Opportunistic Orthogonal Writing on Dirty Paper

Tie Liu, *Student Member, IEEE,* and Pramod Viswanath, *Member, IEEE*

***Abstract*—A simple scheme that achieves the capacity and the reliability function of the wideband Costa dirty-paper channel is proposed. The scheme can be interpreted as an opportunistic version of pulse position modulation (PPM). This interpretation suggests a natural generalization of the scheme which we show to achieve the capacity per unit cost of Gel'fand–Pinsker channels with a zero-cost input letter.**

***Index Terms*—Capacity per unit cost, dirty-paper channel, Gel'fand–Pinsker channel, opportunistic communication, orthogonal signaling, pulse position modulation, pulse position quantization, Wyner–Ziv source coding.**

I. Introduction

THE Shannon capacity of state-dependent discrete memoryless channels where the channel states are noncausally known to the transmitter as side information was characterized by Gel'fand and Pinsker [1].[1] The result was popularized by Costa through his whimsically titled "Writing on Dirty Paper" [3]. Central to this line of research is a powerful technique called *binning*, which promises considerable gain in rate of reliable communication at the expense of increased complexity in the design of encoding algorithm. Several recent works [4]–[7] study the algebraic and coding structure of the random binning scheme used in [1] and [3].

In this paper, we consider the problem of coding for the wideband Costa's dirty-paper channel. Whereas the coding problem for the additive white Gaussian noise (AWGN) channel is involved, there is an explicit scheme in the wideband regime: A simple orthogonal signaling scheme achieves the channel capacity. We ask for the natural extension of this result to the wideband Costa's dirty-paper channel. Our main result is the demonstration of such a scheme which we refer to as opportunistic orthogonal signaling.

We start with an orthogonal set of codewords representing M messages. Each of the codewords is replicated K times so that the overall constellation with MK vectors forms an orthogonal set. Each of the M messages corresponds to a set of K orthogonal signals. To convey a specific message, the encoder transmits the signal (among the set of K orthogonal signals corresponding to the selected message) that has the largest correlation with the interference. An equivalent way of seeing this scheme is as opportunistic pulse position modulation (PPM). Standard PPM involves a single pulse that conveys information based on the position where it is nonzero. Here, every K of the pulse positions corresponds to one message, and the encoder opportunistically chooses the position of the pulse (among K possibilities once the desired message to be conveyed is picked) where the interference is the *largest*. The decoder first picks the most likely position of the transmit pulse (among MK possible choices) using the standard largest amplitude detector. Next, it picks the message corresponding to the set in which the most likely pulse occurred. Choosing K large allows the encoder to harness the opportunistic gain afforded by the knowledge of the additive interference. On the other hand, decoding gets harder as K increases since the number of possible pulse positions, MK, also grows with K. We elaborate on this tradeoff in Sections II and III and show that the correct choice of K allows opportunistic orthogonal signaling to achieve both the capacity and the reliability function of the wideband Costa's dirty-paper channel.

Each bin in the binning scheme can be thought of as a quantizer for the interference, while the codewords in each of the bins put together form a good channel code; this is the *nested coding* interpretation [4] of the abstract binning scheme. In Section IV we show that opportunistic PPM fits this interpretation. We first point out a simple vector quantizer for the wideband Gaussian source which we refer to as pulse position quantization (PPQ). Next, we observe that opportunistic PPM is a combination of PPQ (a good low-rate vector quantizer for the wideband Gaussian source) and PPM (a good channel code for the wideband AWGN channel). This interpretation suggests a natural generalization of the opportunistic PPM scheme to Gel'fand–Pinsker channels with an input cost constraint: Use a cost-efficient vector quantizer to form the codewords within a bin such that all the codewords put together form a cost-efficient channel code. Cost-efficient vector quantizers and channel codes studied by Verdú [13] form the basic constituents of this generalized opportunistic PPM scheme which we show to achieve the capacity per unit cost of Gel'fand–Pinsker channels with a zero-cost input letter; this is done in Section V. The natural source-coding analog of opportunistic PPM is in the study of low-rate quantization of the wideband Gaussian source with the decoder having noncausal access to a noisy version of the source; this is the topic of Section VI.

II. Writing on Wideband Dirty Paper

Consider the continuous-time Costa's dirty-paper channel

$$y(t) = x(t) + s(t) + n(t) \tag{1}$$

where $x(t)$, $s(t)$ and $n(t)$ are the transmit signal, the interference and the background noise. $s(t)$ and $n(t)$ are independent

Manuscript received January 18, 2005; revised June 6, 2005. The material in this paper was presented in part at the International Conference on Signal Processing and Communications, Bangalore, India, December 2004.

The authors are with the Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign, Urbana, IL 61801 USA (e-mail: tieliu@ifp.uiuc.edu; pramodv@uiuc.edu).

Communicated by M. Médard, Associate Editor for Communications.
Digital Object Identifier 10.1109/TIT.2006.872979

[1]The same result was independently derived by El Gamal and Heegard [2] around the same time.

white Gaussian processes with two-sided power spectral density $N_s/2$ and $N_0/2$, respectively. In this channel, $s(t)$ is *non-causally* known at the input side, and this knowledge can be used to encode the message over the entire block. On the other hand, $s(t)$ is only known *statistically* to the receiver at the output side of the channel. We consider communication in the wideband regime: $x(t)$ is power limited but not bandwidth limited. We use the standard $\frac{\varepsilon_b}{N_0}|_{\min}$, the minimum energy per bit normalized by N_0 for reliable communication, as our primary performance criterion. Since the maximum information rate that can be reliably transmitted at is monotonically increasing with the bandwidth, determining $\frac{\varepsilon_b}{N_0}|_{\min}$ of one channel is equivalent to seeking the wideband limit of its capacity. As shown in [3], the capacity of Costa's dirty-paper channel is the same as that of the zero-interference AWGN channel. Therefore, the minimum energy per bit for reliable communication over Costa's dirty-paper channel is the same as that of the zero-interference AWGN channel: $\frac{\varepsilon_b}{N_0}|_{\min} = \log_e 2 = -1.59$ dB. Below, we formally state the opportunistic PPM scheme and show that it achieves this value of minimum energy per bit for reliable communication over the wideband Costa's dirty-paper channel.

Standard PPM involves a single pulse (of total energy ε_s) that conveys information based on the position when it is nonzero. Here, each message $m \in \mathcal{I}_M \stackrel{\text{def}}{=} \{1, \dots, M\}$ is associated with multiple (say K) subpulse positions. The transmitter opportunistically chooses the subpulse position (among K possibilities) where the interference is the largest. The receiver first picks the most likely position of the transmit subpulse (among MK possibilities) using the standard largest amplitude detector. It then claims the message to be the one that corresponds to the pulse position within which the most likely subpulse occurred. A depiction of this encoding/decoding process is shown in Fig. 1.

This scheme can be equivalently described with the following discrete-time representation.

A. Transmitter

Associate each message $m \in \mathcal{I}_M$ with K orthogonal vectors

$$\boldsymbol{x}(m,k) = (0, \dots, \sqrt{\varepsilon_s}, \dots, 0), \qquad k \in \mathcal{I}_K \tag{2}$$

where the only nonzero entry $\sqrt{\varepsilon_s}$ is in the $((m-1)K+k)th$ position. Given a message m and an interference vector $\boldsymbol{s} = (s_1, \dots, s_{MK})$, choose the position that corresponds to the largest among $\{s_{(m-1)K+k} : k \in \mathcal{I}_K\}$ to transmit. That is, the actual transmit vector is $\boldsymbol{x}(m, k^*)$ where

$$k^* = \arg \max_{1 \le k \le K} s_{(m-1)K+k}. \tag{3}$$

B. Channel

The channel corrupts the transmit vector $\boldsymbol{x}(m, k^*)$ by superimposing two independent random vectors $\boldsymbol{s}$ and $\boldsymbol{n}$. The entries of $\boldsymbol{s}$ and $\boldsymbol{n}$ are independent and identically distributed (i.i.d.) Gaussian variables with zero mean and variance $N_s/2$ and $N_0/2$, respectively.

C. Receiver

The correlation demodulator has a bank of MK outputs $\boldsymbol{r} = (r_1, \dots, r_{MK})$. For the branch where the transmit subpulse is nonzero, the output is

$$r_{(m-1)K+k^*} = \sqrt{\varepsilon_s} + s_{(m-1)K+k^*} + n_{(m-1)K+k^*} \tag{4}$$

where

$$s_{(m-1)K+k^*} = \max_{1 \le k \le K} s_{(m-1)K+k}. \tag{5}$$

Otherwise, the output is

$$r_{(i-1)K+j} = s_{(i-1)K+j} + n_{(i-1)K+j}. \tag{6}$$

The estimated message is given by

$$\hat{m} = \arg \max_{1 \le i \le M} \left(\max_{1 \le j \le K} r_{(i-1)K+j} \right). \tag{7}$$

An error occurs if and only if $\hat{m}$ is not unique or it is unique but $\hat{m} \neq m$.

D. Error Analysis

Choosing K large allows the transmitter to harness the opportunistic gain afforded by the knowledge of the additive interference. On the other hand, decoding gets harder as K increases because the number of rival codewords, $(M-1)K$, also grows with K. The following lemma [8, pp. 264–265] taken from the theory of order statistics allows for a precise characterization of the opportunistic gain.

Lemma 1: Suppose $s_1, \dots, s_K$ are i.i.d. Gaussian variables with zero mean and variance $N_s/2$ ($\mathcal{N}(0, N_s/2)$) and $s_{k^*} = \max_{1 \le k \le K} s_k$. Then

$$\sqrt{N_s \log_e K} \left(s_{k^*} - \sqrt{N_s \log_e K} \right) \tag{8}$$

converges in distribution to a limiting random variable with cumulative distribution function

$$\exp\left(-e^{-x}\right), \quad -\infty < x < \infty \tag{9}$$

in the limit as $K \to \infty$. Furthermore, the moments of (8) converge to the corresponding moments of the limiting distribution (9).

Since $\sqrt{N_s \log_e K}$ tends to infinity in the limit as $K \to \infty$, by Lemma 1, $s_{k^*} - \sqrt{N_s \log_e K}$ tends to zero in distribution (and, equivalently, in probability) in the limit as $K \to \infty$.

The error probability is clearly independent of the message being transmitted. Assume $m = 1$, and all the probabilities below will be tacitly understood to be conditioned on that event. The error probability

$$\begin{aligned} P_e &= \mathbb{P}(\hat{m} \neq 1) && (10)\\ &\le \mathbb{P}\left(\max_{2 \le i \le M} \max_{1 \le j \le K} r_{(i-1)K+j} \ge r_{k^*} \right) && (11)\\ &\le \mathbb{P}(r_{k^*} \le \gamma) + \mathbb{P}\left(\max_{2 \le i \le M} \max_{1 \le j \le K} r_{(i-1)K+j} \ge \gamma \right) && (12) \end{aligned}$$

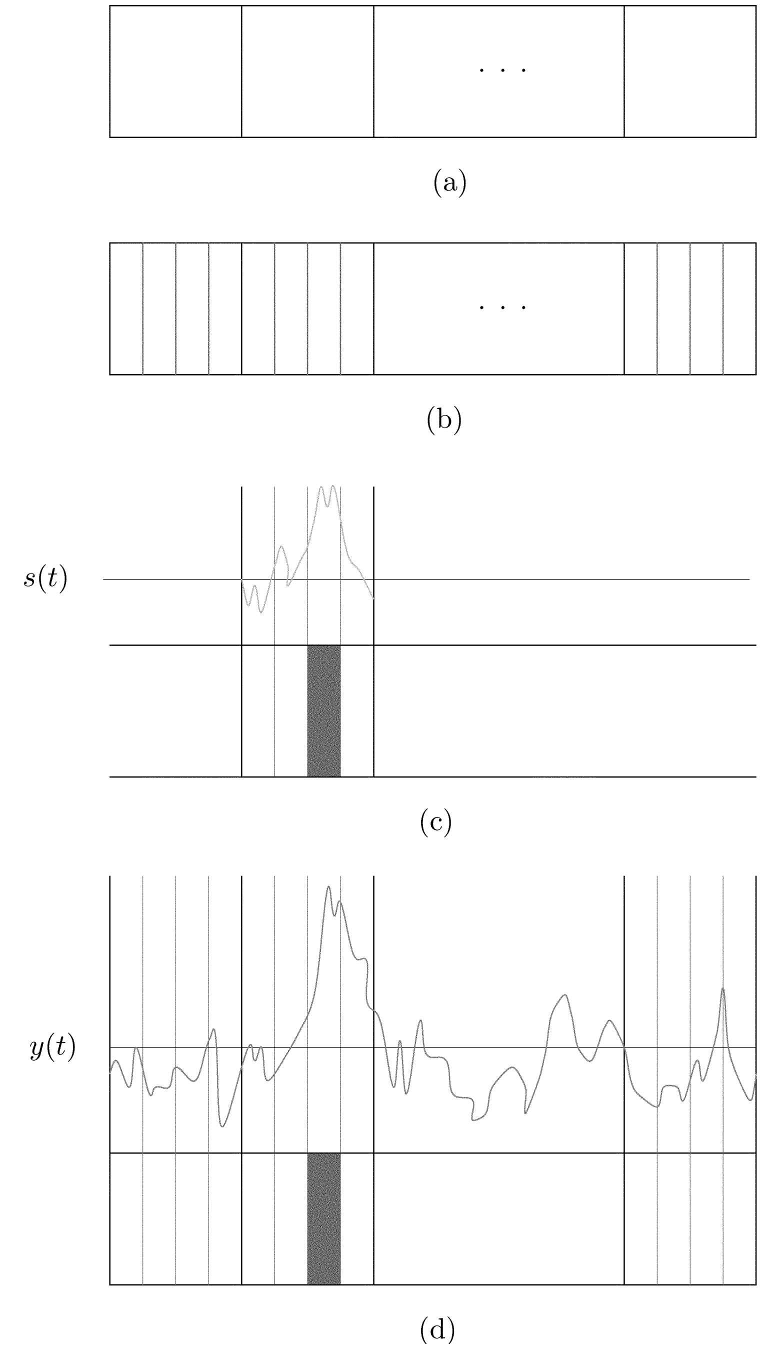

Fig. 1. Opportunistic PPM: (a) M pulses correspond to M messages. (b) Each pulse splits into K subpulses. (c) The transmitter chooses the subpulse position (within the pulse position corresponding to the selected message m) where the interference $s(t)$ is the largest. (d) The receiver picks the subpulse position where the received signal $y(t)$ is the largest and claims the message to be the one corresponding to the pulse containing the subpulse being picked out.

for any real number γ. The right-hand side of (12) has an operational meaning: It is the error probability of decoding by performing MK independent binary hypothesis tests on the outputs $r_{(i-1)K+j}$, $(i,j) \in \mathcal{I}_M \times \mathcal{I}_K$, with γ being the threshold of the tests. Decoding via binary hypothesis testing is generally suboptimal. However, as we shall see shortly, if the threshold γ of the tests is appropriately chosen, it suffices for the desired result.

