α-entropy," *Kybernetika (Prague)*, vol. 3, pp. 30–35, 1967.

[8] A. Rényi, "On measures of entropy and information," in *Proc. 4th Berkeley Symp. Mathematical Statistics and Probability*, vol. I, pp. 547–561, Berkeley, CA: Univ. California Press, 1961.

[9] M. P. Schützenberger, *Contribution aux applications statistiques de la thorie de l'information*. Paris, France: Publ. Inst. Statist. Univ. Paris 3, 1954, vol. 3, 1–2, pp. 3–117.

[10] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, pp. 379–423, 623–656, 1948.

[11] B. D. Sharma and H. C. Gupta, "Entropy as an optimal measure, Information theory," in *Proc. Int. CNRS Colloq., Cachan, 1977)*, Paris, French, 1978, pp. 151–159, Colloq. Internat. CNRS, 276, CNRS.

[12] F. Topsøe, "Game-theoretical equilibrium, maximum entropy and minimum information discrimination, Maximum entropy and Bayesian methods," *Fund. Theor. Phys.*, vol. 53, pp. 15–23, 1992, Published by Kluwer Acad., Paris, France, 1993.

[13] C. Tsallis, "Possible generalization of Boltzmann-Gibbs statistics," *J. Statist. Phys.*, vol. 52, no. 1–2, pp. 479–487, 1988.

[14] C. Tsallis, R. S. Mendes, and A. R. Plastino, "The role of constraints within generalized nonextensive statistics," *Phys. A*, vol. 261, pp. 534–554, 1998.

Coding for the Feedback Gel'fand–Pinsker Channel and the Feedforward Wyner–Ziv Source

Neri Merhav, *Fellow, IEEE*, and Tsachy Weissman, *Member, IEEE*

***Abstract*—We consider both channel coding and source coding, with perfect past feedback/feedforward, in the presence of side information. It is first observed that feedback does not increase the capacity of the Gel'fand–Pinsker channel, nor does feedforward improve the achievable rate–distortion performance in the Wyner–Ziv problem. We then focus on the Gaussian case showing that, as in the absence of side information, feedback/feedforward allows to efficiently attain the respective performance limits. In particular, we derive schemes via variations on that of Schalkwijk and Kailath. These variants, which are as simple as their origin and require no binning, are shown to achieve, respectively, the capacity of Costa's channel, and the Wyner–Ziv rate distortion function. Finally, we consider the finite-alphabet setting and derive schemes for both the channel and the source coding problems that attain the fundamental limits, using variations on schemes of Ahlswede and Ooi and Wornell, and of Martinian and Wornell, respectively.**

***Index Terms*—Dirty paper, feedback, feedforward, Schalkwijk–Kailath scheme, side information, source–channel duality.**

I. INTRODUCTION

That feedback does not increase the capacity of a memoryless channel, yet can dramatically simplify the schemes for achieving it, is a well known fact (cf. [6] and the literature survey therein). More recently, an analogous phenomenon was shown to hold for the dual problem of lossy source coding with perfect past feedback, also known as "feedforward," at the decoder [12], [7], [5], a problem arising in contexts as diverse as prediction theory, remote sensing, and control.

In this work, we revisit these problems to accommodate the presence of side information. As is the case for problems without feedback/feed-

Manuscript received September 1, 2005. The material in this correspondence was presented in part at the IEEE International Symposium on Information theory, Adelaide, Australia, September 2005.

N. Merhav is with the Department of Electrical Engineering, Technion-Israel Institute of Technology, Technion City, Haifa 32000, Israel (e-mail: merhav@ee.technion.ac.il).

T. Weissman is with the Department of Electrical Engineering, Stanford University, Stanford, CA 94305-9510 USA (e-mail: tsachy@stanford.edu).

Communicated by K. Kobayashi, Associate Editor for Shannon Theory.

Digital Object Identifier 10.1109/TIT.2006.880040

forward, the only scenarios with fundamental limits, and achieving schemes, that are not directly implied from those known for the absence of side information are, respectively, the presence of side information only at the encoder, and only at the decoder, for the channel coding and the source coding problems.

Our first observation in this context is that the fact that feedback/feedforward does not improve the fundamental performance limits carries over to these cases where side information is present. To see this, consider first channel coding for the Gel'fand–Pinsker channel [4] with feedback, by which we mean the following: The channel state information S^n is available to the sender, and the memoryless channel has transition probability $p(y|x,s)$ that depends on the input X and the state S. S_i are assumed independent and identically distributed (i.i.d.) $\sim p(s)$. For a message index $W \in \{1,2,\ldots,2^{nR}\}$, the ith channel input is of the form $X_i(W, S^n, Y^{i-1})$, i.e., it is allowed to depend on the past channel output symbols. Decoding is, as usual, based on the channel output Y^n.

