The Duality Between Information Embedding and Source Coding With Side Information and Some Applications

Richard J. Barron, *Member, IEEE*, Brian Chen, *Member, IEEE*, and Gregory W. Wornell, *Senior Member, IEEE*

***Abstract*—Aspects of the duality between the information-embedding problem and the Wyner–Ziv problem of source coding with side information at the decoder are developed and used to establish a spectrum new results on these and related problems, with implications for a number of important applications. The single-letter characterization of the information-embedding problem is developed and related to the corresponding characterization of the Wyner–Ziv problem, both of which correspond to optimization of a common mutual information difference. Dual variables and dual Markov conditions are identified, along with the dual role of noise and distortion in the two problems.**

For a Gaussian context with quadratic distortion metric, a geometric interpretation of the duality is developed. From such insights, we develop a capacity-achieving information-embedding system based on nested lattices. We show the resulting encoder–decoder has precisely the same decoder–encoder structure as the corresponding Wyner–Ziv system based on nested lattices that achieves the rate-distortion limit.

For a binary context with Hamming distortion metric, the information-embedding capacity is developed, along with its relationship to the corresponding Wyner–Ziv rate-distortion function. In turn, an information-embedding system for this case based on nested linear codes is constructed having an encoder–decoder that is identical to the decoder–encoder structure for the corresponding system that achieves the Wyner–Ziv rate-distortion limit.

Finally, based on these results, a simple layered joint source–channel coding system is developed with a perfectly symmetric encoder–decoder structure. Its application and performance is discussed in a broadcast setting in which there is a need to control the fidelity experienced by different receivers. Among other results, we show that such systems and their multilayer extensions retain attractive optimality properties in the Gaussian-quadratic case, but not in the binary-Hamming case.

***Index Terms*—Coding with side information, data hiding, digital watermarking, hybrid coding and transmission, information embedding, joint source–channel coding, Slepian–Wolf coding, Wyner–Ziv coding.**

Manuscript received January 21, 2000; revised October 28, 2002. This work was supported in part by the Air Force Office of Scientific Research under Grant F49620-96-1-0072, the Army Research Laboratory under Cooperative Agreement DAAL01-96-2-0002, the National Science Foundation under Grant CCR-0073520, and under grants from MIT Lincoln Laboratory Advanced Concepts Committee and Microsoft Research. The material in this paper was presented in part at the IEEE International Symposium on Information Theory, Washington, DC, June 2001.

The authors are with the Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA 02139 USA (e-mail: rjbarron@alum.mit.edu; bchen@alum.mit.edu; gww@mit.edu).

Communicated by I. Csiszár, Associate Editor for Shannon Theory.

Digital Object Identifier 10.1109/TIT.2003.810639

I. Introduction

INFORMATION embedding concerns the reliable transmission of information embedded into a host signal, and has an increasingly wide array of applications, from digital watermarking, data hiding, and steganography, to backward-compatible digital upgrading of communications infrastructure [7], [6]. Likewise, source coding with side information has a growing spectrum of applications, ranging from new low-power sensor networks to the upgrading of legacy communications infrastructure [28], [1].

This paper develops the natural duality between information embedding, which can be reinterpreted as a problem of *channel* coding with side information at the *encoder* [7], and the problem of *source* coding with side information at the *decoder*, the most important instance of which is the well-known "Wyner–Ziv" problem [34]. Exploiting this duality, several new results and interesting insights with practical implications are obtained, including several in the context of mixed analog–digital transmission.

Fig. 1 depicts the information-embedding scenario of interest. The n-dimensional vector $\boldsymbol{X}$ is the "host" signal, and the message M is the information to be embedded, which is independent of $\boldsymbol{X}$. The encoder uses both the host and the message to create a "composite" signal $\boldsymbol{W}$ that is suitably close to the host $\boldsymbol{X}$. The composite signal passes through a probabilistic channel, the output of which, $\boldsymbol{Y}$, is reliably decoded to retrieve the embedded message M.[1] In our model for information embedding, each element of the host $\boldsymbol{X}$ is drawn in an independent and identically distributed (i.i.d.) manner from the distribution $p_X(x)$, and the channel is memoryless and characterized by the transition density $p_{Y|W}(y|w)$.[2] The specific information-embedding problem is as follows: if the distortion between the host and composite signal is constrained to be at most d, what is the maximum rate R of reliable communication that can be supported by the embedding given a particular transmission channel?