Fix $0 < \delta < N_s/N_0$ and let the threshold

$$\gamma = \sqrt{\varepsilon_s} + \sqrt{N_s \log_e K} - \delta\sqrt{N_0 \log_e M}. \qquad (13)$$

Define the energy per bit as $\varepsilon_b \stackrel{\text{def}}{=} \varepsilon_s / \log_2 M$. Choose the number of subpulses K associated with each message as

$$\log_e K = \frac{N_s}{N_0} \log_e M. \qquad (14)$$

We now show that, for any $\frac{\varepsilon_b}{N_0} > (1+\delta)^2 \log_e 2$, the error probability P_e can be made as small as possible if we allow the number of messages M to be arbitrarily large. By (4), the first term on the right-hand side of (12) can be written as

$$\begin{aligned} &\mathbb{P}(r_{k^*} \le \gamma) \\ &= \mathbb{P}\left(s_{k^*} + n_{k^*} \le \sqrt{N_s \log_e K} - \delta\sqrt{N_0 \log_e M}\right) \end{aligned} \qquad (15)$$

$$\leq \mathbb{P}\left(s_{k^*} - \sqrt{N_s \log_e K} \leq -\delta\right) + \mathbb{P}\left(n_{k^*} \leq -\delta\left(\sqrt{N_0 \log_e M} - 1\right)\right). \quad (16)$$

By Lemma 1, $s_{k^*} - \sqrt{N_s \log_e K}$ converges to 0 in probability. So we have

$$\mathbb{P}\left(s_{k^*} - \sqrt{N_s \log_e K} \leq -\delta\right) \to 0 \quad (17)$$

in the limit as $K \to \infty$. n_{k^*} is distributed as $\mathcal{N}(0, N_0/2)$, which gives

$$\mathbb{P}\left(n_{k^*} \leq -\delta\left(\sqrt{N_0 \log_e M} - 1\right)\right) \to 0 \quad (18)$$

in the limit as $M \to \infty$. By (14), $M \to \infty$ implies $K \to \infty$. We conclude from (16)–(18) that

$$\mathbb{P}(r_{k^*} \leq \gamma) \to 0 \quad (19)$$

in the limit as $M \to \infty$.

The second term on the right-hand side of (12) can be bounded from above as follows. Let $\mathcal{I}'_M \overset{\text{def}}{=} \mathcal{I}_M - \{1\}$. Since $r_{(i-1)K+j}$, $(i,j) \in \mathcal{I}'_M \times \mathcal{I}_K$, are i.i.d. as $\mathcal{N}(0, (N_s + N_0)/2)$, we have

$$\mathbb{P}(r_{(i-1)K+j} \geq \gamma) = Q\left(\frac{\gamma}{\sqrt{(N_s+N_0)/2}}\right) < \exp\left(-\frac{\gamma^2}{N_s+N_0}\right), \quad \forall\, (i,j) \in \mathcal{I}'_M \times \mathcal{I}_K. \quad (20)$$

It follows that

$$\mathbb{P}\left(\max_{2\leq i\leq M} \max_{1\leq j\leq K} r_{(i-1)K+j} \geq \gamma\right) \leq \sum_{i=2}^{M}\sum_{j=1}^{K} \Pr(r_{(i-1)K+j} \geq \gamma) < MK \exp\left(\frac{-\gamma^2}{N_s+N_0}\right). \quad (21)$$

Substituting (13) and (14) into (21), we arrive at (22) shown at the bottom of the page. This exponential upper bound tends to zero in the limit as $M \to \infty$ so long as

$$\frac{\varepsilon_b}{N_0} > (1+\delta)^2 \log_e 2. \quad (23)$$

Combining (22) with (19), we conclude that the error probability P_e can be made arbitrarily small for sufficiently large M as long as (23) stands. Note that δ can be made arbitrarily close to zero. Thus, reliable communication is achieved by the opportunistic PPM scheme with $\frac{\varepsilon_b}{N_0}$ arbitrarily close to $\log_e 2$. The following theorem summarizes this result.

Theorem 2: Opportunistic orthogonal signaling achieves $\frac{\varepsilon_b}{N_0}|_{\min} = \log_e 2 = -1.59$ dB for reliable communication over the wideband Costa's dirty-paper channel.

III. Ramifications of Opportunistic PPM

A. Comments on "Zero Rate Loss" of Opportunistic PPM

We give some insights into why opportunistic PPM achieves the same $\frac{\varepsilon_b}{N_0}|_{\min}$ for the wideband Costa's dirty-paper channel as that of standard PPM for the wideband zero-interference AWGN channel. These insights provide the intuition on how to extend opportunistic PPM to wideband dirty-paper channels with i.i.d. non-Gaussian "dirt" and, more generally, to Gel'fand–Pinsker channels with an input cost constraint. A byproduct is a natural view of (14) being the correct choice of K as the number of subpulses associated with each message.

To simplify the notation, we use symbol "$\doteq$" to represent equality in the exponential scale of $\log_e M$. To be specific, we use $f(M) \doteq g(M)$ if and only if

$$\lim_{M\to\infty} \frac{\log_e(f(M)/g(M))}{\log_e M} = 0.$$

In light of the error analysis in Section II, the effective signaling amplitude in opportunistic PPM is

$$\sqrt{\varepsilon'_s} = \sqrt{\varepsilon_b \log_e M} + \sqrt{N_s \log_e K} \quad (24)$$

where $\sqrt{N_s \log_e K}$ is the opportunistic gain afforded by the Gaussian tail of the interference distribution. The error probability $P_e^{(\mathrm{OPPM})}$ of opportunistic PPM can be written as

$$P_e^{(\mathrm{OPPM})} \doteq MK\mathbb{P}\left(s + n \geq \sqrt{\varepsilon'_s}\right) \quad (25)$$

$$\doteq MK\mathbb{P}\left(s \geq \frac{N_s}{N_s+N_0}\sqrt{\varepsilon'_s}\right)\mathbb{P}\left(n \geq \frac{N_0}{N_s+N_0}\sqrt{\varepsilon'_s}\right). \quad (26)$$

Here, we drop all the subscripts and use generic s and n to represent the interference and the noise. The order equality (26) follows from the fact that both s and n are Gaussian, and Gaussian distribution is a stable law under convolution. The error probability of standard PPM in the wideband zero-interference AWGN channel [11, pp. 379–383] is

$$P_e^{(\mathrm{PPM})} \doteq M\mathbb{P}\left(n \geq \sqrt{\varepsilon_b \log_2 M}\right). \quad (27)$$

Now, if there exists a choice of K such that

$$K\mathbb{P}\left(s \geq \frac{N_s}{N_s+N_0}\sqrt{\varepsilon'_s}\right) \doteq 1 \quad \text{and} \quad \frac{N_0}{N_s+N_0}\sqrt{\varepsilon'_s} = \sqrt{\varepsilon_b \log_2 M} \quad (28)$$

can be simultaneously satisfied, then opportunistic PPM and standard PPM would achieve the same $\frac{\varepsilon_b}{N_0}$. Note that (28) is equivalent to requiring that the boost of signaling strength by the opportunistic gain completely cancel the detrimental effect of having more competing codewords. It is a matter of simple algebra to verify that (14) indeed gives the uniquely correct choice

$$\mathbb{P}\left(\max_{2\leq i\leq M} \max_{1\leq j\leq K} r_{(i-1)K+j} \geq \gamma\right) < \exp\left(-\frac{\left(\sqrt{\frac{\varepsilon_b}{N_0}\log_2 e} - \delta - 1\right)\left(\sqrt{\frac{\varepsilon_b}{N_0}\log_2 e} - \delta + 1 + \frac{2N_s}{N_0}\right)}{1 + \frac{N_s}{N_0}} \log_e M\right). \quad (22)$$

of K, so that both requirements in (28) are satisfied at the same time.

B. Extension to i.i.d. Non-Gaussian "Dirt"

The above analysis suggests that the success of the opportunistic PPM scheme hinges on the Gaussian tail of both the interference and the noise distribution. However, a surprising result, shown by Cohen and Lapidoth [9, Sec. II-D], is that the capacity of any dirty-paper channel with i.i.d. non-Gaussian "dirt" (which needs to have a finite second moment) is also the same as that of the zero-interference AWGN channel. Carrying this argument to the wideband regime, we immediately come to the conclusion that the minimum energy per bit normalized by N_0 for reliable communication over any wideband dirty-paper channel with i.i.d. non-Gaussian "dirt" is also -1.59 dB. In light of the discussion in Section III.A, applying opportunistic PPM directly to such channels cannot achieve this value of $\frac{\varepsilon_b}{N_0}$ due to the non-Gaussian tail of the interference distribution. However, it turns out that there is a simple remedy to the basic opportunistic PPM scheme so that $\frac{\varepsilon_b}{N_o}|_{\min} = -1.59$ dB can still be achieved.

Consider the continuous-time wideband dirty-paper channel

$$y(t) = x(t) + s(t) + n(t) \tag{29}$$

where the channel input $x(t)$ is power limited but not bandwidth limited; the interference $s(t)$ is an independent but *non-Gaussian* process with two-sided power spectral density $N_s/2$; and $n(t)$ is the usual white Gaussian noise with two-sided power spectral density $N_0/2$ and is assumed to be independent of $s(t)$. In the basic form of the opportunistic PPM scheme, the length of the pulse is irrelevant: The only aspect of the pulse that affects the calculation is its energy. This is because both the interference and noise are Gaussian, and their statistics remain unchanged under an averaging operation. With non-Gaussian interference, however, the length of the pulse also has a role to play. In particular, we can use the central limit theorem (CLT) to make the effective interference look like Gaussian. This scheme can be equivalently described with the following two-dimensional discrete-time representation, see Fig. 2.

The K transmit signals associated with the message m are $N \times MK$ matrices

$$X(m,k) = \begin{pmatrix} 0 & \cdots & \sqrt{\frac{\varepsilon_s}{N}} & \cdots & 0 \\ 0 & \cdots & \sqrt{\frac{\varepsilon_s}{N}} & \cdots & 0 \\ \cdot & & \cdot & & \cdot \\ \cdot & & \cdot & & \cdot \\ 0 & \cdots & \sqrt{\frac{\varepsilon_s}{N}} & \cdots & 0 \end{pmatrix}, \quad k \in \mathcal{I}_K \tag{30}$$

where the $((m-1)K+k)$th column is the only nonzero one. A total energy of ε_s has been evenly split within the subpulse of length N, so each entry in the $((m-1)K+k)$th column is equal to $\sqrt{\varepsilon_s/N}$.

Based on its noncausal knowledge on the interference $s(t)$, the transmitter chooses the actual transmit signal (among K possibilities) according to some opportunistic rule (which will become clear shortly). Upon the reception of $r_{l,(i-1)K+j}$, $l \in \mathcal{I}_N$, $(i,j) \in \mathcal{I}_M \times \mathcal{I}_K$, the decoder performs the following CLT type averaging on each of the columns:

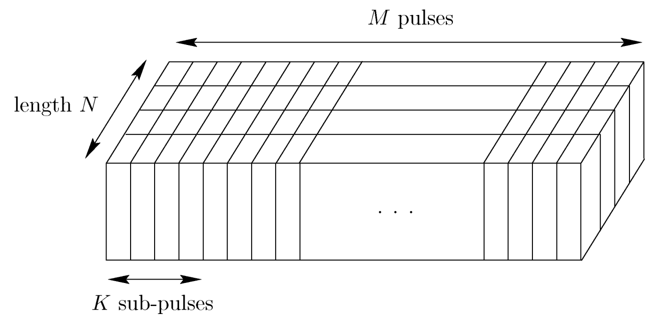

Fig. 2. Generalized opportunistic PPM: M pulses correspond M messages. Each pulse splits into K subpulses of length N. The block length of this transmission scheme is MKN.

$$\overline{r}_{(i-1)K+j} \stackrel{\text{def}}{=} \frac{1}{\sqrt{N}} \sum_{l=1}^{N} r_{l,(i-1)K+j} \tag{31}$$

$$= \frac{1}{\sqrt{N}} \sum_{l=1}^{N} x_{l,(i-1)K+j} + \underbrace{\frac{1}{\sqrt{N}} \sum_{l=1}^{N} s_{l,(i-1)K+j}}_{\stackrel{\text{def}}{=} \overline{s}_{(i-1)K+j}} + \underbrace{\frac{1}{\sqrt{N}} \sum_{l=1}^{N} n_{l,(i-1)K+j}}_{\stackrel{\text{def}}{=} \overline{n}_{(i-1)K+j}}. \tag{32}$$

For the branch where the subpulse is nonzero, the averaging output is

$$\overline{r}_{(m-1)K+k^*} = \sqrt{\varepsilon_s} + \overline{s}_{(m-1)K+k^*} + \overline{n}_{(m-1)K+k^*}. \tag{33}$$

Otherwise, the averaging output is

$$\overline{r}_{(i-1)K+j} = \overline{s}_{(i-1)K+j} + \overline{n}_{(i-1)K+j}. \tag{34}$$

Now it should be clear what the opportunistic rule should be. The transmitter should choose the transmit subpulse position k^* (among K possibilities) where the averaging interference is the largest

$$k^* = \arg \max_{1 \le k \le K} \overline{s}_{(m-1)K+k}. \tag{35}$$

Note that the averaging preserves the signaling strength and the Gaussianity of the noise distribution. Moreover, by the CLT, the averaging interference $\overline{s}_{(i-1)K+j}$ can be made arbitrarily close to a Gaussian distribution if N_s is positive, finite and N, the length of the subpulse, is sufficiently large. Since both the encoding and the decoding are based on the averages for which the effective interference and noise are Gaussian, the proposed scheme achieves the same $\frac{\varepsilon_b}{N_0}|_{\min}$ as that by the basic opportunistic PPM scheme in the wideband Costa's dirty-paper channel.

We note that this extension of the basic opportunistic PPM scheme is a substantial abuse of the degrees of freedom: The block length for transmitting M messages is MKN which tends to infinity in the limit as the length of averaging $N \to \infty$.