Observation 1: Feedback does not increase the capacity of the Gel'fand–Pinsker channel.

Proof: One need merely observe that the original converse proof of Gel'fand and Pinsker [4] is general enough so as to include feedback. In other words, $U_i - (X_i, S_i) - Y_i$, where $U_i = (W, S_{i+1}^n, Y^{i-1})$, continues to be a Markov chain even in the presence of feedback. To see this, note that $P(y_i|w, s^n, y^{i-1}) = p(y_i|x_i, s_i)$ so $(W, S^n, Y^{i-1}) - (X_i, S_i) - Y_i$ is a Markov chain and, therefore, since U_i is a deterministic function of (W, S^n, Y^{i-1}), so is $U_i - (X_i, S_i) - Y_i$. Q.E.D.

Though our interest in this work, and the schemes we develop, are for the case of noncausal state information (SI), we mention in passing that a similar conclusion applies also for the Shannon channel with causal SI, where the ith channel input is of the form $X_i(W, S^i, Y^{i-1})$. The independence between U_i and S_i in the causal case is readily verified to persevere in the presence of feedback, implying: Feedback does not increase the capacity of the Shannon channel (with causal SI).

Moving to the source coding analogue, consider the problem of Wyner–Ziv source coding [14] with feedforward: The source and side information are generated as independent drawings of the pair (X_i, Y_i). Encoding, as in the original problem, is done by mapping the sequence X^n into $T \in \{1,2,\ldots,2^{nR}\}$. The ith reconstruction this time is of the form $\hat{X}_i(T, Y^n, X^{i-1})$, i.e., allowed to depend also on the past, nonquantized, past source symbols. This setting is the extension of the source coding with feedforward problem [12], [7], [5] to the case of side information at the decoder.

Observation 2: Feedforward does not improve the rate distortion tradeoff in the Wyner–Ziv problem.

Proof: Here too, the original converse proof carries over essentially unchanged. Specifically, in the notation of [3, Sec. 14.9], we only need to add X^{i-1} to W_i, resulting in $W_i = (T, Y^{i-1}, Y_{i+1}^n, X^{i-1})$. The converse proof of [3, Theorem 14.9.1] carries over verbatim (erasing line (14.298) therein), since $W_i - X_i - Y_i$ continues to form a Markov chain under this modified W_i. Q.E.D

Given Observations 1 and 2, it is natural to ask whether, similarly as in the absence of side information, feedback/feedforward can lead to simple schemes for attaining the fundamental limits. For the Gaussian case, we answer this question in the affirmative in the next section. More specifically, we present efficient schemes that exploit feedback/feedforward to achieve the capacity of Costa's channel [2], and the Wyner–Ziv function for a source which is a Gaussian-noise-corrupted version of the side information. Our schemes, which are variations on those of Schalkwijk and Kailath [9], [8], are as efficient as their origin and, in particular, do not require binning. In Section III, we consider the finite-alphabet setting and derive a scheme for the Gel'fand–Pinsker channel with feedback, building on the ideas of [1], [6] . We also derive a scheme for the dual problem of Wyner–Ziv coding with feedforward, by extending the approach of [5]. Our schemes for the finite-alphabet setting rely on Slepian–Wolf coding [11], and thus we make no claim at this point regarding the efficiency with which they can be implemented (in comparison to the efficiency of practical schemes for the Gel'fand–Pinsker channel and the Wyner–Ziv problem in the absence of feedback/feedforward). They are, however, conceptually simple and suggest another view on the information-theoretic formulas of the Gel'fand–Pinsker capacity and the Wyner–Ziv rate–distortion function. They also shed light on yet another aspect of the duality between source coding and channel coding with side information.

II. Variations on the Schalkwijk–Kailath Scheme

A. Writing on Dirty Paper With On-Line Proofreading: Costa's Channel With Feedback

Consider the channel $Y_i = X_i + S_i + Z_i$, where $\{S_i\}$ is an interference signal (with $ES_i = 0$ and $\sigma_S^2 = ES_i^2 < \infty$) known to the encoder, and $\{Z_i\}$ is zero-mean, i.i.d. Gaussian noise with variance σ_Z^2. Let the transmission power be limited to P. We now describe a modified version of the scheme of [8] for coding with feedback, which achieves the capacity $C = \frac{1}{2}\log(1 + P/\sigma_Z^2)$. Moreover, for every $R < C$, the error probability is identical to that of the original scheme, as if S_i were identically zero, namely, it decays double-exponentially rapidly with $C - R$.