The dashed line in Fig. 1 represents a less interesting variant of information embedding whereby the host is also known to the decoder. Wolfowitz [31] originally derived capacity for this system without the distortion constraint, i.e., capacity with side information at the encoder and decoder. For the purposes of this

[1]The decoder can also extract $\boldsymbol{W}$ from $\boldsymbol{Y}$, thereby reconstructing the original host to within distortion d.

[2]For watermarking problems, a variety of attack channel models of the form considered in [23], [7], and [12] are also of particular interest, although we do not consider such channels in this paper.

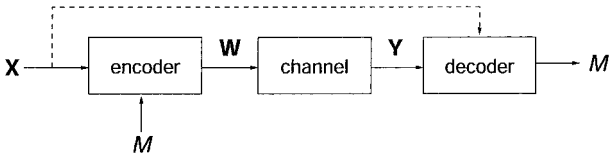

Fig. 1. The information-embedding model. The signals $\boldsymbol{X}$, M, $\boldsymbol{W}$, and $\boldsymbol{Y}$ are, respectively, the host, embedded information, composite signal, and channel output. The dashed line represents side information at the decoder, which may or may not be present, depending on the application.

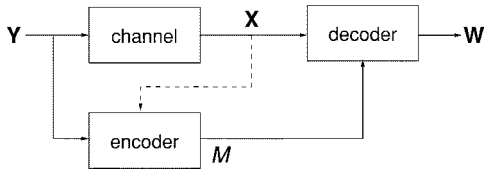

Fig. 2. The source coding with side information model. The signals $\boldsymbol{Y}$, M, $\boldsymbol{X}$, and $\boldsymbol{W}$ are respectively the source, digital encoding, channel output, and decoded source. The dashed line represents side information at the encoder, which may or may not be present, depending on the application.

paper, we refer to the case where the host is also known at the decoder as "private" information embedding, and to the case where the host is not also known at the decoder as "public" information embedding. While we examine both forms of embedding, we emphasize public information embedding in our development, and when there is no risk of confusion we use the term "information embedding" generically to refer to this case.

Fig. 2 depicts the source coding with side information problem of interest. The n-dimensional source vector $\boldsymbol{Y}$ passes through a probabilistic channel, producing the side information $\boldsymbol{X}$. The encoder produces the message M from the source $\boldsymbol{Y}$ that the decoder uses in conjunction with $\boldsymbol{X}$ to produce a suitably accurate reconstruction $\boldsymbol{W}$ of $\boldsymbol{Y}$. In our model for source coding with side information, the source is drawn i.i.d. from $p_Y(y)$, and the channel is memoryless with transition density $p_{X|Y}(x|y)$. For this problem, the question is: given a particular side information channel, what is the minimum rate R that is required at the output of the encoder to ensure that distortion between the source and reconstruction is at most d?

The dashed line in Fig. 2 represents a less interesting variant of the source coding with side information problem whereby the side information is also known to the encoder. When the side information is also known to the encoder, achievable performance is easily characterized in terms of a familiar conditional rate-distortion function [5], [21]. When the side information is not also known to the encoder, we have the problem considered by Wyner and Ziv [34]. While we develop dualities associated with both forms of the source-coding problem, we emphasize the Wyner–Ziv version, and when there is no risk of confusion we describe both versions generically as the "Wyner–Ziv problem."