Therefore, this scheme should be mostly thought of as an achievability proof of $\frac{\mathcal{E}_b}{N_0}|_{\min} = -1.59$ dB for the wideband dirty-paper channel with i.i.d. non-Gaussian "dirt". In Section V, we shall give an alternative scheme which achieves the same $\frac{\mathcal{E}_b}{N_0}|_{\min}$ but with a much more efficient use of the available bandwidth. The final remark here is that the above extension works for arbitrary independent "dirt" as long as the CLT stands. A general sufficient condition for the CLT to stand is the Feller–Lindeberg condition [10, Theorem 3.18]. Our extension achieves the minimum energy per bit for reliable communication over any dirty-paper channel in which the law of the "dirt" satisfies the Feller–Lindeberg condition.

C. The Error Exponents

We have shown that the capacity of the wideband Costa's dirty-paper channel is the same as that of the wideband zero-interference AWGN channel and is achieved by opportunistic PPM. A natural question to ask next is how opportunistic PPM performs in terms of the channel reliability. The upper and lower bounds on the error probability of orthogonal signaling for the infinite-bandwidth AWGN channel have been derived and shown to coincide in the exponential scale [11, pp. 378–383]. So the wideband AWGN channel is one of very few channels whose reliability function has been completely characterized. In the following theorem, we derive an exponential upper bound on the error probability of opportunistic PPM for the wideband Costa's dirty-paper channel. The derived exponents (asymptotically) coincide with the reliability function of the wideband AWGN channel for all rates up to channel capacity. Since the error exponents of the wideband Costa's dirty-paper channel cannot exceed that of the wideband zero-interference AWGN channel, the exponents we derived completely characterize the reliability function of the wideband Costa's dirty-paper channel.

Theorem 3: The error probability of opportunistic PPM in the wideband Costa's dirty-paper channel can be bounded from above as

$$P_e < (2+o(1))\exp\left(-\left(E\left(\frac{\mathcal{E}_b}{N_0}\right)-o(1)\right)\log_e M\right) \quad (36)$$

where $o(1)$ tends to zero in the limit as $M\to\infty$, and $E(x)$ is the reliability function of the wideband zero-interference AWGN channel:

$$E(x) = \begin{cases} \left(\sqrt{x}-\sqrt{\log_e 2}\right)^2, & \log_e 2 < x \le 4\log_e 2 \\ \frac{1}{2}x - \log_e 2, & x \ge 4\log_e 2. \end{cases} \quad (37)$$

In light of the (coarse) error analysis in Section II, there are two typical ways for opportunistic PPM to make an error.

- *Encoding error:* the opportunistic gain $s_{(m-1)K+k^*}$ is not large enough.
- *Decoding error:* either $n_{(m-1)K+k^*}$ is too small, or $s_{(i-1)K+j} + n_{(i-1)K+j}$ is too large for some $(i,j)\in\mathcal{I}_M\times\mathcal{I}_K$ with $i\neq m$.

It turns out that whereas the decoding error probability decays exponentially with $\log_e M$, the encoding error probability decays *superexponentially* with $\log_e M$. Therefore, for sufficiently large M, the dominating error event is the decoding error which we show to possess the same decay rate as that of standard PPM for the AWGN channel if the number of subpulse positions associated with each message is correctly chosen. The formal proof amounts to making this argument mathematically precise; the details are deferred to Appendix A.

IV. INSIGHTS FROM OPPORTUNISTIC PPM

A. Connections to the Binning Scheme

As we have seen, both random binning and opportunistic PPM achieve the capacity of wideband dirty-paper channels. Common to both schemes is to associate multiple codewords to each message, and the encoder picks one of them based on the knowledge of the additive interference. This statement is further strengthened by the following observation.

Observation: To achieve the capacity of the wideband dirty-paper channel with i.i.d. Gaussian/non-Gaussian "dirt," the number of codewords in each bin in the random binning scheme is the same as the number of subpulse positions associated with each message in opportunistic PPM.

Denote by K the number of codewords in each bin in the random binning scheme. Recall from [3, Sec. II] and [9, Sec. II-D] that

$$\log_e K = \lim_{W\to\infty} 2WI(U;S). \quad (38)$$

Here, W is the real bandwidth of communication, and S is the interference (possibly non-Gaussian) with zero mean and variance $N_s W$. The auxiliary variable is $U = X+\alpha S$ where X is distributed as $\mathcal{N}(0,P)$

$$\alpha = \frac{P}{P+N_0W} \sim \frac{P}{N_0W} \quad (39)$$

in the limit as $W\to\infty$, and $N_0/2$ is the two-sided power spectral density of the additive white Gaussian noise. We have

$$\log_e K = \lim_{W\to\infty} 2WI\left(X+\frac{P}{N_0W}S;S\right) \quad (40)$$

$$= \frac{2P^2N_s}{N_0^2}\lim_{t\to 0} t^{-1}I\left(X+\sqrt{t}\widetilde{S};\widetilde{S}\right) \quad (41)$$

$$= \frac{PN_s}{N_0^2} \quad (42)$$

irrespective of the distribution of the interference S [12, Lemma 5.2.1]. Here, $\widetilde{S} \stackrel{\text{def}}{=} \frac{1}{\sqrt{N_S W}}S$, so it has a unit variance, and $t \stackrel{\text{def}}{=} \frac{P^2N_s}{N_0^2W}$. Let M be the maximum number of messages that can be reliably transmitted. By definition, $\log_e M$ is the capacity of the wideband dirty-paper channel, i.e.,

$$\log_e M = \frac{P}{N_0}. \quad (43)$$

Substituting (43) into (42), we have

$$\log_e K = \frac{N_s}{N_0}\log_e M. \quad (44)$$

Comparing (44) with (14), our observation is confirmed.

In light of the above observation, one may wonder if opportunistic PPM can be thought of as a structured binning scheme. In particular, structured binning such as algebraic binning [4] can be interpreted as a nested coding scheme: Each bin is a good quantizer and all bins put together form a good channel code. It is natural to ask if opportunistic PPM fits this interpretation. It is well known that orthogonal codes are capacity-achieving channel codes for the wideband AWGN channel. Next, we show that orthogonal codes are also good quantizers for the wideband Gaussian source.

B. Pulse Position Quantization (PPQ)

Suppose $s(t)$ is an ideal bandlimited Gaussian process with real bandwidth W and two-sided power spectral density $N_S/2$. The degrees of freedom per unit time for this source are asymptotically $2W$ for large W [11, p. 373]. Suppose we have a total rate budget of r bits per unit time. Classical rate-distortion theory states that the minimum *total* mean-squared error distortion for reproducing this source is

$$d = 2WD\left(\frac{r}{2W}\right) \tag{45}$$

where

$$D(R) = \left(\frac{N_s}{2}\right) 2^{-2R} \tag{46}$$

is the quadratic Gaussian distortion-rate function. We are interested in the case where the bandwidth of the source $W \to \infty$. Then, the total mean-squared error distortion

$$d = \lim_{W\to\infty} N_s W 2^{-\frac{r}{W}} = \infty \tag{47}$$

no matter how large the rate budget r is as long as it is finite. This suggests that the classical distortion measure is no longer useful in the wideband regime. However, if we consider the reward function

$$\delta_D(R) = D(0) - D(R) \tag{48}$$

the distortion reduction per degree of freedom by describing it using R bits, then the maximum mean-squared error distortion *reduction* per unit time that can be obtained from a rate budget of r bits per unit time is

$$\delta_d = \lim_{W\to\infty} 2W\delta_D(r) = \lim_{W\to\infty} N_s W\left(1-2^{-\frac{r}{W}}\right) = rN_s \log_e 2. \tag{49}$$

This is *one* most efficient way of using the bit budget for the wideband Gaussian source; the corresponding bit efficiency is $\delta_d/r = N_s \log_e 2$. As we shall see shortly, orthogonal codes can achieve this value of bit efficiency as well. Therefore, they are also good quantizers for the wideband Gaussian source.

Let $\boldsymbol{s} = (s_1, \ldots, s_K)$ be the time sample of $s(t)$ where $K = 2W$. Then s_k, $k \in \mathcal{I}_K$, are i.i.d. as $\mathcal{N}(0, N_s/2)$. To encode the source, the encoder simply looks at the entries of $\boldsymbol{s}$ and transmit the index k^* for which s_{k^*} is the largest. By symmetry, k^* is uniformly distributed over $\mathcal{I}_K$. So we need a total of $\log_2 K$ bits to describe it. The reconstruction uses an orthogonal code. Given k^*, the decoder produces a reconstruction

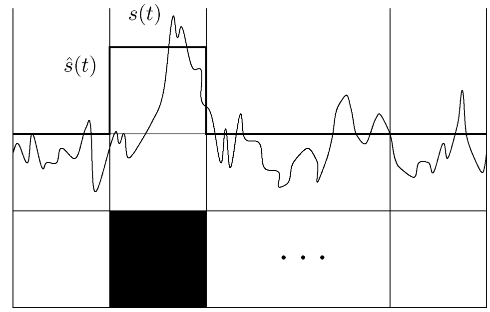

Fig. 3. PPQ: The encoder informs the decoder the position where the source is the largest. The decoder reconstructs the source by $\sqrt{N_s \log_e K}$ at the informed position and 0 otherwise.

$$\hat{\boldsymbol{s}} = (0, \ldots, \sqrt{N_s \log_e K}, \ldots, 0) \tag{50}$$

with the only nonzero entry $\sqrt{N_s \log_e K}$ in the k^*th position. If the encoder does not provide any description about the source, the best reconstruction letter is 0. So the total distortion reduction of the above quantization scheme is

$$\begin{aligned}\delta_d &= \mathbb{E}\left[(s_{k^*} - 0)^2\right] - \mathbb{E}\left[(s_{k^*} - \sqrt{N_s \log_e K})^2\right] &(51)\\ &= N_s \log_e K + 2\,\mathbb{E}\left[\sqrt{N_s \log_e K}\left(s_{k^*} - \sqrt{N_s \log_e K}\right)\right] &(52)\\ &= N_s \log_e K + o(\log_e K) &(53)\end{aligned}$$

in the limit as $K \to \infty$. Here, (53) follows from that $\sqrt{N_s \log_e K}\left(s_{k^*} - \sqrt{N_s \log_e K}\right)$ converges to a limiting random variable in the mean according to Lemma 1. Thus, the bit efficiency of the above quantization scheme is $N_s \log_e K / \log_2 K = N_s \log_e 2$, which is the highest efficiency possible for the wideband Gaussian source. A depiction of the above encoding/decoding procedure is shown in Fig. 3.

Note that the nature of the above quantization scheme is not to quantize the amplitude of each degree of freedom (as suggested by the classical rate-distortion theory), but rather to inform the decoder the position of the degree of freedom which would cause the largest distortion. We call this scheme pulse position quantization (PPQ), as a counterpart of PPM in channel coding. PPQ is one best quantization scheme for the wideband Gaussian source under the measure of total mean-squared error distortion reduction per information bit. Therefore, opportunistic PPM can be thought of as an explicit binning scheme, with orthogonal codes serving as both channel code and quantizers.

C. Cost-Efficient Coding and Low-Rate Quantization

Verdú's capacity per unit cost framework [13] is the natural generalization of the wideband AWGN channel to general discrete memoryless channels with an input cost constraint. The key feature of this abstraction is that the limitation is put on the input cost rather than on the number of degrees of freedom. It is shown in [13] that the capacity per unit cost (the precise definition of which is in the next section) can be computed as

$$C = \sup_X \frac{I(X;Y)}{\mathbb{E}[b(X)]} \tag{54}$$

where X and Y are the channel input and output, and $b(\cdot)$ is a function that assigns a cost to each letter in the input alphabet. For the most important case where the input alphabet contains a zero-cost letter labeled as "0," the capacity per unit cost is given by

$$C = \sup_x \frac{D(P_{Y|X=x} \| P_{Y|X=0})}{b(x)}. \tag{55}$$

Note that, in (55), the optimization is over the input alphabet as opposed to (54) where it is over the input distribution. This greatly simplifies the calculation of the capacity per unit cost for general discrete memoryless channels. Moreover, (55) can be achieved by the following generalized PPM scheme: M messages correspond to M pulse positions. The length of each pulse is N. When a specific message is chosen, the transmitter sends a pulse with each letter identically equal to x in the corresponding position and "0," otherwise. Instead of using the maximum-likelihood decoding, the decoder performs M independent binary hypothesis tests on the transmit position. Using Stein's lemma, Verdú [13] showed that (55) is indeed achievable if we choose

$$M \doteq \exp(ND(P_{Y|X=x} \| P_{Y|X=0})). \tag{56}$$

Here (and from now on), we use symbol "$\doteq$" to represent equality in the exponential scale of N.

In [13], Verdú also considered the problem of low-rate quantization as a counterpart of cost-efficient channel coding in the rate-distortion theory. This can be seen as a generalization of PPQ to arbitrary wideband sources. Classical rate-distortion theory states that the minimum number of bits that needs to be transmitted per source letter so as to reproduce the source X with average distortion not exceeding D is

$$R(D) = \inf_{P_{Y|X}} I(X;Y) \tag{57}$$

where the infimum is over all conditional probabilities $P_{Y|X}$ such that $\mathbb{E}[d(X,Y)] \leq D$. Here, the nonnegative function $d(\cdot,\cdot)$ assigns a penalty to each input-output pair. Let D_{max} be the minimum distortion that can be achieved by representing the source with a single letter:

$$D_{\max} = \mathbb{E}[d(X, v_X)] \tag{58}$$

with

$$v_X = \arg\min_y \mathbb{E}[d(X,y)]. \tag{59}$$

In the low-rate regime, we are interested in finding the level of distortion reduction from $D_{\max}$ that can be achieved by any clever coding scheme. If we consider the reward function $D_{max} - d(x,y)$, then the minimum number of bits necessary to get one reward unit is the slope of the rate-distortion function (57) at $D_{\max}$

$$R'(D_{\max}) = \lim_{D \uparrow D_{max}} \frac{R(D)}{D_{\max} - D} \tag{60}$$

$$= \inf_{P_W \ll P_X} \frac{D(P_W \| P_X)}{\mathbb{E}[d(W,v_X) - d(W,v_W)]} \tag{61}$$

where by $P_W \ll P_X$, we mean P_W is absolutely continuous over P_X. Furthermore, (61) can be achieved by the following generalized PPQ scheme. Fix an arbitrary source distribution P_W. Given M length-N source vectors, the encoder looks for one that is P_W-typical and informs the decoder its position. If we choose

$$M \doteq \exp(ND(P_W \| P_X)) \tag{62}$$

with high probability, the encoder is able to find such a source vector. The decoder decides each letter should be represented by v_W if belongs to the P_W-typical source vector or by v_X, otherwise. In this way, we are able to use $\log_e M$ nats to get a reward of $N\mathbb{E}[d(W,v_X) - d(W,v_W)]$. The quantization efficiency (61) is thus achieved.