Initialization: Define $\alpha = \sqrt{1 + P/\sigma_Z^2}$ and $g = \sqrt{P/\sigma_Z^2}$. Given a message $m = 0, 1, \ldots, M-1, M = 2^{nR}$, let $\theta = (m + 1/2)/M$. Given $S^n = (S_1, \ldots, S_n)$ define $\psi_2 = S_1/\alpha$, and for $i = 2, 3, \ldots, n$, compute recursively

$$\psi_{i+1} = \psi_i + \left(1 - \frac{1}{\alpha^2}\right)\frac{S_i}{\alpha^{i-1}g}.$$

Finally, let $\theta' = \theta + \psi_{n+1}$.

Recursion: For $i = 1$, set $X_{1,1} = 0.5$ and transmit $\alpha(X_{1,1} - \theta')$. At the receiver, compute $X_{2,1} = X_{1,2} = X_{1,1} - \frac{Y_1}{\alpha}$ and send $X_{2,1}$ back to the transmitter. For $i = 2, 3, \ldots, n$, transmit $\alpha^{i-1}g(X_{i,1} - \theta' + \psi_i)$. At the receiver, compute $X_{i,2} = X_{i,1} - \frac{Y_i}{\alpha^{i-1}g}$, then update

$$X_{(i+1),1} = \frac{1}{\alpha^2}X_{i,1} + \left(1 - \frac{1}{\alpha^2}\right)X_{i,2}$$

and (for $i < n$) send $X_{(i+1),1}$ back to the transmitter.

Finally, decode m by quantizing $X_{(n+1),1}$ to its message interval.

Analysis: First, note that

$$\begin{aligned} X_{2,1} &= X_{1,1} - \frac{\alpha(X_{1,1} - \theta') + S_1 + Z_1}{\alpha} \\ &= \theta' - \frac{S_1}{\alpha} - \frac{Z_1}{\alpha} \\ &= \theta' - \psi_2 - \frac{Z_1}{\alpha}. \end{aligned} \tag{1}$$

We now argue that for all $i \geq 2, X_{i,1} = \theta' - \psi_i - \phi_i$, where $\{\psi_i\}$ are defined as above, and $\{\phi_i\}$ are defined by $\phi_2 = Z_1/\alpha$ and by the recursion

$$\phi_{i+1} = \frac{1}{\alpha^2}\phi_i + \left(1 - \frac{1}{\alpha^2}\right)\frac{Z_i}{\alpha^{i-1}g}, \qquad i = 2, 3, \ldots, n.$$

We prove this by induction: For $i = 2$, this has been shown already in (1). Assuming now that the hypothesis is true for a given $i \geq 2$, then

$$\begin{aligned} X_{(i+1),1} &= \frac{1}{\alpha^2} X_{i,1} + \left(1 - \frac{1}{\alpha^2}\right) X_{i,2} \\ &= \frac{1}{\alpha^2}(\theta' - \psi_i - \phi_i) + \left(1 - \frac{1}{\alpha^2}\right) \\ &\quad \times \left[X_{i,1} - \frac{\alpha^{i-1} g(X_{i,1} - \theta' + \psi_i) + S_i + Z_i}{\alpha^{i-1} g}\right] \\ &= \frac{1}{\alpha^2}(\theta' - \psi_i - \phi_i) + \left(1 - \frac{1}{\alpha^2}\right)\left[\theta' - \psi_i - \frac{S_i + Z_i}{\alpha^{i-1} g}\right] \\ &= \theta' - \left[\psi_i + \left(1 - \frac{1}{\alpha^2}\right)\frac{S_i}{\alpha^{i-1} g}\right] \\ &\quad - \left[\frac{1}{\alpha^2}\phi_i + \left(1 - \frac{1}{\alpha^2}\right)\frac{Z_i}{\alpha^{i-1} g}\right] \\ &= \theta' - \psi_{i+1} - \phi_{i+1} \end{aligned} \tag{2}$$

confirming the induction hypothesis for $i + 1$. Thus, for $i = n + 1$, we get

$$X_{(n+1),1} = \theta' - \psi_{n+1} - \phi_{n+1} = \theta - \phi_{n+1}. \tag{3}$$

But ϕ_{n+1} is exactly the estimation error variable in [8], whose variance has been shown to be σ_Z^2/α^{2n}. Thus, the decision made by this scheme is identical to that of Schalkwijk's scheme (with $S^n = 0$) for every realization of the noise sequence. Obviously, the error performance is then the same too.