As Figs. 1 and 2 suggest, there is a one-to-one correspondence between variables in the information-embedding and Wyner–Ziv problems. Indeed, our notation is chosen so as to identify the correspondence between variables in the two problems that arises out of the duality, as we will discuss.[3] For the moment, it suffices to observe that the information-embedding *encoder* has exactly the same input variables ($\boldsymbol{X}$ and M) and output variable ($\boldsymbol{W}$) as the *decoder* for the Wyner–Ziv problem. Furthermore, the information-embedding *decoder* has the same input variable ($\boldsymbol{Y}$) and output variable (M) as the *encoder* for the Wyner–Ziv problem. As we illustrate in some key contexts of interest, this is not a coincidence: the two problems are, in fact, duals in the sense that an optimal encoder–decoder for one problem is an optimal decoder–encoder pair for the other.

In developing the deeper connection, we show that, in general, the capacities and rate-distortion limits for the two problems are closely related, and can be expressed in terms of an optimization over the same mutual information difference with respect to the free parameters in each problem. Moreover, we show that distortion and channel noise play dual roles in the two problems.

In addition to our own work [1], [6], [2], there has been growing interest in aspects of the subject of this paper in recent times, and an expanding set of results and insights. Su, Eggers, and Girod [30] consider the Gaussian-quadratic special case and have a similar geometric interpretation to ours. Chiang and Cover [9], [17], [16] expand the scope of the duality beyond the information-embedding context. Chou, Pradhan, and Ramchandran describe aspects of the duality in [10] and investigate it further in [24], [25].

An outline of the paper is as follows. After establishing some basic notation in Section II, we develop and relate the basic single-letter characterizations for the two problems in Section III. For the information-embedding problem, we generalize a result of Gel'fand and Pinsker to include a distortion constraint and an arbitrary metric; other versions of this problem are considered by Moulin and O'Sullivan [23]. In Appendixes I and II, we provide the proofs for the coding theorems for public and private information embedding, respectively, emphasizing the duality with the corresponding Wyner–Ziv problem. We then discuss the duality between the resulting mutual information optimization and Markov conditions for the information-embedding and Wyner–Ziv problems. We further examine the dual relationship between distortion and channel noise in the two problems, developing the correspondence between the noise-free and distortion-free special cases of each problem. Among other insights, we discuss the resulting duality between a version of Slepian–Wolf encoding and information embedding for noise-free channels.

Section IV examines the duality further in the case of Gaussian contexts with a quadratic distortion metric. In this case, we relate the information embedding capacity and Wyner–Ziv rate-distortion function geometrically, which emphasizes the dual relationship between distortion and channel noise in the two problems. We then proceed to build deterministic information-embedding systems based on nested lattices that achieve capacity at high signal-to-distortion ratio (SDR). Such systems are also developed independently by Erez, Shamai, and Zamir in [18]. We show how the resulting encoder–decoder pair

[3]Throughout this paper it will be clear through context whether a variable to which we refer corresponds to the information-embedding problem or the Wyner–Ziv problem.

has the *identical* structure as the associated decoder–encoder pair for a deterministic Wyner–Ziv embedding system based a nested lattices, which achieves the rate-distortion limit at high signal-to-noise ratio (SNR), and which is a nondithered version of the Wyner–Ziv solution developed by Zamir and Shamai [36]. We further develop, in Appendix III, a Wyner–Ziv code based on nested lattices with dithering that achieves the rate-distortion limit at any SNR.

Section V examines the duality further in the case of binary contexts with a Hamming distortion metric. In this case, we use our results in Section III to compute the information-embedding capacities (Appendix IV), and highlight the close relationship—between both the proofs and the final expressions—to the corresponding Wyner–Ziv rate-distortion function developed in [34]. We then proceed to build deterministic information-embedding systems for this case based on nested linear codes that achieve capacity, and show how the resulting encoder–decoder pair has the identical structure as the associated decoder–encoder pair developed in [28] for achieving the Wyner–Ziv rate-distortion limit. In the noise-free special case, the information-embedding system we construct is the dual of Wyner's Slepian–Wolf code construction [32].