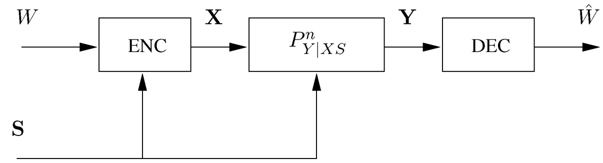

Fig. 4. Gel'fand–Pinsker channel.

V. Capacity per Unit Cost for Gel'fand–Pinsker Channels

A. Preliminaries

We have shown that opportunistic PPM achieves the minimum energy per bit for reliable communication over the wideband Costa's dirty-paper channel and that opportunistic PPM can be thought of as a nested combination of PPM and PPQ. In this section, we generalize this result to abstract Gel'fand–Pinsker channels with an input cost constraint. Speaking of generalization, Verdú's cost-efficient coding and low-rate quantization schemes are natural extensions of the basic PPM and PPQ schemes. Along this line of thinking, one is tempted to think that the nested combination of Verdú's cost-efficient coding and low-rate quantization may achieve the capacity per unit cost for general Gel'fand–Pinsker channels. The main result of this section is to show that this is indeed the case. We start with some preliminaries on Gel'fand–Pinsker channels and the capacity per unit cost.

Referring to Fig. 4, a Gel'fand–Pinsker channel is a discrete channel with input alphabet $\mathcal{X}$, output alphabet $\mathcal{Y}$, the set of states $\mathcal{S}$, and is determined by the set of conditional probabilities $P_{Y|XS}(y|x,s)$, $y \in \mathcal{Y}$, $x \in \mathcal{X}$, $s \in \mathcal{S}$ and the probability distribution $P_S(s)$ over $\mathcal{S}$. The channel is memoryless and stationary. That is, $P_{\boldsymbol{Y}|\boldsymbol{X}\boldsymbol{S}}(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{s})$ and $P_{\boldsymbol{S}}(\boldsymbol{s})$ are given by

$$P_{\boldsymbol{Y}|\boldsymbol{X}S}(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{s}) = P^n_{Y|XS}(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{s}), \quad P_{\boldsymbol{S}}(\boldsymbol{s}) = P^n_S(\boldsymbol{s}). \tag{63}$$

The states $\boldsymbol{s}$ are *noncausally* known to the encoder. The decoder, on the other hand, only knows the statistics of $\boldsymbol{s}$.

An (n, M, ν, ϵ) code is one in which the block length is equal to n; the number of messages is equal to M; each codeword $\boldsymbol{x}(\boldsymbol{s}, m)$, $m \in \mathcal{I}_M$, $\boldsymbol{s} \in \mathcal{S}^n$, satisfies the constraint

$$\sum_{i=1}^{n} b(x_i(\boldsymbol{s}, m)) \leq \nu \tag{64}$$

where $b: \mathcal{X} \to \mathcal{R}^+ \stackrel{\text{def}}{=} [0, +\infty)$ is a function that assigns a cost to each input letter; and the average (over equiprobable messages) probability of correctly decoding the message is better than $1 - \epsilon$. The following definition of capacity per unit cost is equivalent to [13, Definition 2].

Definition 4: Given $0 \leq \epsilon < 1$, a nonnegative number R is an ϵ-achievable rate per unit cost if for every $\gamma > 0$, there exists a positive integer n_0 and $\beta > 0$ such that if $n > n_0$, then an (n, M, ν, ϵ) code can be found with $\nu = n\beta$ and $\log_e M > \nu(R - \gamma)$. R is achievable per unit cost if it is ϵ-achievable per unit cost for all $0 < \epsilon < 1$, and the capacity per unit cost is the maximum achievable rate per unit cost.

B. Main Results

Consider a Gel'fand–Pinsker channel with *finite* input, output and state alphabets. Suppose there is a free input letter "0," i.e., $b(0) = 0$, in $\mathcal{X}$. Denote by $\mathcal{A}$ the set of all pairs (X_0, V) of random variables taking values from $\mathcal{X}$ and $\mathcal{S}$ such that $P_V \ll P_S$. For any $A = (X_0, V) \in \mathcal{A}$, let

$$R(A) \stackrel{\text{def}}{=} \frac{D(P_{Y_0} \| P_{Y|X=0}) - D(P_V \| P_S)}{\mathbb{E}[b(X_0)]} \tag{65}$$

where

$$P_{Y_0}(y) \stackrel{\text{def}}{=} \sum_{x} \sum_{s} P_{Y|XS}(y|x, s) P_{X_0 V}(x, s). \tag{66}$$

The following theorem is our main result on the capacity per unit cost for Gel'fand–Pinsker channels with a zero-cost input letter.

Theorem 5: Suppose there is a free letter "0" in the input alphabet $\mathcal{X}$. The capacity per unit cost of a Gel'fand–Pinker channel with finite input, output and state alphabets is

$$C = \max_{A \in \mathcal{A}} R(A). \tag{67}$$

The proof of the achievability is constructive: A generalization of the basic opportunistic PPM achieves (67). This generalization can be thought of as a nested combination of Verdú's cost-efficient coding and low-rate quantization schemes and is outlined as follows. M pulses correspond to M messages. Each pulse splits into K length-N subpulses. (This is akin to the generalized opportunistic PPM scheme for the wideband dirty-paper channel with i.i.d. non-Gaussian "dirt," see Fig. 2. However, this is also where the similarity ends.) Given a message m, the encoder looks at the channel state vectors in the K subpulse positions associated with m and choose one that is P_V-typical (Verdú's low-rate quantization). If we choose

$$K \doteq \exp(N D(P_V \| P_S)) \tag{68}$$

with high probability, the encoder is able to find such a state vector. For the subpulse position where the state vector is P_V-typical, the encoder sends a length-N codeword independently and identically chosen according to $P_{X_0|V=s}$ where s is the state realization in the corresponding position. Otherwise, it sends the free letter "0" (Verdú's cost-efficient coding). The cost of the transmission is thus $N\mathbb{E}[b(X_0)]$. The decoder performs MK independent binary hypothesis tests on each received subpulse. Note that, by (66), the received subpulse is i.i.d. as P_{Y_0} if the transmitted subpulse is nonzero and is i.i.d. as $P_{Y|X=0}$, otherwise. If we choose

$$MK \doteq \exp(N D(P_{Y_0} \| P_{Y|X=0})) \tag{69}$$

with high probability, the decoder can correctly figure out the position of the nonzero transmit subpulse. It then decides the message to be the one that corresponds to the pulse containing the nonzero subpulse. Using the above encoding/decoding procedure, the capacity per unit (67) is achieved.

The proof of the converse is based on calculus of mutual information; the details of the proof are deferred to Appendix B.

This theorem has the following simple addenda, which may be useful for calculating the capacity per unit cost for some specific Gel'fand–Pinsker channels. The proof follows from properties of divergence and is included in Appendix C for completeness.

Corollary 6: To evaluate C, one may take in (67) the maximum for pairs $A \in \mathcal{A}$ with deterministic $P_{X_0|V}$ for a given probability distribution P_V.

Note that, even with the help of Corollary 6, the computational advantage of computing the capacity per unit cost (55) of a point-to-point discrete memoryless channel no longer exists. One still has to optimize over a certain distribution, e.g., P_V, to obtain the capacity per unit cost of a Gel'fand–Pinsker channel with a zero-cost input letter.

To prove the converse part of Theorem 5, we need the following result as the starting point. This result is slightly more general than Theorem 5, in that it does not require the existence of a zero-cost input letter. Denote by $\mathcal{A}_1$ the set of all triples (U, S, X) of random variables (U is an auxiliary variable with values in an arbitrary finite set $\mathcal{U}$) with the joint distribution $P_{USX}(u, s, x)$ such that the marginal distribution of S is equal to the state distribution $P_S(s)$. To any triple $A_1 = (U, S, X) \in \mathcal{A}_1$, we assign the quadruple (U, S, X, Y) of random variables by

$$P_{USXY}(u, s, x, y) = P_{USX}(u, s, x) P_{Y|XS}(y|x, s) \tag{70}$$

for instance, $U \to (X, S) \to Y$ forms a Markov chain. Here, $P_{Y|XS}(y|x, s)$ is transition probability of the Gel'fand–Pinker channel. For any $A_1 \in \mathcal{A}_1$, let

$$R_1(A_1) \stackrel{\text{def}}{=} \frac{I(U; Y) - I(U; S)}{\mathbb{E}[b(X)]}. \tag{71}$$

Theorem 7: The capacity per unit cost of a Gel'fand–Pinsker channel with finite input, output, and state alphabets is

$$C = \max_{A_1 \in \mathcal{A}_1} R_1(A_1). \tag{72}$$

The achievable part of the above theorem is barely new. It essentially says that the capacity per unit cost of a Gel'fand–Pinsker channel can be achieved by the binning scheme [1]. The converse is a bit more involved than that in [1] because the cost constraint on the codeword must be taken into account. This issue can be resolved by introducing a time-sharing random variable. The idea of using time-sharing random variable to incorporate an input constraint into the side-information problem was due to Willems [14]. The proof of this theorem is in Appendix D. For completeness, we also have the following corollary; the proof is a simple consequence of [1, Proposition 1].

Corollary 8: To evaluate C, one may take in (72) the maximum for triples $A_1 \in \mathcal{A}_1$ with deterministic $P_{X|US}$ for a given $P_{U|S}$ and $|\mathcal{U}| \leq |\mathcal{X}| + |\mathcal{S}|$.

C. Continuous Alphabets

Theorem 5 can be extended to the case of continuous alphabets using traditional arguments. Given probability measures P_V, P_S, P_{Y_0} and $P_{Y|X=0}$, divergences are defined as [16, Ch. 2.3]

$$\begin{aligned} D(P_V\|P_S) &= \sup D(P_{\overline{V}}\|P_{\overline{S}}) \\ D(P_{Y_0}\|P_{Y|X=0}) &= \sup D(P_{\overline{Y}_0}\|P_{\overline{Y|X=0}}) \end{aligned} \tag{73}$$

where the supremos are over all finite partitions of infinite alphabets $\mathcal{X} - \{0\}$, $\mathcal{S}$ and $\mathcal{Y}$, yielding probability distributions $P_{\overline{V}}$, $P_{\overline{S}}$, $P_{\overline{Y}_0}$ and $P_{\overline{Y|X=0}}$ over finite alphabets.

Assume that $\mathcal{X} = \mathcal{S} = \mathcal{Y}$ are finite-dimension Euclidean spaces or compact subsets thereof and that the density functions f_S and $f_{Y|XS}$ satisfy certain regularity conditions including being bounded and continuous over their domain. The cost function is assumed to be continuous. Under these assumptions, the divergences on the left-hand sides of (73) can be rewritten as

$$D(P_V\|P_S) = \int f_V(s) \log_e \frac{f_V(s)}{f_S(s)} ds \tag{74}$$

$$D(P_{Y_0}\|P_{Y|X=0}) = \int f_{Y_0}(y) \log_e \frac{f_{Y_0}(y)}{\int f_{Y|XS}(y|0,s) f_S(s) ds} dy \tag{75}$$

where

$$f_{Y_0}(y) = \int\int f_{Y|XS}(y|x,s) f_{X_0 V}(x,s) dx\, ds. \tag{76}$$

For any $\epsilon > 0$, select a finite partition of alphabets such that

$$D(P_V\|P_S) - \epsilon < D(P_{\overline{V}}\|P_{\overline{S}}) \leq D(P_V\|P_S) \tag{77}$$

$$D(P_{Y_0}\|P_{Y|X=0}) - \epsilon < D(P_{\overline{Y}_0}\|P_{\overline{Y|X=0}}) \leq D(P_{Y_0}\|P_{Y|X=0}) \tag{78}$$

and

$$\left|\mathbb{E}[b(\overline{X}_0)] - \mathbb{E}[b(X_0)]\right| < \epsilon. \tag{79}$$

The existence of such a partition is guaranteed by our regularity assumptions. Let

$$R_\epsilon(A) \overset{\text{def}}{=} \frac{D(P_{\overline{Y}_0}\|P_{\overline{Y|X=0}}) - D(P_{\overline{V}}\|P_{\overline{S}})}{\mathbb{E}[b(\overline{X}_0)]}. \tag{80}$$

By (77)–(79), the capacity per unit cost of a continuous Gel'fand–Pinsker channel with a zero-cost input letter is

$$C \overset{\text{def}}{=} \lim_{\epsilon \downarrow 0} \max_{A \in \mathcal{A}} R_\epsilon(A) = \sup_{A \in \mathcal{A}} R(A). \tag{81}$$

For compact $\mathcal{A}$, "sup" can be replaced by "max," and Theorem 5 applies for continuous alphabets as well.

A few examples are now in order.

Example 1 (Costa's Dirty-Paper Channel): Consider the discrete-time Costa's dirty-paper channel

$$Y_i = X_i + S_i + N_i \tag{82}$$

where the channel input X_i can be an arbitrary real number, and $\{S_i\}$, $\{N_i\}$ are independent i.i.d. Gaussian processes with zero mean and variance $N_s/2$ and $N_0/2$, respectively. S_i, $i = 1, 2, \ldots$, are assumed to be known noncausally to the input side of the channel. The cost is on the power of the input letter, i.e., $b(x) = x^2$. Let $X_0 = x_0$ almost surely for some $x_0 > 0$ and V be distributed as $\mathcal{N}(x_0 N_s/N_0, N_s/2)$. The divergence between two Gaussian distributions is given by

$$\begin{aligned} &D(\mathcal{N}(m_1, \sigma_1^2)\|\mathcal{N}(m_0, \sigma_0^2)) \\ &\quad = \log_e \frac{\sigma_0}{\sigma_1} + \frac{1}{2}\left(\frac{\sigma_1^2}{\sigma_0^2} - 1\right) + \frac{(m_1 - m_0)^2}{2\sigma_0^2}. \end{aligned} \tag{83}$$

By Theorem 5, we have (84) shown at the bottom of the page, which is an achievable capacity per unit cost for Costa's dirty-paper channel. This result is equivalent to the $\frac{\varepsilon_b}{N_0}|_{\min} = \log_e 2 = -1.59$ dB result for the wideband Costa's dirty-paper channel. We thus conclude that the above choice of (X_0, V) is an optimal one.