As for the transmission power, we will distinguish again between $i = 1$ and $i \geq 2$. For $i = 1$, the transmission power is approximately $\alpha^2(1/12 + \mathrm{Var}\{\psi_{n+1}\})$, where $1/12$ approximates the variance of θ as one corresponding to the uniform distribution in $[0, 1]$, and $\mathrm{Var}\{\psi_{n+1}\}$ is bounded independently of n since ψ_{n+1} is a linear combination of $\{S_i\}$ with coefficients that decay exponentially with i. As for $i \geq 2$, the transmission power is

$$\begin{aligned} \alpha^{2(i-1)} g^2 E(X_{i,1} - \theta' + \psi_i)^2 &= \alpha^{2(i-1)} g^2 E\phi_i^2 \\ &= \alpha^{2(i-1)} g^2 \frac{\sigma_Z^2}{\alpha^{2(i-1)}} = \sigma_Z^2 g^2 = P, \end{aligned} \tag{4}$$

where the second equality has been proved in [8] (and can also easily be seen by induction, using the recursive definition of $\{\phi_i\}$). Thus, except for $i = 1$, the transmission power is P at all times, which means that for large n the total average power tends to P.

At the point, a few comments are in order.

1. We have seen that in the presence of feedback, it is possible to achieve capacity with a simple scheme, without binning.
2. While in the absence of feedback [2], the idea is not to "fight" the interference by trying to pre-cancel it but rather to harness it to our own benefit, here the pre-canceling approach seems to be fruitful. This is manifested both at the transmitter, where the contribution of $\{S_i\}$ to the estimation error to be transmitted is canceled in order to save power, and in the definition of θ', which shifts θ by an amount (ψ_{n+1}) which pre-cancels the contribution of $\{S_i\}$ to the error of the final estimator.
3. As mentioned earlier, operatively, this scheme gives exactly the same estimation and decoding as in [8] for every realization of the noise process, and as if S^n were nonexistent $(S^n = 0)$.
4. Similarly to the non-feedback case, the probability law of $\{S_n\}$ is immaterial. The only requirement is that $\sigma_S^2 < \infty$ to assure that the expected power used at time $i = 1$ is finite.
5. Note that the noncausal dependence of the transmission on S^n is only via one number ψ_{n+1}.

B. A Scheme for Wyner–Ziv Coding With Feedforward

Consider first rate distortion coding with feedforward in the absence of side information [12]. Let $\{X_i\}_{i=1}^l$ be i.i.d. $\mathcal{N}(0, \sigma^2)$ and, for a given positive real β, let

$$Y = -\sum_{k=2}^{l} \sqrt{\beta^2 - 1}\,\beta^{-(k+1)} X_k - \beta^{-1} X_1. \tag{5}$$

Let $\hat{Y}$ be the quantized version of Y using a uniform scalar quantizer on the interval $[-\Delta/2, \Delta/2]$ with M levels (truncating values outside the interval). The encoder describes $\hat{Y}$ to the decoder by giving index $I(\hat{Y})$ of the quantization cell. The decoder reconstructs as follows: $\hat{X}_1 = -\beta\hat{Y}$, $\hat{X}_2 = \sqrt{\beta^2 - 1}(\hat{X}_1 - X_1)$, $\hat{X}_i = \beta\hat{X}_{i-1} - (\beta^2 - 1)\beta^{-1} X_{i-1}$ for $i = 3, \ldots, l$. It was shown in [7] that, for $l \geq 1$

$$\frac{1}{l}\sum_{i=1}^{l} E(X_i - \hat{X}_i)^2 = \frac{E(Y - \hat{Y})^2 \beta^{2l}}{l} + \frac{\sigma^2(l\beta^2 - \beta^2)}{l\beta^4}. \tag{6}$$

To see how this scheme attains the rate distortion function, fix the rate R (so $M = 2^{Rl}$) and a small $\varepsilon > 0$ throughout. Take $\beta = 2^{R-2\varepsilon}$ and $\Delta = 2^{l\varepsilon}$. We note the following.

1. $\sum_{k=2}^{\infty}(\beta^2 - 1)\beta^{-2(k+1)} < \infty$ so the variance of Y is bounded (does not exceed a fixed value) regardless of l.
2. $\Pr\{Y \notin [-\Delta/2, \Delta/2]\}$ is diminishing with l (in fact, double exponentially rapidly since Y is Gaussian with bounded variance and is Δ exponentially growing with l).
3. In $[-\Delta/2, \Delta/2]$, we are performing uniform quantization with resolution $\Delta/M = 2^{-(R-\varepsilon)l}$.
4. The two previous items imply that $E(Y - \hat{Y})^2 \leq c(\Delta/M)^2 = c2^{-2(R-\varepsilon)l}$ for an l-independent constant c (in fact, a high-resolution quantization argument will give the more refined $E(Y - \hat{Y})^2 \sim \frac{1}{12}2^{-2(R-\varepsilon)l}$).
5. Substituting into (6), we get, as l grows large, that the first term on the right-hand side diminishes, while the second one converges to $\frac{\sigma^2}{\beta^2} = \sigma^2 2^{-2(R-2\varepsilon)}$, which is the distortion-rate function (up to the small ε factor).