Finally, in Section VI, we exploit our results in the development of a new class of layered joint source–channel coding systems from the interconnection of information embedding and Wyner–Ziv subsystems. The new systems can be used in a broadcast setting in which one wants to control the fidelity available to different groups of users. We show that, with our construction, no price need be paid for this extra functionality in the Gaussian-quadratic case, but that there is a cost in the binary-Hamming case. A unique feature of our coding system is that the encoder and decoder share *identical* structure.

Section VII contains some concluding remarks.

II. Notation

In terms of general notation, the components of a length n random vector $\boldsymbol{V}$ are denoted $V_1, \ldots, V_n$. In turn, we use V_j^k to denote a vector comprised of the jth through kth components of $\boldsymbol{V}$, where if the subscript is omitted, j is implicitly 1; whence $\boldsymbol{V} = V_1^n = V^n$. A script $\mathcal{V}$ is used to denote the alphabet of the random variable V. Except when otherwise indicated (i.e., in Gaussian scenarios), all random variables in this paper take on values from finite alphabets. We use $D(u, \hat{u}) > 0$ to denote a general distortion measure. The expressions $I(\cdot;\cdot)$, $H(\cdot)$, and $H(\cdot|\cdot)$ denote Shannon's mutual information, entropy, and conditional entropy, respectively. All logarithms in this paper are to be interpreted base-2.

III. Single-Letter Characterizations of Capacity and Rate Distortion

In this section, we describe the single-letter expressions for the distortion-constrained public information embedding capacity and the Wyner–Ziv rate-distortion function. We compare these expressions to those when the host (respectively, source) $\boldsymbol{X}$ is known at the decoder (respectively, encoder).

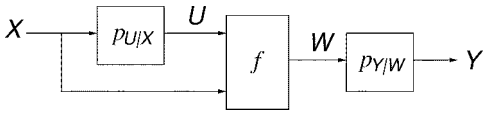

Fig. 3. Illustration of the variable relationship in the single-letter characterization of information embedding, where U is the auxiliary random variable.

A. Public Information Embedding Capacity

The capacity of public information embedding subject to an embedding distortion constraint is denoted $C^{\mathrm{IE}}(d)$. It is defined as the maximum achievable rate for communicating a message M such that $\Pr(\hat{M} \neq M)$ is arbitrarily small and $E[\frac{1}{n}\sum_{k=1}^{n} D(X_k, W_k)]$ is arbitrarily close to d for sufficiently large n.

The following result is a generalization of that of Gel'fand and Pinsker [20] and Heegard and El Gamal [22], which consider the problem without a distortion constraint.

Claim 1: For general distortion measures $D(\cdot,\cdot)$, the capacity $C^{\mathrm{IE}}(d)$ can be expressed in the form

$$C^{\mathrm{IE}}(d) = \sup I(Y; U) - I(U; X) \tag{1}$$

where the supremum is taken over all distributions $p_{U|X}(u|x)$ and functions $f\colon \mathcal{U} \times \mathcal{X} \mapsto \mathcal{W}$ satisfying

$$E[D(X, W)] \leq d \qquad \text{where } W = f(U, X) \tag{2}$$

where U is an auxiliary random variable.

The relationship between the primary and auxiliary random variables in this single-letter characterization is depicted in Fig. 3. To prove Claim 1, we begin by using an extension of the reasoning in [20] to show that the rate $I(Y; U) - I(U; X)$ is achievable. The basic encoder and decoder construction is as follows. A random codebook $\mathcal{C}$ is generated with 2^{nR_1} i.i.d. codewords $\boldsymbol{U} \sim \prod_{i=1}^{n} p_U(U_i)$, where $R_1 = I(Y; U) + \epsilon$. The codewords are distributed randomly into 2^{nR_2} bins, where $R_2 = I(Y; U) - I(U; X) - 2\epsilon$. At the encoder, the embedded information M specifies the bin which is used to code the source. The encoder finds the codeword $\boldsymbol{U}$ in that bin that is jointly distortion typical with the host $\boldsymbol{X}$ and transmits it.[4] The decoder looks for the code vector in all of $\mathcal{C}$ that is jointly typical with the channel output $\boldsymbol{Y}$. The bin index of that code vector is the decoded information $\hat{M}$. That this encoder and decoder structure has the requisite properties is straightforward as shown in [1]. It remains only to show the converse, which is provided in Appendix I and relies on the concavity of $C^{\mathrm{IE}}(d)$.