Example 2 (Estimation-Theoretic Lower Bound): Consider Gel'fand–Pinsker channels with $\mathcal{X} = \mathcal{S} = \mathcal{Y}$ being the whole real line and the cost function $b(x) = x^2$. Let the family of random variables $(X_0^{(\theta)}, V^{(\theta)})$ be such that $X_0^{(\theta)} = \theta$ almost surely and $P_V^{(\theta)}$ converges to P_S in distribution in the limit as

$$\frac{D\left(\mathcal{N}\left(\frac{N_s+N_0}{N_0}x_0, \frac{N_s+N_0}{2}\right)\|\mathcal{N}\left(0, \frac{N_s+N_0}{2}\right)\right) - D\left(\mathcal{N}\left(\frac{N_s}{N_0}x_0, \frac{N_s}{2}\right)\|\mathcal{N}\left(0, \frac{N_s}{2}\right)\right)}{x_0^2} = \frac{1}{N_0} \tag{84}$$

$\theta \to 0$. The capacity per unit cost of such channels can be bounded from below as

$$C \geq \lim_{\theta \to 0} \frac{D\left(P_{Y_0}^{(\theta)} \| P_{Y|X=0}\right) - D\left(P_V^{(\theta)} \| P_S\right)}{\theta^2} \tag{85}$$

where

$$P_{Y_0}^{(\theta)}(y) = \int P_{Y|XS}(y|\theta, s)\, dP_V^{(\theta)}(s). \tag{86}$$

Under our regularity assumptions, the following asymptotic result on divergence is known [16, Ch. 2.6]:

$$\lim_{\theta \to 0} \frac{D\left(P_V^{(\theta)} \| P_S\right)}{\theta^2} = \frac{1}{2} J_0\left(P_V^{(\theta)}\right) \tag{87}$$

where $J_0\left(P_V^{(\theta)}\right)$ is the Fisher information for estimating θ from $V^{(\theta)}$, i.e.

$$J_\theta\left(P_V^{(\theta)}\right) \stackrel{\text{def}}{=} \mathbb{E}\left[\left(\frac{\partial}{\partial \theta} \log_e f_V(V;\theta)\right)^2\right] \tag{88}$$

evaluated at $\theta = 0$, and $f_V(v;\theta)$ is the density function of $V^{(\theta)}$. Similarly, $Y_0^{(\theta)}$ converges in distribution to Y conditioned on $X = 0$. So we have

$$\lim_{\theta \to 0} \frac{D\left(P_{Y_0}^{(\theta)} \| P_{Y|X=0}\right)}{\theta^2} = \frac{1}{2} J_0\left(P_{Y_0}^{(\theta)}\right). \tag{89}$$

Substituting (87) and (89) into (85), we obtain the following estimation-theoretic lower bound for the capacity per unit cost of Gel'fand–Pinsker channels with real alphabets and quadratic cost function:

$$C \geq \frac{1}{2}\left(J_0\left(P_{Y_0}^{(\theta)}\right) - J_0\left(P_V^{(\theta)}\right)\right). \tag{90}$$

Example 3 (Dirty-Paper Channels With i.i.d. Non-Gaussian "Dirt"): We now apply the estimation-theoretic lower bound (90) to dirty-paper channels with i.i.d. non-Gaussian "dirt." The channel model is the same as (82), except that the "dirt" S_i, $i = 1, 2, \ldots$, are i.i.d. but non-Gaussian. Consider the choice $V^{(\theta)} = \alpha\theta + S$ where α is an optimization parameter. We obtain from (90) that the capacity per unit cost

$$C \geq \frac{1}{2}\left((\alpha+1)^2 J(S+N) - \alpha^2 J(S)\right) \tag{91}$$

where $J(X)$ is the Fisher information of random variable X with respect to a translation parameter, i.e.,

$$J(X) \stackrel{\text{def}}{=} \mathbb{E}\left[\left(\frac{d}{dX} \log_e f_X(X)\right)^2\right]. \tag{92}$$

Here, we have dropped the subscripts and used generic S and N to represent the interference and noise, respectively. Choosing α to maximize the right-hand side of (92), we obtain

$$C \geq \frac{J(S)J(S+N)}{2(J(S) - J(S+N))} \tag{93}$$

where the optimal choice of α is $\frac{J(S+N)}{J(S)-J(S+N)}$. It is well known that

$$J(X) \geq \frac{1}{\mathrm{Var}(X)} \tag{94}$$

where $\mathrm{Var}(X)$ denotes the variance of X, and the equality holds if and only X is Gaussian. In light of Cohen and Lapidoth's result [9] (and our discussion in Section III-B), the capacity per unit cost of dirty-paper channels with i.i.d. non-Gaussian "dirt" is

$$C = \frac{1}{N_0} = \frac{1}{2\mathrm{Var}(N)} = \frac{1}{2} J(N) \tag{95}$$

where the last equality follows from the Gaussianity of N. Substituting (95) into (93), we obtain

$$\frac{1}{J(S+N)} \geq \frac{1}{J(S)} + \frac{1}{J(N)} \tag{96}$$

which is the special case of the celebrated Fisher information inequality (FII) [17] with one of the participating random variables fixed to be Gaussian. It is known that FII holds with equality if and only if both participating random variables are Gaussian. Therefore, the estimation-theoretic lower bound with the proposed $V^{(\theta)} = \alpha\theta + S$ is generally not tight for dirty-paper channels with i.i.d. non-Gaussian "dirt". Note that this does not exclude the existence of other choices of $V^{(\theta)}$ such that the estimation-theoretic lower bound (90) is tight for this problem. Unfortunately, we have not been able to find an explicit (X_0, V) which is optimal for dirty-paper channels with i.i.d. non-Gaussian "dirt." We put this implication into the following lemma.

Lemma 9: Suppose N is Gaussian with zero mean, S has zero mean and finite second moment, and N and S are statistically independent. Then we have

$$\sup_{P_{X_0 V}} \frac{D(X_0 + V + N \| S + N) - D(V \| S)}{\mathbb{E}[X_0^2]} = \frac{1}{\mathrm{Var}(N)} \tag{97}$$

where the supreme is over all pairs (X_0, V) of random variables such that (X_0, V) is independent of N and the marginal distributions satisfy $P_V \ll P_S$.

VI. Low-Rate Quantization for the Wyner–Ziv Problem

In this section, we extend PPQ to low-rate source coding with side information. Referring to Fig. 5, the traditional rate-distortion problem with side information at the decoder was considered by Wyner and Ziv [18]. The main objective of this section is

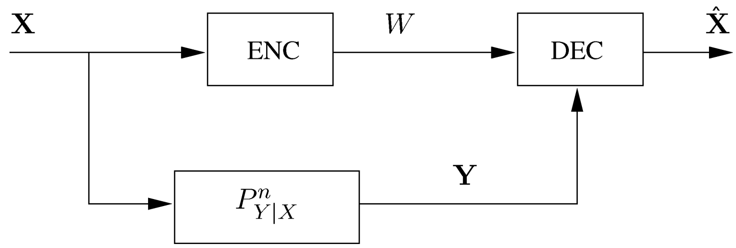

Fig. 5. Wyner–Ziv source coding.

to show that opportunistic version of PPQ achieves the low-rate slope of the Wyner–Ziv rate-distortion function.

A. The Gaussian Case

We first treat the special case where the source and side-information letters (X_i, Y_i), $i = 1, 2, \ldots$, are i.i.d. as $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$ where

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{pmatrix}. \tag{98}$$

Wyner [19] showed that if the distortion measure is a squared one, i.e., $d(x, y) = (x - y)^2$, the rate-distortion function with side information Y_i, $i = 1, 2, \ldots$, available *only* at the decoder is

$$R(D) = \frac{1}{2}\log_e \frac{(1-\rho^2)\sigma_X^2}{D}, \quad 0 \le D \le (1-\rho^2)\sigma_X^2. \tag{99}$$

Thus, the low-rate slope is equal to

$$R'(D_{\max}) = \lim_{D \uparrow D_{\max}} \frac{R(D)}{D_{\max} - D} = \frac{1}{2(1-\rho^2)\sigma_X^2}. \tag{100}$$

We note that both (99) and (100) are the same as those obtained as if the side information is also available at the encoder.

Opportunistic PPQ: Without loss of generality, assume X_i and Y_i are positively correlated, i.e., $\rho > 0$. Consider the following low-rate quantization scheme: Given a string of MK source letters $X_1, \ldots, X_{MK}$, randomly partition them into M bins. The encoder picks the one (among all MK source letters) that is the *largest* and informs the decoder the bin number it belongs to; this needs a total of $\log_e M$ nats for description. The decoder looks at the side-information letters Y_i, $i = 1, \ldots, MK$, in that bin and picks (within that bin) one that is the largest (which we shall denote by Y_{i^*}). The reconstruction uses the following rule:

$$\hat{X}_i = \begin{cases} \sigma_X\sqrt{2\log_e(MK)}, & i = i^* \\ \frac{\rho\sigma_X}{\sigma_Y} Y_i, & \text{otherwise.} \end{cases} \tag{101}$$

Performance Analysis: By symmetry, we may assume that X_1 is the largest among all X_i, $i = 1, \ldots, MK$, and that it belongs to the first bin. We first show that, with high probability, Y_1 is the largest among all Y_i's in the first bin (so the decoder can correctly pick out X_1 using the informed bin number and the side information). By our assumption, X_1 is the largest among MK independent random variables identically distributed as $\mathcal{N}(0, \sigma_X^2)$. By Lemma 1, $X_1 - \sigma_X\sqrt{2\log_e(MK)}$ converges to 0 in probability in the limit as $MK \to \infty$. Since X_1 and Y_1 are jointly distributed as $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$, we may write $Y_1 = (\rho\sigma_Y/\sigma_X)X_1 + N_1$ where N_1 is distributed as $\mathcal{N}(0, (1-\rho^2)\sigma_Y^2)$ and is independent of X_1. It follows that, with high probability, Y_1 will be close to $\rho\sigma_Y\sqrt{2\log_e(MK)}$ for large MK. If we choose

$$\log_e K = \frac{\rho^2}{1-\rho^2}\log_e M \tag{102}$$

we will have $\rho\sigma_Y\sqrt{2\log_e(MK)} = \sigma_Y\sqrt{2\log_e K}$, which is the opportunistic gain by choosing the largest among K random variables i.i.d. as $\mathcal{N}(0, \sigma_Y^2)$. Therefore, with high probability, Y_1 is the largest among all Y_i in the first bin. Now that the decoder has successfully figured out the largest one among all X_i's, by Lemma 1, the total distortion reduction (relative to no message being sent to the decoder) is

$$\mathbb{E}\left[\left(X_1 - \frac{\rho\sigma_X}{\sigma_Y}Y_1\right)^2\right] - \mathbb{E}\left[\left(X_1 - \sigma_X\sqrt{2\log_e(MK)}\right)^2\right] = 2(1-\rho^2)\sigma_X^2\log_e M + o(\log_e M) \tag{103}$$

in the limit as $M \to \infty$. Thus, the bit efficiency

$$\frac{\log_e M}{2(1-\rho^2)\sigma_X^2\log_e M} = \frac{1}{2(1-\rho^2)\sigma_X^2} \tag{104}$$

is achieved. This is the highest efficiency possible that can be achieved by any low-rate quantizer. Similar to opportunistic PPM, the success of the above quantization scheme critically hinges on the fact that the source and the side information are jointly Gaussian. We call this scheme opportunistic PPQ, as a counterpart of opportunistic PPM in the rate-distortion theory.

Theorem 10: Opportunistic PPQ achieves the low-rate slope of the Gaussian Wyner–Ziv rate-distortion function.

B. The General Case

In this section, we extend the basic opportunistic PPQ scheme to the general case where the source and the side-information letters (X_i, Y_i), $i = 1, 2, \ldots$, are i.i.d. as P_{XY} and the distortion measure $d(x, y)$ is an abstract one. (In case of continuous alphabets, P_{XY} and $d(x, y)$ need to satisfy certain regularity conditions.) It is shown in [18] that the traditional rate-distortion function with decoder side information only is

$$R(D) = \inf_{P_{W|X}} \inf_{f} I(X;W) - I(Y;W) \tag{105}$$

where the minimizations are over all conditional distributions $P_{W|X}$ such that $W \to X \to Y$ forms a Markov chain and all functions $f : \mathcal{W} \times \mathcal{Y} \to \mathcal{X}$ such that

$$\mathbb{E}[d(X, f(W, Y))] \le D \tag{106}$$

and the outer infimum is over all functions $f : \mathcal{W} \times \mathcal{Y} \to \mathcal{X}$. The main result of this section is the following theorem.

Theorem 11: The low-rate slope of the general Wyner–Ziv rate-distortion function is

$$R'(D_{max}) = \lim_{D \uparrow D_{\max}} \frac{R(D)}{D_{\max} - D} = \inf_{P_V \ll P_X} \frac{D(P_V \| P_X) - D(P_{Y_0} \| P_Y)}{\mathbb{E}[d(V, \hat{x}(Y_0)) - d(V, \hat{v}(Y_0))]} \quad (107)$$

where

$$P_{Y_0} = \int P_{Y|X}(y|x) dP_V(x) \quad (108)$$

and

$$\hat{x} = \arg \min_{f:\mathcal{Y}\to\mathcal{X}} \mathbb{E}[d(X, f(Y))]$$
$$\hat{v} = \arg \min_{f:\mathcal{Y}\to\mathcal{X}} \mathbb{E}[d(V, f(Y_0))]. \quad (109)$$

The low-rate slope (107) can be achieved by the following generalized opportunistic PPQ scheme: Given a string of MKN source letters, the encoder randomly partitions them into M bins with each bin containing K length-N source vectors. The encoder looks (among all MK source vectors) for one that is P_V-typical. If we choose

$$MK \doteq \exp(ND(P_V \| P_X)) \quad (110)$$

with high probability, the encoder is able to find such a source vector. It then informs the decoder the bin number to which the P_V-typical source vector belongs; this needs a total of $\log_e M$ nats for description. Given the informed bin number, the decoder chooses (among all length-N side-information vectors in that bin) one that is P_{Y_0}-typical. It then claims that the source vector in the corresponding position is the one that is P_V-typical. By (108), if the source vector is P_V-typical, with high probability, the corresponding side-information vector will be P_{Y_0}-typical. So if we choose

$$K \doteq \exp(ND(P_{Y_0} \| P_Y)) \quad (111)$$

with high probability, the decoder can correctly figure out the position of the P_V-typical source vector. Finally, the decoder reconstructs letters x_i in the P_V-typical source vector by $\hat{v}(y_i)$ and by $\hat{x}(y_i)$, otherwise. The total distortion reduction (relative to no message being sent to the decoder) of this scheme is

$$N\mathbb{E}[d(V, \hat{x}(Y_0)) - d(V, \hat{v}(Y_0))]. \quad (112)$$

By (110)–(112), we conclude that the bit efficiency (107) is achieved by the above generalized opportunistic PPQ scheme.