Performance analysis for our scheme below will rely also on the following.

Claim 1: The scheme described is robust in the sense that if the decoder receives any index $\tilde{I}$ such that $\log|I(\hat{Y}) - \tilde{I}| = o(l)$, then the distortion converges, as for the original scheme, to $\sigma^2 2^{-2(R-2\varepsilon)}$.

Proof: The distance between the centers of two adjacent quantization cells is $2^{-(R-\varepsilon)l}$, so, letting $\tilde{Y}$ denote the value of $\hat{Y}$ that the decoder assumes based on $\tilde{I}$, $|\hat{Y} - \tilde{Y}| \leq 2^{-(R-\varepsilon+o(1))l}$. The error in reconstruction due to this discrepancy can increase from one component to the next by a factor of $\beta = 2^{R-2\varepsilon}$, so the overall distance between the reconstruction based on $\tilde{I}$ and that based on I is diminishing (this is why $\beta = 2^{R-2\varepsilon}$ rather than $\beta = 2^{R-\varepsilon}$ was taken). Q.E.D.

Consider now the Wyner–Ziv problem with perfect feedforword on the past source symbols at the decoder. Assume that

1. $\{Y_i\}$ is an arbitrarily distributed side-information signal available only at the decoder;
2. $\{X_i\}$, the source signal, is given by $X_i = Y_i + N_i$, where $\{N_i\}$ is i.i.d. $\mathcal{N}(0, \sigma^2)$, independent of $\{Y_i\}$.

Consider next the following scheme for this setting.

- Encoder: operate *exactly* as encoder associated with (6).
- Decoder:
  1. Add $\sum_{k=2}^{l} \sqrt{\beta^2 - 1}\,\beta^{-(k+1)} Y_k + \beta^{-1} Y_1$ to received $\hat{Y}$.
  2. Input the result into the decoder using $\{N_i\}$ as the feedforward sequence (which is possible since at time i X_{i-1} is revealed, and Y_{i-1} is of course known).

3. Let the reconstruction be given by $\hat{X}_i = Y_i + \hat{N}_i$, where $\hat{N}_i$ is output of the decoder from the previous stage.

Claim 2: As $l \to \infty$, the distortion of the scheme described converges to $\sigma^2 2^{-2(R-2\varepsilon)}$.

Proof: Since

$$\begin{aligned} Y &= -\sum_{k=2}^{l} \sqrt{\beta^2-1}\beta^{-(k+1)} X_k - \beta^{-1} X_1 \\ &= -\sum_{k=2}^{l} \sqrt{\beta^2-1}\beta^{-(k+1)} Y_k \\ &\quad - \beta^{-1} Y_1 - \sum_{k=2}^{l} \sqrt{\beta^2-1}\beta^{-(k+1)} N_k - \beta^{-1} N_1 \end{aligned}$$

assuming $Y \in [-\Delta/2, \Delta/2]$, the quantization resolution implies $|Y - \hat{Y}| \le \Delta/M = 2^{-(R-\varepsilon)l}$, so the input (index) given to the decoder in the second stage is within 1 from what it would have received had encoding been performed (with scheme in (6) directly on the $\{N_i\}$ sequence. Claim 1 implies then that the distortion between $\{N_i\}$ and $\{\hat{N}_i\}$, hence also between $\{X_i\}$ and $\{\hat{X}_i\}$, is essentially $\sigma^2 2^{-2(R-2\varepsilon)}$. It only remains to argue that our assumption $Y \in [-\Delta/2, \Delta/2]$ was justified. To this end, observe that

$$\mathrm{Var}\{Y\} \le \mathrm{Var}\left\{\sum_{k=2}^{l} \sqrt{\beta^2-1}\beta^{-(k+1)} Y_k + \beta^{-1} Y_1\right\} \tag{7}$$

$$+ \sigma^2 \left[\sum_{k=2}^{\infty} (\beta^2-1)\beta^{-2(k+1)} + \beta^{-2}\right], \tag{8}$$

so as long as $\sum_{k=2}^{l} \sqrt{\beta^2-1}\beta^{-(k+1)} Y_k + \beta^{-1} Y_1$ has expectation and variance growing subexponentially with l, which is the case for all but the wildest processes, since $\Delta = 2^{l\varepsilon}$, $\Pr\{Y \in [-\Delta/2, \Delta/2]\}$ is overwhelmingly small. Q.E.D.