Finally, observe that since U is an auxiliary random variable, the characterization of the physical channel in this problem is such that $U \rightarrow (W, X) \rightarrow Y$ form a Markov chain, whence

$$I(Y; U|W, X) = 0. \tag{3}$$

[4]The main difference between this achievability proof and that of [20] is that joint distortion typicality—not just joint typicality—is required to meet the embedding distortion constraint.

B. Private Information Embedding Capacity

The corresponding result to Claim 1 for private information embedding subject to a distortion constraint with arbitrary metric is summarized as follows.

Claim 2: The private information embedding capacity, denoted $C^{\mathrm{IE}}_{\mathrm{priv}}(d)$, is given by

$$C^{\mathrm{IE}}_{\mathrm{priv}}(d) = \sup I(Y; W|X) \tag{4}$$

where the supremum is taken over all $p_{W|X}(w|x)$ such that $E[D(X, W)] \leq d$.

A proof is provided in Appendix II. The construction for achievability involves the use of a set of codebooks, each of which is a capacity-achieving codebook for a particular host value $X = x \in \mathcal{X}$. The total achievable rate is thus the expected value of the conditional capacities over X, and the average distortion is the expected value of the distortions over all the codebooks. The converse exploits the concavity of $C^{\mathrm{IE}}_{\mathrm{priv}}(d)$.

The public and private information embedding capacities are related by

$$C^{\mathrm{IE}}(d) \leq C^{\mathrm{IE}}_{\mathrm{priv}}(d) \tag{5}$$

where equality in (5) holds if and only if the maximizing distribution for X, Y, U, W in (1) also maximizes the argument on the right-hand side of (4), and if with this distribution

$$I(X; U|Y) = 0 \tag{6}$$

i.e., $U \rightarrow Y \rightarrow X$ form a Markov chain.

To verify (6), we first obtain, by expanding $I(Y, X; U)$, two different ways using the chain rule

$$I(Y; U) - I(U; X) = I(Y; U|X) - I(U; X|Y) \tag{7}$$

where U is any auxiliary random variable such that (2) is satisfied. Likewise applying the chain rule to $I(Y; U, W|X)$ we obtain

$$I(Y;U|X) - I(Y;W|X) = I(Y;U|W,X) - I(Y;W|U,X). \tag{8}$$

However, from (3), the first term on the right-hand side of (8) is zero, and since W is a deterministic function of U and X, the second term on the right-hand side of (8) is also zero. Thus, (8) implies $I(Y; U|X) = I(Y; W|X)$ and (7) can be rewritten as

$$I(Y; W|X) = [I(Y; U) - I(U; X)] + I(X; U|Y). \tag{9}$$

Comparing (9) with (1) and (4), we obtain the stated necessary and sufficient conditions for the public and private embedding capacities to be equal.

C. Rate-Distortion Function With Side Information at the Decoder

In [34], Wyner and Ziv define the rate-distortion function with side information at the decoder, denoted $R^{\mathrm{WZ}}_{Y|X}(d)$, as the minimum data rate at which M can be transmitted such that when n is large the average distortion $E[\frac{1}{n}\sum_{k=1}^{n} D(Y_k, W_k)]$ is arbitrarily close to d.

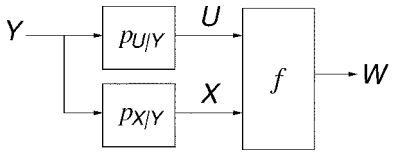

Fig. 4. Illustration of the variable relationships in the single-letter characterization of source coding with side information, where U is the auxiliary random variable.