To establish the converse part, we first expand the mutual information terms on the right-hand side of (105) as

$$I(X; W) = \int D(P_{X|W=w} \| P_X) dP_W(w) \quad (113)$$
$$I(Y; W) = \int D(P_{Y|W=w} \| P_Y) dP_W(w). \quad (114)$$

It follows by (115)–(118) as shown at the bottom of the page, where $\hat{x}(y)$ is defined in (109). Let (V, Y_0) be a new pair of random variables such that $P_{VY_0}(x, y) = P_{XY|W}(x, y|w^*)$. The marginal distributions satisfy

$$P_V(x) = \int P_{VY_0}(x, y) dy = \int P_{XY|W}(x, y|w^*) dy = P_{X|W}(x|w^*) \quad (119)$$
$$P_{Y_0}(y) = \int P_{Y|X}(y|x) dP_V(x) = \int P_{Y|X}(y|x) dP_{X|W}(x|w^*) = P_{Y|W=w^*}(y). \quad (120)$$

In (120), the first equality follows from (108), the second equality follows from (119), and the last equality follows from the Markov chain $W \to X \to Y$. Furthermore

$$\int d(x, \hat{x}(y)) dP_{XY|W}(x, y|w^*) = \int d(x, \hat{x}(y)) dP_{VY_0}(x, y) = \mathbb{E}[d(V, \hat{x}(Y_0))]. \quad (121)$$

Given an arbitrary $P_{W|X}$, to minimize

$$\mathbb{E}[d(X, f(W, Y))] = \int \left(\int d(x, f(w, y)) dP_{XY|W}(x, y|w) \right) dP_W(w) \quad (122)$$

it is necessary for f to minimize

$$\int d(x, f(w, y)) dP_{XY|W}(x, y|w)$$

for all $w \in \mathcal{W}$. In particular, f needs to minimize

$$\int d(x, f(w^*, y)) dP_{XY|W}(x, y|w^*) = \int d(x, f(w^*, y)) dP_{VY_0}(x, y) = \mathbb{E}[d(V, \hat{v}(Y_0))]. \quad (123)$$

$$R'(D_{\max}) \geq \inf_{P_{W|X}} \inf_f \frac{\int \left(D(P_{X|W=w} \| P_X) - D(P_{Y|W=w} \| P_Y) \right) dP_W(w)}{\int d(x, \hat{x}(y)) dP_{XY}(x, y) - \int d(x, f(w, y)) dP_{WXY}(w, x, y)} \quad (115)$$
$$= \inf_{P_{W|X}} \inf_f \frac{\int \left(D(P_{X|W=w} \| P_X) - D(P_{Y|W=w} \| P_Y) \right) dP_W(w)}{\int \int \left(d(x, \hat{x}(y)) - d(x, f(w, y)) \right) dP_{XY|W}(x, y|w) dP_W(w)} \quad (116)$$
$$\geq \inf_{P_{W|X}} \inf_f \inf_w \frac{D(P_{X|W=w} \| P_X) - D(P_{Y|W=w} \| P_Y)}{\int \left(d(x, \hat{x}(y)) - d(x, f(w, y)) \right) dP_{XY|W}(x, y|w)} \quad (117)$$
$$= \inf_{P_{W|X}} \inf_f \frac{D(P_{X|W=w^*} \| P_X) - D(P_{Y|W=w^*} \| P_Y)}{\int \left(d(x, \hat{x}(y)) - d(x, f(w^*, y)) \right) dP_{XY|W}(x, y|w^*)} \quad (118)$$

Substituting (119)–(121) and (123) into (118), we obtain the desired reverse inequality

$$R'(D_{\max}) \geq \inf_{P_V \ll P_X} \frac{D(P_V\|P_X) - D(P_{Y_0}\|P_Y)}{\mathbb{E}[d(V,\hat{x}(Y_0)) - d(V,\hat{v}(Y_0))]}. \quad (124)$$

VII. Concluding Remarks

We propose a simple orthogonal signaling scheme that achieves the capacity per unit cost of Gel'fand–Pinsker channels with a zero cost input letter. The scheme is by nature opportunistic and can be interpreted as *structured* binning in the wideband regime. As a special case, we *explicitly* construct an opportunistic PPM scheme which we show to achieve both the capacity and the reliability function of the wideband Costa's dirty-paper channel. The source-coding counterparts (except for the source-coding exponents) of the above results have also been found. These new results exhibit some interesting connections to estimation theory.

What has been exclusively considered in this paper is a hypothetical communication scenario where bandwidth is not a commodity (and hence can be abused without incurring any penalty). As a future direction, it would be interesting to explore simple signaling schemes which are not only capacity achieving but also spectrally efficient (in the sense of [20]). Such schemes will be useful for the design of precoding algorithms in practical wideband communication systems. It would also be interesting to evaluate the performance of the proposed scheme in the wideband wireless downlink where the base station has no access to fading realizations.

Appendix A
Proof of Theorem 3

We now derive an exponential upper bound on the error probability of opportunistic PPM for the wideband Costa's dirty-paper channel. The derivation is quite long so we divide it into several steps.

Step 1. Assume $m = 1$, and all the probabilities below will be tacitly understood to be conditioned on that event. The error probability of opportunistic PPM can be bounded from above as

$$\begin{aligned} P_e &= \mathbb{P}(\hat{m} \neq 1) && (125)\\ &= \mathbb{P}(\hat{m} \neq 1, s_{k^*} \leq \xi) + \mathbb{P}(\hat{m} \neq 1, s_{k^*} \geq \xi) && (126)\\ &\leq \mathbb{P}(s_{k^*} \leq \xi) + \mathbb{P}(\hat{m} \neq 1 | s_{k^*} \geq \xi) && (127)\end{aligned}$$

where the threshold of the opportunistic gain

$$\xi = \sqrt{N_s(\log_e K - 2\log_e \log_e K)}. \quad (128)$$

The two probability terms on the right-hand side of (127) represent the probability of encoding and decoding error, respectively.

Step 2. The probability of encoding error can be written as

$$\begin{aligned}\mathbb{P}(s_{k^*} \leq \xi) &= \mathbb{P}\left(\max_{1\leq k\leq K} s_k \leq \xi\right)\\ &= \left(1 - Q\left(\sqrt{2(\log_e K - 2\log_e \log_e K)}\right)\right)^K. \quad (129)\end{aligned}$$

The exponential decay rate with respect to $\log_e K$ can be estimated as follows:

$$\begin{aligned} -\frac{\log_e \mathbb{P}(s_{k^*} \leq \xi)}{\log_e K} &= -\frac{K\log_e\left(1 - Q\left(\sqrt{2(\log_e K - 2\log_e\log_e K)}\right)\right)}{\log_e K} && (130)\\ &\sim \frac{KQ\left(\sqrt{2(\log_e K - 2\log_e \log_e K)}\right)}{\log_e K} && (131)\\ &\sim \frac{K\exp(-\log_e K + 2\log_e\log_e K)}{\log_e K\sqrt{4\pi(\log_e K - 2\log_e\log_e K)}} && (132)\\ &= \frac{\log_e K}{\sqrt{4\pi(\log_e K - 2\log_e\log_e K)}} && (133)\end{aligned}$$

in the limit as $K \to \infty$. Here, (131) follows from the fact that $\log_e(1+x) \sim x$ in the limit as $x \to 0$, and (132) follows from the fact that $Q(x) \sim \frac{e^{-x^2/2}}{x\sqrt{2\pi}}$ in the limit as $x \to \infty$. Note that the right-hand side of (133) tends to infinity in the limit as $K \to \infty$. We conclude that the probability of encoding error $\mathbb{P}(s_{k^*} \leq \xi)$ decays *superexponentially* with $\log_e K$.

Step 3. The conditional probability of decoding error can be bounded from above as

$$\begin{aligned}&\mathbb{P}(\hat{m} \neq 1 | s_{k^*} = s)\\ &= \int_{-\infty}^{\infty}\left(1 - \prod_{i=2}^{M}\prod_{j=1}^{K}\mathbb{P}\left(r_{(i-1)K+j} < r\right)\right) f_{r_{k^*}|s_{k^*}}(r|s)dr && (134)\\ &= \int_{-\infty}^{\infty}\left(1 - \left(1 - Q\left(\frac{r}{\sqrt{(N_s+N_0)/2}}\right)\right)^{(M-1)K}\right) f_{r_{k^*}|s_{k^*}}(r|s)dr && (135)\\ &\leq \int_{-\infty}^{\gamma} f_{r_{k^*}|s_{k^*}}(r|s)dr\\ &\quad + \int_{\gamma}^{\infty} MK\,Q\left(\frac{r}{\sqrt{(N_s+N_0)/2}}\right) f_{r_{k^*}|s_{k^*}}(r|s)dr && (136)\\ &\leq \int_{-\infty}^{\gamma} f_{r_{k^*}|s_{k^*}}(r|s)dr\\ &\quad + \int_{\gamma}^{\infty} MK\exp\left(-\frac{r^2}{N_s+N_0}\right) f_{r_{k^*}|s_{k^*}}(r|s)dr && (137)\end{aligned}$$

where the threshold of decoding

$$\gamma = \sqrt{(N_s+N_0)\log_e(MK)}. \quad (138)$$

Here, (136) follows from the well-known inequality $1-(1-x)^n \leq \min\{1, nx\}$ for $0 \leq x \leq 1$ and positive integer n. Conditioned on $s_{k^*} = s$, $r_{k^*} = \sqrt{\varepsilon_s} + s_{k^*} + n_{k^*}$ is distributed as $\mathcal{N}\left(\sqrt{\varepsilon_s} + s, N_0/2\right)$. So the two integrals on the right-hand side of (137) can be bounded from above as shown in (139)–(142) at the bottom of the following page. Substituting (139) and (142) into (137), we obtain (143) shown at the bottom of the following page. Next we get (144)–(145) shown at the bottom of the following page, where (144) is due to the fact that larger opportunistic gain s_{k^*} adds to the signaling strength and hence can only help decoding. Note from (128) that

$$\xi = \sqrt{(N_s - o(1)) \log_e K}$$

in the limit as $K \to \infty$. Choosing K to minimize the right-hand side of (145) (ignoring the o(1) terms), we obtain

$$\mathbb{P}(\hat{m} \neq 1 | s_{k^*} \geq \xi) < 2 \exp\left(-\left(E\left(\frac{\varepsilon_b}{N_0}\right) - o(1)\right) \log_e M\right) \tag{146}$$

in the limit as $K \to \infty$. Here, $E(x)$ is the reliability function of the wideband AWGN channel

$$E(x) = \begin{cases} \left(\sqrt{x} - \sqrt{\log_e 2}\right)^2, & \log_e 2 < x \leq 4 \log_e 2 \\ \frac{1}{2}x - \log_e 2, & x \geq 4 \log_e 2. \end{cases} \tag{147}$$

The optimum choice of K is given by

$$\log_e K = \begin{cases} \frac{N_s}{N_0} \log_e M, & \log_e 2 < \frac{\varepsilon_b}{N_0} \leq 4 \log_e 2 \\ \left(\frac{1}{4 \log_e 2} \frac{\varepsilon_b}{N_0}\right) \frac{N_s}{N_0} \log_e M, & \frac{\varepsilon_b}{N_0} \geq 4 \log_e 2. \end{cases} \tag{148}$$

Step 4. By (148), $M \to \infty$ implies $K \to \infty$. Combining Steps 1)–3), we obtain

$$P_e < (2 + o(1)) \exp\left(-\left(E\left(\frac{\varepsilon_b}{N_0}\right) - o(1)\right) \log_e M\right) \tag{149}$$

in the limit as $M \to \infty$. This completes the proof of Theorem 3.

APPENDIX B
PROOF OF THEOREM 5

A. *The Achievability*

To prove the achievability, we will explicitly construct an (n, M, ν, ϵ) code such that

$$\log_e M > \nu(R(A) - \gamma) \tag{150}$$

for any given $A \in \mathcal{A}$, $\gamma > 0$, $0 < \epsilon < 1$ and $n \geq n_0$ for some positive integer n_0. The proof is rather long so we divide it into several steps.

Step 1. By Corollary 6, one may assume that $P_{X_0|V}$ is deterministic. That is, $P_{X_0|V}$ takes value 0 or 1. So we suppose that there is a mapping $x_0 : \mathcal{V} \to \mathcal{X}$ such that $x_0 = x_0(v)$ if and only if $P_{X_0|V}(x_0|v) = 1$.

Step 2. In the sequel, we will need the notion of typical sequences. The following definition and results are gathered from [15, Ch. 1.2].

Definition 12: Denote by $N(a|\boldsymbol{x})$ the number of occurrences of $a \in \mathcal{X}$ in $\boldsymbol{x} \in \mathcal{X}^n$. For any distribution P_X on $\mathcal{X}$, a sequence $\boldsymbol{x} \in \mathcal{X}^n$ is called P_X-typical with constant δ if

$$\left|\frac{1}{n} N(a|\boldsymbol{x}) - P_X(a)\right| \leq \delta, \quad \forall\, a \in \mathcal{X} \tag{151}$$

and, in addition, no $a \in \mathcal{X}$ with $P_X(a) = 0$ occurs in $\boldsymbol{x}$. The set of such sequences will be denoted by $T_X^n(\delta)$.