Comments:

1. The scheme is as simple as the channel coding one, with no binning required.
2. This scheme achieves the conditional rate–distortion function for the case where the side information is available at both encoder and decoder, for an arbitrarily distributed side information process $\{Y_i\}$.
3. Observation 2, combined with the previous item, implies that for the regular Wyner–Ziv problem, in the case where the pairs (X_i, Y_i) are i.i.d., with Y_i arbitrarily distributed and $X_i = Y_i + N_i$ for N_i Gaussian and independent of Y_i, there is no loss due to the absence of side information at the encoder. This fact can be deduced also directly from the single-letter expression. Indeed, the argument used in [13, Sec. 3] to show that the Wyner–Ziv function coincides with the conditional rate distortion function when (X_i, Y_i) are jointly Gaussian is readily seen to carry over to this more general case.
4. The results of [12] can be shown to imply, for an arbitrarily distributed SI process $\{Y_i\}$, and source given by $X_i = Y_i + N_i$, for i.i.d. (but arbitrarily distributed) process $\{N_i\}$, that feedforward does not help for source coding with SI on both sides. Combined with the second item, this implies that the Wyner–Ziv performance in the presence of feedforward for an arbitrarily distributed SI process and source given by $X_i = Y_i + N_i$, for N_i i.i.d. Gaussian, coincides with that for SI at both sides. Furthermore, we have just shown a simple scheme attaining optimum performance for this case which is no less simple than had the SI been available at the encoder as well. Thus, not only is there no loss for not knowing the SI at the encoder in terms of the fundamental limit, there is also no loss in the simplicity of the scheme attaining it.
5. Noncausal dependence of decoding on the SI in the above scheme is only once, in the first step, for computing $\sum_{k=2}^{l} \sqrt{\beta^2-1}\beta^{-(k+1)} Y_k + \beta^{-1} Y_1$. The reconstruction in the remaining steps uses the SI causally.

III. Finite Alphabets

A. A Scheme for the Gel'Fand–Pinsker Channel With Feedback

Consider the finite-alphabet setting of the Gel'fand–Pinsker channel, as described in the Introduction. Let S, U, X, Y have a capacity-achieving distribution, namely, a distribution achieving $\max_{p(u|s),f}[I(U;Y) - I(U;S)]$, where $X = f(U,S)$. Consider the following scheme of coding with feedback for the Gel'fand–Pinsker channel,[1] building on the ideas of [1], [6].

- *Transmitter:* Maps the N message bits into the sequence U^{n_1}, $n_1 = N/H(U|S)$, where U^{n_1} is the output of the decoder corresponding to an optimal Slepian–Wolf encoder of U^{n_1} for side information S^{n_1}, when receiving the N message bits as input from the encoder and observing the side information S^{n_1}.
- Sends X^{n_1} through the channel, where $X_i = f(U_i, S_i), 1 \le i \le n_1$.
- *Channel:* Corrupts X^{n_1} according to $p(y|s,x)$.
- *Receiver:* Feeds channel output Y^{n_1} back to the transmitter.
- *Transmitter:* Using Y^{n_1}, compresses U^{n_1} into $n_1 H(U|Y)$ new data bits.
- Maps these bits into the sequence $U_{n_1+1}^{n_1+n_2}$, for $n_2 = n_1 H(U|Y)/H(U|S)$, by letting $U_{n_1+1}^{n_1+n_2}$ be the output of the decoder corresponding to an optimal Slepian–Wolf encoder of U for side information S (for n_2-tuples), when receiving the $n_1 H(U|Y)$ new data bits as input from the encoder and observing the side information $S_{n_1+1}^{n_1+n_2}$.
- Sends $X_{n_1+1}^{n_1+n_2}$ through the channel, where $X_i = f(U_i, S_i)$, $n_1 + 1 \le i \le n_1 + n_2$.
- *Channel:* Corrupts $X_{n_1+1}^{n_1+n_2}$ according to $p(y|s,x)$.
- *Receiver:* Feeds channel output $Y_{n_1+1}^{n_1+n_2}$ back to the transmitter.
- *Transmitter:* Using $Y_{n_1+1}^{n_1+n_2}$, compresses $U_{n_1+1}^{n_1+n_2}$ into $n_2 H(U|Y)$ new data bits,

and so on. After k iterations of this process, letting $l_k = \sum_{i=1}^{k} n_i$, use a simplistic termination code for conveying the n_k-tuple $U_{l_{k-1}+1}^{l_k}$ to the decoder, allowed to be based also on $Y_{l_{k-1}+1}^{l_k}$ that will be available from the feedback. Thus, in effect, this termination code needs to communicate $\approx n_k H(U|Y) = N[H(U|Y)/H(U|S)]^k$ additional information bits.