Their main result is the following:

$$R^{\mathrm{WZ}}_{Y|X}(d) = \inf I(Y; U) - I(U; X) \tag{10}$$

where the infimum is taken over all $p_{U|Y}(u|y)$ and functions $f\colon \mathcal{U} \times \mathcal{X} \mapsto \mathcal{W}$ such that

$$U \rightarrow Y \rightarrow X \text{ is a Markov chain} \tag{11}$$

and

$$E[D(Y, W)] \leq d, \qquad \text{where } W = f(U, X) \tag{12}$$

where U is an auxiliary random variable. The relationship between the primary and auxiliary random variables in this single-letter characterization is depicted in Fig. 4.

Note that the objective functions on the right-hand sides of (10) and (1) are identical, as occurs in the case of the duality between source and channel coding without side information [15]. Condition (11), i.e., X and U are conditionally independent given Y, implies (c.f. (6))

$$I(X; U|Y) = 0 \tag{13}$$

which using (7) simplifies (10) to

$$R^{\mathrm{WZ}}_{Y|X}(d) = \inf I(Y; U|X). \tag{14}$$

The achievability proof [15] and that used for the information-embedding problem [1] are mirrors of each other. Indeed, the Wyner–Ziv encoder (respectively, decoder) is used in precisely the same manner as the information-embedding decoder (respectively, encoder). Likewise, whereas the converse for information embedding relies on the concavity of $C^{\mathrm{IE}}(d)$, the converse for the Wyner–Ziv problem relies on the convexity of $R^{\mathrm{WZ}}_{Y|X}(d)$.

D. Conditional Rate-Distortion Function

Source coding with the side information known at the decoder *and* encoder is the dual of private information embedding. As shown by Berger [5] and Gray [21], the achievable rate is given by the conditional rate-distortion function

$$R_{Y|X}(d) = \inf I(Y; W|X) \tag{15}$$

where the infimum is taken over all distributions $p_{W|X}(w|x)$ such that $E[D(Y, W)] \leq d$.

The proof of this result mirrors the proof of private information embedding described in Appendix II. In particular, achievability of the conditional rate-distortion function is proven by a "switching" argument; for each $x \in \mathcal{X}$, an optimal rate-distortion codebook is used to code the source samples Y_i for all i such that $X_i = x$. The total rate is thus the expectation over X of the marginal rate-distortion functions, and the distortion is the

expected value of the distortions over all the codebooks. Likewise, in the same way that the converse for private information embedding exploits the concavity of $C^{\mathrm{IE}}_{\mathrm{priv}}(d)$, the converse for the conditional rate-distortion problem exploits the convexity of $R_{Y|X}(d)$.

The Wyner–Ziv and conditional rate-distortion problems are related by

$$R_{Y|X}(d) \le R^{\mathrm{WZ}}_{Y|X}(d) \tag{16}$$

where, as shown in [34], equality in (16) holds if and only if the minimizing distribution for X, Y, U, W in (10) also minimizes the objective function on the right-hand side of (15), and if with this distribution we have (c.f. (3))

$$I(Y;\, U|W,\, X) = 0 \tag{17}$$

i.e., $U \to (X,\, W) \to Y$ form a Markov chain.

E. Duality of Necessary and Sufficient Conditions

The relationships described in the preceding subsections reveal an important aspect of the duality between information embedding and source coding with side information.

To summarize, with information embedding (respectively, source coding with side information), for the embedding capacity (respectively, rate-distortion function) to be the same whether or not the host (respectively, side information) is known at the decoder (respectively, encoder), the optimizing distributions for X, Y, U, W must first be the same with or without the signal X known at the encoder (respectively, decoder).

The duality manifests itself in the remaining necessary condition. For information embedding this is the Markov condition (6), which for the Wyner–Ziv problem is automatically satisfied (c.f. (13)). Similarly, for the Wyner–Ziv problem, the remaining necessary condition is the Markov constraint (17), which for the case of information embedding is automatically satisfied (c.f. (3)).