Lemma 13: If $\boldsymbol{x} \in T_X^n(\delta)$ and $\boldsymbol{y} \in T_{Y|X}^n(\delta'|\boldsymbol{x})$, then $(\boldsymbol{x}, \boldsymbol{y}) \in T_{XY}^n(\delta + \delta')$. Consequently, $\boldsymbol{y} \in T_Y^n(\delta'')$ for $\delta'' = (\delta + \delta')\,|\mathcal{X}|$.

$$\int_{-\infty}^{\gamma} f_{r_{k^*}|s_{k^*}}(r|s) dr = Q\left(\frac{\sqrt{\varepsilon_s} + s - \gamma}{\sqrt{N_0/2}}\right) < \begin{cases} 1, & s \leq \gamma - \sqrt{\varepsilon_s} \\ \exp\left(-\frac{(\sqrt{\varepsilon_s} + s - \gamma)^2}{N_0}\right), & s \geq \gamma - \sqrt{\varepsilon_s}, \end{cases} \tag{139}$$

$$\begin{aligned} &\int_{\gamma}^{\infty} MK \exp\left(-\frac{r^2}{N_s + N_0}\right) f_{r_{k^*}|s_{k^*}}(r|s) dr \\ &= \int_{\gamma}^{\infty} \exp\left(\frac{\gamma^2 - r^2}{N_s + N_0}\right) \frac{1}{\sqrt{\pi N_0}} \exp\left(-\frac{\left(r - \sqrt{\varepsilon_s} - s\right)^2}{N_0}\right) dr & (140) \\ &= \sqrt{\frac{N_s + N_0}{N_s + 2N_0}} \exp\left(\frac{\gamma^2}{N_s + N_0} - \frac{\left(\sqrt{\varepsilon_s} + s\right)^2}{N_s + 2N_0}\right) Q\left(\frac{\gamma - \frac{N_s + N_0}{N_s + 2N_0}\left(\sqrt{\varepsilon_s} + s\right)}{\sqrt{\frac{(N_s + N_0)N_0}{2(N_s + 2N_0)}}}\right) & (141) \\ &< \begin{cases} \exp\left(-\frac{(\sqrt{\varepsilon_s} + s - \gamma)^2}{N_0}\right), & s \leq \frac{N_s + 2N_0}{N_s + N_0}\gamma - \sqrt{\varepsilon_s} \\ \exp\left(-\frac{(\sqrt{\varepsilon_s} + s)^2}{N_s + 2N_0} + \frac{\gamma^2}{N_s + N_0}\right), & s \geq \frac{N_s + 2N_0}{N_s + N_0}\gamma - \sqrt{\varepsilon_s}. \end{cases} & (142) \end{aligned}$$

$$\mathbb{P}(\hat{m} \neq 1 | s_{k^*} = s) < \begin{cases} 2 \exp\left(-\frac{(\sqrt{\varepsilon_s} + s - \gamma)^2}{N_0}\right), & \gamma - \sqrt{\varepsilon_s} \leq s \leq \frac{N_s + 2N_0}{N_s + N_0}\gamma - \sqrt{\varepsilon_s} \\ 2 \exp\left(-\frac{(\sqrt{\varepsilon_s} + s)^2}{N_s + 2N_0} + \frac{\gamma^2}{N_s + N_0}\right), & s \geq \frac{N_s + 2N_0}{N_s + N_0}\gamma - \sqrt{\varepsilon_s}. \end{cases} \tag{143}$$

$$\begin{aligned} &\mathbb{P}(\hat{m} \neq 1 | s_{k^*} \geq \xi) \leq \mathbb{P}(\hat{m} \neq 1 | s_{k^*} = \xi) & (144) \\ &\quad < \begin{cases} 2 \exp\left(-\frac{(\sqrt{\varepsilon_s} + \xi - \gamma)^2}{N_0}\right), & \gamma - \xi \leq \sqrt{\varepsilon_s} \leq \frac{N_s + 2N_0}{N_s + N_0}\gamma - \xi \\ 2 \exp\left(-\frac{(\sqrt{\varepsilon_s} + \xi)^2}{N_s + 2N_0} + \frac{\gamma^2}{N_s + N_0}\right), & \sqrt{\varepsilon_s} \geq \frac{N_s + 2N_0}{N_s + N_0}\gamma - \xi \end{cases} & (145) \end{aligned}$$

Lemma 14: Suppose P_X and $P_{\widetilde{X}}$ are probability distributions on $\mathcal{X}$.

1) For any $\delta > 0$, $P_X^n(T_X^n(\delta)) \to 1$ as $n \to \infty$.
2) Fix $\eta > 0$. Then for sufficiently small $\delta > 0$ and sufficiently large n, we have

$$\exp(-n(D(P_X\|P_{\widetilde{X}}) + \eta)) \leq P_{\widetilde{X}}^n(T_X^n(\delta)) \leq \exp(-n(D(P_X\|P_{\widetilde{X}}) - \eta)). \quad (152)$$

Step 3. Given $A \in \mathcal{A}$ and $\gamma > 0$, let

$$M = \exp(N(D(P_{Y_0}\|P_{Y|X=0}) - D(P_V\|P_S) - \eta)) \quad (153)$$
$$K = \exp(N(D(P_V\|P_S) + 2\eta)) \quad (154)$$

for some sufficiently large N and

$$\eta = \frac{(\mathbb{E}[b(X_0)])^2}{R(A)}\gamma^2.$$

Associate the message $m \in \mathcal{I}_M$ with K signal matrices of size $N \times MK$:

$$X(m,k) = \begin{pmatrix} 0 & \cdots & x_{1,(m-1)K+k} & \cdots & 0 \\ 0 & \cdots & x_{2,(m-1)K+k} & \cdots & 0 \\ \cdot & & \cdot & & \cdot \\ \cdot & & \cdot & & \cdot \\ 0 & \cdots & x_{N,(m-1)K+k} & \cdots & 0 \end{pmatrix}, \quad k \in \mathcal{I}_K. \quad (155)$$

Here, each column of the matrix represents a subpulse, and the only nonzero transmit subpulse in $X(m,k)$ is the $((m-1)K+k)$th column. Organize the channel states accordingly as $S = (\boldsymbol{s}_1^t, \ldots, \boldsymbol{s}_{MK}^t)$. The encoder observes the columns of S and chooses among them one that is P_V-typical with a sufficiently small constant δ. If such a vector does not exist, the encoder claims an encoding failure and sends an all-zero codeword. Otherwise, suppose $\boldsymbol{s}_{(m-1)K+k^*}$ is one that is P_V-typical. The encoder chooses $X(m,k^*)$ in which the nonzero column is

$$\boldsymbol{x}_{(m-1)K+k^*} = \left(x_0(s_{1,(m-1)K+k^*}), \ldots, x_0(s_{N,(m-1)K+k^*})\right) \quad (156)$$

as the actual codeword and sends it through the channel. Note that such a transmission scheme is by nature opportunistic: The encoder only spends its cost when the instants become favorite for transmission. By Lemma 13, $\boldsymbol{x}_{(m-1)K+k^*}$ is P_{X_0}-typical with constant $|\mathcal{S}|\delta$. Therefore, the cost of the transmission satisfies

$$\sum_{l=1}^{N} b\left(x_{l,(m-1)K+k^*}\right) \leq \nu \quad (157)$$

where

$$\nu = N(\mathbb{E}[b(X_0)] + |\mathcal{X}||\mathcal{S}|b_{max}\delta) \quad (158)$$

and $b_{max} = \max_{a\in\mathcal{X}} b(a)$. Choosing $\delta < \frac{\mathbb{E}[b(X_0)]}{R(A)|\mathcal{X}||\mathcal{S}|b_{\max}}\gamma$, we obtain from (153) and (158) that

$$\frac{\log_e M}{\nu} = \frac{D(P_{Y_0}\|P_{Y|X=0}) - D(P_V\|P_S) - \eta}{\mathbb{E}[b(X_0)] + |\mathcal{X}||\mathcal{S}|b_{\max}\delta} > R(A) - \gamma. \quad (159)$$

Step 4. Given the channel output matrix $Y = (\boldsymbol{y}_1^t, \ldots, \boldsymbol{y}_{MK}^t)$ in which each column represents a receive subpulse, the decoder performs MK independent binary hypothesis tests on the columns of Y. An H_1 is claimed whenever $\boldsymbol{y}_{(i-1)K+j}$, $(i,j) \in \mathcal{I}_M \times \mathcal{I}_K$, is P_{Y_0}-typical with constant $\delta' = 2|\mathcal{X}||\mathcal{S}|\delta$; otherwise, an H_0 is claimed. If there is only one column for which H_1 has been claimed, the message associated with the chosen subpulse position gives an estimate of the message $\hat{m}$. An error occurs if H_1 has been claimed more than once, or H_1 has been claimed only once but $\hat{m} \neq m$.

Step 5. The error probability is clearly independent of the message being transmitted. Assume $m = 1$, and all the probabilities will be tacitly understood to be conditioned on that event. The probability of decoding error can be bounded from above as:

$$\begin{aligned} P_e &= \mathbb{P}(\hat{m} \neq 1) \\ &\leq \prod_{j=1}^{K} \mathbb{P}(\boldsymbol{s}_j \notin T_V^N(\delta)) + \mathbb{P}(\boldsymbol{y}_{k^*} \notin T_{Y_0}^N(\delta')|\boldsymbol{s}_{k^*} \in T_V^N(\delta)) \\ &\quad + \sum_{i=2}^{M}\sum_{j=1}^{K} \mathbb{P}(\boldsymbol{y}_{(i-1)K+j} \in T_{Y_0}^N(\delta')). \end{aligned} \quad (160)$$

We next show that all three terms on the right-hand side of (160) vanish for sufficiently small δ and sufficiently large N. First, $\boldsymbol{s}_j$, $j \in \mathcal{I}_K$, are i.i.d. as P_S^N. By Part 2 of Lemma 14, we have

$$\mathbb{P}(\boldsymbol{s}_j \in T_V^N(\delta)) = P_S^N(T_V^N(\delta)) \geq \exp(-N(D(P_V\|P_S) + \eta)) \quad (161)$$

for sufficiently small δ and sufficiently large N. It follows that

$$\begin{aligned} &\prod_{j=1}^{K} \mathbb{P}(\boldsymbol{s}_j \notin T_V^N(\delta)) \\ &= \prod_{j=1}^{K} \left(1 - \mathbb{P}(\boldsymbol{s}_j \notin T_V^N(\delta))\right) && (162) \\ &\leq \left(1 - \exp\left(-N\left(D\left(P_V\|P_S\right) + \eta\right)\right)\right)^K && (163) \\ &= \left(1 - \exp\left(-N\left(D\left(P_V\|P_S\right) + \eta\right)\right)\right)^{\exp(N(D(P_V\|P_S)+2\eta))} && (164) \\ &\sim \exp\left(-\exp\left(N\eta\right)\right) && (165) \end{aligned}$$

which tends to zero in the limit as $N \to \infty$. Here, (165) is due to the well-known limit $\lim_{x\to 0}(1-x)^{1/x} = e^{-1}$. Second, for any $\boldsymbol{s}_{k^*} \in T_V^N(\delta)$, we have

$$\begin{aligned} &\mathbb{P}(\boldsymbol{y}_{k^*} \notin T_{Y_0}^N(\delta')|\boldsymbol{s}_{k^*}) \\ &= \mathbb{P}(\boldsymbol{y}_{k^*} \notin T_{Y_0}^N(\delta')|\boldsymbol{x}_{k^*}, \boldsymbol{s}_{k^*}) && (166) \\ &= 1 - \mathbb{P}(\boldsymbol{y}_{k^*} \in T_{Y_0}^N(\delta')|\boldsymbol{x}_{k^*}, \boldsymbol{s}_{k^*}) && (167) \\ &\leq 1 - P_{Y|XS}^N(T_{Y|XS}^N(\delta|\boldsymbol{x}_{k^*}, \boldsymbol{s}_{k^*})|\boldsymbol{x}_{k^*}, \boldsymbol{s}_{k^*}) && (168) \end{aligned}$$

where the last inequality follows from Lemma 13. By Part 1 of Lemma 14

$$P_{Y|XS}^N(T_{Y|XS}^N(\delta|\boldsymbol{x}_{k^*}, \boldsymbol{s}_{k^*})|\boldsymbol{x}_{k^*}, \boldsymbol{s}_{k^*}) \to 1 \quad (169)$$

in the limit as $N \to \infty$. It thus follows from (168) that

$$\mathbb{P}(\boldsymbol{y}_{k^*} \notin T_{Y_0}^N(\delta')|\boldsymbol{s}_{k^*}) \to 0, \quad \forall\, \boldsymbol{s}_{k^*} \in T_V^N(\delta) \quad (170)$$

in the limit as $N \to \infty$. Finally, $\boldsymbol{y}_{(i-1)K+j}$, $(i,j) \in \mathcal{I}'_M \times \mathcal{I}_K$, are i.i.d. as $P^N_{Y|X=0}$. By Part 2 of Lemma 14, we have

$$\begin{aligned}&\mathbb{P}(\boldsymbol{y}_{(i-1)K+j} \in T^N_{Y_0}(\delta'))\\ &\quad = P^N_{Y|X=0}(T^N_{Y_0}(\delta')) \le \exp\left(-N\left(D(P_{Y_0}\|P_{Y|X=0}) - \frac{\eta}{2}\right)\right)\end{aligned} \tag{171}$$

for sufficiently small δ (and hence sufficiently small δ') and sufficiently large N. So we have

$$\begin{aligned}\sum_{i=2}^{M}\sum_{j=1}^{K}&\mathbb{P}(\boldsymbol{y}_{(i-1)K+j} \in T^N_{Y_0}(\delta'))\\ &\le MK\exp\left(-N\left(D(P_{Y_0}\|P_{Y|X=0}) - \frac{\eta}{2}\right)\right)\\ &= \exp\left(-\frac{\eta}{2}N\right)\end{aligned} \tag{172}$$

which tends to zero in the limit as $N \to \infty$. To summarize, there exist δ_0 and N_0 (depending on δ_0) such that $P_e \le \epsilon$ for any $0 < \delta \le \delta_0$ and $N \ge N_0$.

We now conclude from Steps 1) to 5) that $R(A)$ is an achievable capacity per unit cost for any $A \in \mathcal{A}$.