Decoding: Let $\hat{U}_{l_{k-1}+1}^{l_k}$ denote the decoder's estimated version of $U_{l_{k-1}+1}^{l_k}$, and let $\boldsymbol{b_k}$ be the binary $n_k H(U|S)$-tuple obtained by taking the output of the Slepian–Wolf encoder (used at the kth stage of the encoding) when this n_k-tuple is used as its input. Let $\hat{U}_{l_{k-2}+1}^{l_{k-1}}$ be the conditional entropy decoding of an n_{k-1}-tuple of the source U given the corresponding n_{k-1}-tuple of Y as side information, for the Y sequence $Y_{l_{k-2}+1}^{l_{k-1}}$ and the binary encoding $\boldsymbol{b}_k$. Now feed the n_{k-1}-tuple $\hat{U}_{l_{k-2}+1}^{l_{k-1}}$ into the Slepian–Wolf encoder used at the $k-1$th stage of the encoding, and let $\boldsymbol{b_{k-1}}$ be the binary $n_{k-1} H(U|S)$-tuple obtained

[1] Throughout, we ignore integer constraints, writing, e.g., $N/H(U|S)$ rather than $\lceil N/H(U|S) \rceil$.

at its output. Continue this process for k iterations, until obtaining the binary N-tuple $\boldsymbol{b}_1$, letting that be the decoded message bits.

Analysis: The overall number of channel uses is

$$\begin{aligned} l_k + L &= \frac{N}{H(U|S)} \sum_{i=1}^{k} [H(U|Y)/H(U|S)]^{i-1} + L \\ &= \frac{N}{H(U|S)} \frac{1 - [H(U|Y)/H(U|S)]^k}{1 - [H(U|Y)/H(U|S)]} + L \\ &\leq \frac{N}{H(U|S) - H(U|Y)} + L \end{aligned}$$

where L denotes the length of the termination code. In other words, assuming $L \ll N$, the number of information bits per channel use is essentially $H(U|S) - H(U|Y) = I(U;Y) - I(U;S)$, the capacity. The probability of decoding error can readily be shown to diminish, taking k small enough so that the probability of an error in the Slepian–Wolf coding at any one of the k steps is negligible, yet large enough so that the length L of the termination code required to reliably transmit the last block (whose length decays exponentially with k) is negligible relative to N.

B. Wyner–Ziv Coding With Feedforward

Assume the Wyner–Ziv setting where source and SI are i.i.d. drawings of (X, Y). We further generate U according to $P_{U|X}$ (so $U - X - Y$), and let $\hat{X} = f(U, Y)$, taking $P_{U|X}$ and f to be achievers of the Wyner–Ziv function.

Shaping Subsystem: Given x^n which is P_X-typical, a shaper $S_{U|X}(\cdot, x^n)$ is a one-to-one mapping from $\{0,1\}^{nH(U|X)}$ into $T_{U|X}[x^n]$ (where $T_{U|X}[x^n]$ denotes the set of u^n's that are jointly typical with x^n). In other words, to every binary $nH(U|X)$-tuple b there corresponds a (different) $u^n = S_{U|X}(b, x^n)$ such that (u^n, x^n) are jointly typical. Let $S^{-1}_{U|X}(\cdot, x^n)$ denote the inverse mapping of $S_{U|X}(\cdot, x^n)$. Existence of shapers follows from elementary facts known from the method of types. Shaping systems can be implemented efficiently via arithmetic coding [6], [5].

Slepian–Wolf Coding: Given a typical u^n, let $C_U(u^n)$ denote the bit sequence of length $nH(U|Y)$ resulting from an essentially optimal Slepian–Wolf encoding of u^n for the presence of side information Y^n at the decoder. For b, a binary sequence of length $nH(U|Y)$, let $C_U^{-1}(b, y^n)$ denote the reconstruction of the corresponding decoder when receiving b from the encoder and the side information sequence is y^n.

Our Scheme: Fixing L, K, we take the length of the source sequence to be

$$n = L \sum_{j=0}^{k-1} [H(U|Y)/H(U|X)]^j.$$

The following scheme builds on the ideas in [5].