B. The Converse

To prove the reverse inequality, we first bound the mutual information $I(U;Y)$ from above as:

$$I(U;Y) = \sum_u P_U(u) D(P_{Y|U=u}\|P_Y) \tag{173}$$

$$= \sum_u P_U(u) D(P_{Y|U=u}\|P_{Y|X=0}) - D(P_{Y|X=0}\|P_Y) \tag{174}$$

$$\le \sum_u P_U(u) D(P_{Y|U=u}\|P_{Y|X=0}). \tag{175}$$

We also have

$$I(U;S) = \sum_u P_U(u) D(P_{S|U=u}\|P_S), \tag{176}$$

$$\mathbb{E}[b(X)] = \sum_u P_U(u)\mathbb{E}[b(X)|U=u]. \tag{177}$$

Substituting (175)–(177) into Theorem 7, we have

$$C \le \max_{A_1} \frac{\sum_u P_U(u)\big(D(P_{Y|U=u}\|P_{Y|X=0}) - D(P_{S|U=u}\|P_S)\big)}{\sum_u P_U(u)\mathbb{E}[b(X)|U=u]} \tag{178}$$

$$\le \max_{A_1}\max_u \frac{D(P_{Y|U=u}\|P_{Y|X=0}) - D(P_{S|U=u}\|P_S)}{\mathbb{E}[b(X)|U=u]} \tag{179}$$

$$= \max_{A_1} \frac{D(P_{Y|U=u^*}\|P_{Y|X=0}) - D(P_{S|U=u^*}\|P_S)}{\mathbb{E}[b(X)|U=u^*]} \tag{180}$$

where, for a given $A_1 \in \mathcal{A}_1$, u^* is defined as

$$u^* \stackrel{\text{def}}{=} \arg\max_u \frac{D(P_{Y|U=u}\|P_{Y|X=0}) - D(P_{S|U=u}\|P_S)}{\mathbb{E}[b(X)|U=u]}. \tag{181}$$

Let $A = (X_0, V)$ be a new random variable such that

$$P_{X_0V}(x,s) = P_{XS|U}(x,s|u^*).$$

The marginal distributions satisfy

$$P_{X_0}(x) = P_{X|U}(x|u^*), \quad P_V(s) = P_{S|U}(s|u^*). \tag{182}$$

Furthermore, we have

$$P_{Y|U}(y|u^*) = \sum_x\sum_s P_{Y|XSU}(y|x,s,u^*)P_{XS|U}(x,s|u^*) \tag{183}$$

$$= \sum_x\sum_s P_{Y|XS}(y|x,s)P_{XS|U}(x,s|u^*) \tag{184}$$

$$= \sum_s\sum_x P_{Y|XS}(y|x,s)P_{X_0V}(x,s) \tag{185}$$

$$= P_{Y_0}(y) \tag{186}$$

where (184) follows from the Markov chain $U \to (X,S) \to Y$, and (186) follows from the definition of $P_{Y_0}(y)$ in Theorem 5. Substituting (182) and (186) into (180), we have the desired reverse inequality

$$C \le \max_A R(A). \tag{187}$$

Appendix C
Proof of Corollary 6

Let s_0 be an arbitrary letter in the state alphabet $\mathcal{S}$. Fix an arbitrary $P_V(\cdot)$ and an arbitrary $P_{X_0|V}(\cdot|s)$ for $s \in \mathcal{S}$, $s \ne s_0$. If we can show that $R(A)$ is maximized by a deterministic $P_{X_0|V}(\cdot|s_0)$, Corollary 6 will follow.

By (66), we have

$$P_{Y_0}(y) = \sum_s\sum_x P_{Y|XS}(y|x,s)P_{X_0|V}(x|s)P_V(s) \tag{188}$$

$$= \sum_x P_{X_0|V}(x|s_0)C_1(x,y) \tag{189}$$

where

$$\begin{aligned}C_1(x,y) \stackrel{\text{def}}{=}\, &P_{Y|XS}(y|x,s_0)P_V(s_0)\\ &+ \sum_{s\ne s_0}\sum_{x'} P_{Y|XS}(y|x',s)P_{X_0|V}(x'|s)P_V(s).\end{aligned} \tag{190}$$

It follows that the divergence $D(P_{Y_0}\|P_{Y|X=0})$ can be bounded from above as shown in (191)–(193) at the top of the following page, where (192) is due to the fact that $\sum_x P_{X_0|V}(x|s_0) = 1$, and (193) follows from the Log-Sum Inequality [15, Lemma 3.1]. The average cost can be written as

$$\mathbb{E}[b(X)] = \sum_s\sum_x b(x)P_{X_0|V}(x|s)P_V(s) \tag{194}$$

$$= \sum_x P_{X_0|V}(x|s_0)C_2(x) \tag{195}$$

where

$$C_2(x) \stackrel{\text{def}}{=} b(x)P_{X_0|V}(x|s_0) + \sum_{s\ne s_0}\sum_{x'} b(x')P_{X_0|V}(x'|s)P_V(s). \tag{196}$$

Substituting (191) and (195) into (65), we obtain (197)–(199) also at the top of the following page. Note that both $C_1(x,y)$ and $C_2(x)$ are independent of the choice of $P_{X_0|V}(\cdot|s_0)$, so is

$$D(P_{Y_0}\|P_{Y|X=0}) = \sum_y P_{Y_0}(y)\log_e \frac{P_{Y_0}(y)}{P_{Y|X=0}(y)} \tag{191}$$

$$= \sum_y \left(\sum_x P_{X_0|V}(x|s_0)C_1(x,y)\right)\log_e \frac{\sum_x P_{X_0|V}(x|s_0)C_1(x,y)}{\sum_x P_{X_0|V}(x|s_0)P_{Y|X=0}(y)} \tag{192}$$

$$\leq \sum_y\sum_x P_{X_0|V}(x|s_0)C_1(x,y)\log_e \frac{C_1(x,y)}{P_{Y|X=0}(y)} \tag{193}$$

$$R(A) \leq \frac{\sum_y\sum_x P_{X_0|V}(x|s_0)C_1(x,y)\log_e \frac{C_1(x,y)}{P_{Y|X=0}(y)} - D(P_V\|P_S)}{\sum_x P_{X_0|V}(x|s_0)C_2(x)} \tag{197}$$

$$= \frac{\sum_x P_{X_0|V}(x|s_0)\left(\sum_y C_1(x,y)\log_e \frac{C_1(x,y)}{P_{Y|X=0}(y)} - D(P_V\|P_S)\right)}{\sum_x P_{X_0|V}(x|s_0)C_2(x)} \tag{198}$$

$$\leq \max_x \frac{\sum_y C_1(x,y)\log_e \frac{C_1(x,y)}{P_{Y|X=0}(y)} - D(P_V\|P_S)}{C_2(x)}. \tag{199}$$

---

the right-hand side of (199). Observe that both inequalities (193) and (199) hold with equality if $P_{X_0|V}(\cdot|s_0)$ is a deterministic one. This completes the proof of Cor. 6.

Appendix D
Proof of Theorem 7

A. The Achievability

To prove the forward part of the theorem, we need to show that $R_1(A_1)$ is an ϵ-achievable rate per unit cost for any $A_1 \in \mathcal{A}_1$ and $0<\epsilon<1$. Given $\gamma>0$, let

$$\gamma_1 = \frac{(\mathbb{E}[b(X)])^2}{R_1(A_1)}\gamma^2 \quad \text{and} \quad \gamma_2 = \frac{\mathbb{E}[b(X)]}{R_1(A_1)}\gamma.$$

Following the random binning argument in [1, Proposition 2], there exists a positive integer n_0 such that if $n > n_0$, an (n,M,ν,ϵ) code can be found with

$$\log_e M > n(I(U;Y) - I(U;S) - \gamma_1) \tag{200}$$

and

$$\nu = n(\mathbb{E}[b(X)] + \gamma_2). \tag{201}$$

The same (n,M,ν,ϵ) code satisfies

$$\frac{\log_e M}{\nu} > R_1(A_1) - \gamma. \tag{202}$$

This holds for any $\gamma>0$. So $R(A_1)$ is ϵ-achievable per unit cost for every $0<\epsilon<1$, and $\max_{A_1\in\mathcal{A}_1} R_1(A_1)$ is an achievable rate per unit cost.

B. The Converse

To prove the converse, we will show that, for small enough ϵ, any (n,M,ν,ϵ) code must satisfy

$$\frac{\log_e M}{\nu} \leq R_1(A_1) \tag{203}$$

for some $A_1 \in \mathcal{A}_1$. To achieve this goal, we relax the cost constraint (64) on each codeword to the following average cost constraint:

$$\mathbb{E}\left[\sum_{i=1}^n b(X_i)\right] \leq \nu \tag{204}$$

where the probability distribution of X_i, $i\in\mathcal{I}_n$, is induced by that of the message W and the states $\boldsymbol{S}$. Such a relaxation only enlarges the class of admissible codes and hence strengthens our converse.

We begin with Fano's inequality which implies that every (n,M,ν,ϵ) code must satisfy

$$(1-\epsilon)\log_e M \leq I(W;\boldsymbol{Y}) + H_2(\epsilon) \tag{205}$$

where $\boldsymbol{Y} = (Y_1,\ldots,Y_n)$ and

$$H_2(\epsilon) \overset{\text{def}}{=} -\epsilon\log_e\epsilon - (1-\epsilon)\log_e(1-\epsilon) \tag{206}$$

is the binary entropy function of ϵ. The random variables W and $\boldsymbol{S}$ are independent. So we have $I(W;\boldsymbol{S})=0$ and

$$(1-\epsilon)\log_e M \leq I(W;\boldsymbol{Y}) - I(W;\boldsymbol{S}) + H_2(\epsilon). \tag{207}$$

We need the following lemma which was proved in [1, Lemma 4].

Lemma 15: There exist auxiliary random variables U_i such that $U_i \to (X_i,S_i) \to Y_i$ forms a Markov chain for all $i\in\mathcal{I}_n$ and

$$I(W;\boldsymbol{Y}) - I(W;\boldsymbol{S}) \leq \sum_{i=1}^n (I(U_i;Y_i) - I(U_i;S_i)). \tag{208}$$

Now define a time-sharing random variable Q that is uniformly distributed over the set $\mathcal{I}_n$ and is independent of all other random variables. Let random variables U_Q, X_Q, S_Q and Y_Q be U_i, X_i, S_i and Y_i, respectively when Q takes value i. With this definition, we have

$$\sum_{i=1}^n (I(U_i;Y_i) - I(U_i;S_i)) = n(I(U_Q;Y_Q|Q) - I(U_Q;S_Q|Q)) \tag{209}$$

$$= n(I(U_Q, Q; Y_Q) - I(Q; Y_Q) - I(U_Q, Q; S_Q) + I(Q; S_Q)) \tag{210}$$

$$\leq n(I(U_Q, Q; Y_Q) - I(U_Q, Q; S_Q)) \tag{211}$$

where (211) is due to $I(Q; S_Q) = 0$ since the distribution of S_i does not depend on i. Similarly, we have $P_{S_Q} = P_S$ and $P_{Y_Q|X_Q S_Q} = P_{Y|XS}$ because P_{S_i} and $P_{Y_i|S_i,X_i}$ are independent of i. Moreover, the Markov chain $U_Q \rightarrow (X_Q, S_Q) \rightarrow Y_Q$ holds because Markov chains $U_i \rightarrow (X_i, S_i) \rightarrow Y_i$ hold for all $i \in \mathcal{I}_n$. Letting $U = (U_Q, Q)$, $X = X_Q$, $S = S_Q$ and $Y = Y_Q$ and gathering (207), (208) and (211), we have

$$(1 - \epsilon) \log_e M \leq n(I(U; Y) - I(U; S)) + H_2(\epsilon). \tag{212}$$

On the other hand, by the average cost constraint (204) and the definitions of Q, X_Q and X, we have

$$\begin{aligned} \nu &\geq \sum_{i=1}^{n} \mathbb{E}[b(X_i)] = \sum_{i=1}^{n} \mathbb{E}[b(X_i)|Q = i] \\ &= n\mathbb{E}[b(X_Q)] = n\mathbb{E}[b(X)]. \end{aligned} \tag{213}$$

Putting together (212) and (213) and letting $\epsilon \downarrow 0$, the desired inequality (203) follows from the fact that $H_2(\epsilon)$ tends to zero in the limit as $\epsilon \downarrow 0$.

Acknowledgment

The authors wish to thank Prof. Ralf Koetter, Prof. Pierre Moulin, and Prof. Sergio Verdú for interesting discussions that led to some ideas in this paper.

References

[1] S. I. Gel'fand and M. S. Pinsker, "Coding for channel with random parameters," *Probl. Contr. Inf. Theory*, vol. 9, no. 1, pp. 19–31, 1980.

[2] A. El Gamal and C. Heegard, "On the capacity of computer memory with defects," *IEEE Trans. Inf. Theory*, vol. IT-29, no. 5, pp. 731–739, Sep. 1983.

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

[4] R. Zamir, S. Shamai (Shitz), and U. Erez, "Nested linear/lattice codes for structured multiterminal binning," *IEEE Trans. Inf. Theory*, vol. 48, no. 6, pp. 1250–1276, Jun. 2002.

[5] T. Philosof, U. Erez, and R. Zamir, "Combined shaping and precoding for interference cancellation at low SNR," in *Proc. IEEE Int. Symp. Information Theory*, Yokohama, Japan, Jun./Jul. 2003, p. 68.

[6] U. Erez and S. ten Brink, "Approaching the dirty paper limit in canceling known interference," in *Proc. Allerton Conf. Communications, Control, and Computing*, Monticello, IL, Oct. 2003, pp. 799–808.

[7] A. Bennatan, D. Burshtein, G. Caire, and S. Shamai (Shitz), "Superposition coding for side-information channels," *IEEE Trans. Inf. Theory*, vol. 52, no. 5, pp. 0000–0000, May 2006.

[8] H. A. David, *Order Statistics*, 1st ed. New York: Wiley, 1970.

[9] A. S. Cohen and A. Lapidoth, "The Gaussian watermarking game," *IEEE Trans. Inf. Theory*, vol. 48, no. 6, pp. 1639–1667, Jun. 2002.

[10] S. R. S. Varadhan, *Probability Theory*. Providence, RI: Amer. Math. Soc., 2001.

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

[12] A. Lapidoth and S. Shamai (Shitz), "Fading channels: How perfect need 'perfect side information' be?," *IEEE Trans. Inf. Theory*, vol. 48, no. 5, pp. 1118–1134, May 2002.

[13] S. Verdú, "On channel capacity per unit cost," *IEEE Trans. Inf. Theory*, vol. 36, no. 5, pp. 1019–1030, Sep. 1990.

[14] F. M. J. Willems, "An information theoretic approach to information embedding," in *Proc. 21st Symp. Information Theory Benelux*, Wassenaar, The Netherlands, May 2000, pp. 255–260.

[15] I. Csiszár and J. Körner, *Information Theory: Coding Theorem for Discrete Memoryless Systems*, 2nd ed. Budapest, Hungary: Akadémiai Kiadó, 1986.

[16] S. Kullback, *Information Theory and Statistics*. New York: Wiley, 1959.

[17] N. M. Blachman, "The convolution inequality for entropy powers," *IEEE Trans. Inf. Theory*, vol. IT-11, no. 2, pp. 267–271, Apr. 1965.

[18] 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.

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

[20] S. Verdú, "Spectral efficiency in the wideband regime," *IEEE Trans. Inf. Theory*, vol. 48, no. 6, pp. 1319–1343, Jun. 2002.