Encoding:

- Initialize $T = 1, l = L, j = 1$, and reverse the input so that $X^n \to (X_n, X_{n-1}, \ldots, X_1)$
- Take the block of source samples X^l and generate a "noisy version" U^l by passing X^l through the "channel" $P_{U|X}$.
- **while** $j < k$ **do:**
- Do Slepian–Wolf encoding of U_T^{T+l} to obtain the binary $lH(U|Y)$-tuple $b = C_U(U_T^{T+l})$. Let $T = T + l + 1, l = L[H(U|Y)/H(U|X)]^j, U_T^{T+l} = S_{U|X}(b, X_T^{T+l})$, and $j = j+1$
- **end while**
- **return** $b = C_U(U_T^{T+l}) = C_U(U_{n-l}^n)$

Decoding:

- Initialize $T = n, j = k - 1$.
- **while** $j \geq 0$ **do:**
- Let $l = L[H(U|Y)/H(U|X)]^j$ and $T = T - l$. Construct $\hat{X}_T^{T+l}$ by letting $\hat{X}_i = f(\hat{U}_i, Y_i)$ for each $T \leq i \leq T + l$, where $\hat{U}_T^{T+l} = C_U^{-1}(b, Y_T^{T+l})$. Obtaining X_T^{T+l} via the feedforward, let $b = S^{-1}_{U|X}(U_T^{T+l}, X_T^{T+l})$. Finally, let $j = j - 1$
- **end while**
- **return** the reversed version of $\hat{X}_1^n$

Performance: For k fixed and L large the Slepian–Wolf decoding is essentially error free, i.e., with high probability, at each of the k cycles of the **while** loop in the decoding $\hat{U}_T^{T+l} = U_T^{T+l}$. Furthermore, the reconstruction $\hat{X}_T^{T+l}$ obtained at each of the k cycles is, with high probability, jointly typical with X_T^{T+l}. Thus, the overall distortion is, with high probability, approximately $E\,\rho(X, \hat{X})$. As for the rate note that, by construction, the number of bits emitted by the encoder is $L[H(U|Y)/H(U|X)]^{k-1} \cdot H(U|Y)$, while the number of source samples encoded is

$$n = L \frac{[H(U|Y)/H(U|X)]^k - 1}{[H(U|Y)/H(U|X)] - 1}.$$

Thus, essentially, for $1 \ll k \ll L$, the rate achieved is

$$R \approx H(U|Y) - H(U|X) = I(U;X) - I(U;Y);$$

the optimal Wyner–Ziv rate.

ACKNOWLEDGMENT

N. Merhav would like to thank S. Shamai and Y. Steinberg for interesting discussions at the early stages of this work.

REFERENCES

[1] R. Ahlswede, "A constructive proof of the coding for discrete memoryless channel with feedback," in *Proc. 6th Prague Conf. Information Theory, Statistical Decision Functions, and Random Processes*, Prague, Czechoslovakia, 1971, pp. 39–50.

[2] M. H. M. Costa, "Writing on dirty paper," *IEEE Trans. Inf. Theory*, vol. IT-29, no. 3, pp. 439–441, May 1983.

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

[4] S. I. Gel'fand and M. S. Pinsker, "Coding for channel with random parameters," *Probl. Contr. and Inf. Theory*, vol. 9, no. I, pp. 19–31, 1980.

[5] E. Martinian and G. W. Wornell, "Source coding with fixed lag side information," in *Proc. 42nd Annu. Allerton Conf. Communications, Control and Computing*, Monticello, IL, Sep. 2004.

[6] J. M. Ooi and G. W. Wornell, "Fast iterative coding techniques for feedback channels," *IEEE Trans. Inf. Theory*, vol. 44, no. 7, pp. 2960–2976, Nov. 1998.

[7] S. S. Pradhan, "Source coding with feedforward: Gaussian sources," in *Proc. IEEE Int. Symp. Information Theory*, Chicago, IL, Jun./Jul. 2004, p. 212.

[8] J. P. M. Schalkwijk, "A coding scheme for additive noise channels with feedback-II: Band-limited signals," *IEEE Trans. Inf. Theory*, vol. IT-12, no. 2, pp. 183–189, Apr. 1966.

[9] J. P. M. Schalkwijk and T. Kailath, "A coding scheme for additive noise channels with feedback-i: No bandwidth constraint," *IEEE Trans. Inf. Theory*, vol. IT-12, no. 2, pp. 183–189, Apr. 1966.

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

[11] D. Slepian and J. K. Wolf, "Noiseless coding of correlated information sources," *IEEE Trans. Inf. Theory*, vol. IT-19, no. 4, pp. 471–480, Jul. 1973.

[12] T. Weissman and N. Merhav, "On competitive prediction and its relationship to rate-distortion theory," *IEEE Trans. Inf. Theory*, vol. 49, no. 12, pp. 3185–3193, Dec. 2003.

[13] A. D. Wyner, "The rate distortion function for source coding with side information at the decoder-II: General sources," *Inf. Contr.*, vol. 38, pp. 60–80, 1978.

[14] A. D. Wyner and J. Ziv, "The rate distortion function for source coding with side information at the decoder," *IEEE Trans. Inf. Theory*, vol. IT-22, no. 1, pp. 1–10, Jan. 1976.