The Markov condition not automatically satisfied by the problem construction may or may not be satisfied. Indeed, in Section IV, we will see that it *is* for both problems in the Gaussian-quadratic case, while in Section V we will see that it *is not* for either problem in the binary-Hamming case.

Unless otherwise noted, for the remainder of this paper, we restrict our attention to the problems of source coding with side information known only at the decoder, and information embedding with side-information known only at the encoder.

F. Noise-Free/Distortion-Free Duality

In this subsection, we examine important limiting cases of the duality between information embedding and Wyner–Ziv coding, corresponding to noise-free and distortion-free scenarios. First, we observe that distortion-free information embedding and noise-free Wyner–Ziv encoding are trivial duals

$$R^{\mathrm{WZ}}_{\text{noise-free}}(d) = 0 = C^{\mathrm{IE}}(0). \tag{18}$$

In the other limiting case—noise-free information embedding and distortion-free Wyner–Ziv coding—the duality is more interesting.

The minimum rate $R^{\mathrm{WZ}}(0)$ required for distortion-free Wyner–Ziv coding follows immediately from an application of the Slepian–Wolf source-coding theorem [29]. In particular, the source can be reproduced exactly at the decoder ($\hat{\boldsymbol{W}} = \boldsymbol{Y}$) if and only if [15, Sec. 14.4]

$$R \ge H(Y|X) = R^{\mathrm{WZ}}(0) \tag{19}$$

where the underlying density $p_{Y|X}(y|x)$ is prescribed by the problem, so no infimum in (19) is required.

To see the duality to noise-free information embedding, we develop the associated capacity in the sequel.

1) Noise-Free Information Embedding Capacity: The maximum rate that can be attained for noise-free information embedding is closely related [6]. In particular, the dual result is as follows: one can reliably embed a message M in the host signal for transmission over an error-free channel if and only if

$$R \le \max H(Y|X) = C^{\mathrm{IE}}_{\text{noise-free}}(d) \tag{20}$$

where the maximum in (20) is over all distributions $p_{Y|X}(y|x)$ such that $E[D(X,\, W)] \le d$.

Equation (20) is verified as follows. We first show that, even with the constraint $U = W$ in (1), the rate $H(Y|X)$ is achievable

$$\begin{aligned} C &= \sup I(Y;\, U) - I(U;\, X) \\ &\ge I(Y;\, Y) - I(Y;\, X) \\ &= H(Y) - [H(Y) - H(Y|X)] \\ &= H(Y|X) \end{aligned} \tag{21}$$

where we have used $U = W = Y$ in the second line. Now, we shall show that the capacity (1) cannot exceed $H(Y|X)$

$$\begin{aligned} I(Y;U) - I(U;X) &= H(U) - H(U|Y) - H(U) + H(U|X) \\ &= H(U|X) - H(U|Y) \\ &\le H(U|X) - H(U|Y,\, X) \\ &= I(U;\, Y|X) \\ &= H(Y|X) - H(Y|U,\, X) \\ &\le H(Y|X). \end{aligned}$$

The third line follows since conditioning decreases entropy. The final line arises since entropy is nonnegative.

It remains only to maximize this resulting rate $R = H(Y|X)$ over all possible choices of $p_{U|X}(u|x)$. Equation (20) is expressed in terms of the equivalent maximization over $p_{Y|X}(y|x)$ since $Y = W = U$.

IV. Gaussian-Quadratic Case

In this section, we examine the information embedding capacity and rate-distortion function in the case of a (continuous-alphabet) Gaussian host and source, respectively, a memoryless Gaussian channel, and a quadratic distortion metric. Our development reveals the duality in the derivations of these bounds and in the codes that achieve them.

A. Gaussian-Quadratic Information Embedding Capacity

Consider an i.i.d. Gaussian host $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{0},\, \sigma_X^2 \boldsymbol{I})$ and a channel that adds white Gaussian noise $\boldsymbol{V} \sim \mathcal{N}(\boldsymbol{0},\, \sigma_V^2 \boldsymbol{I})$ that is

independent of $\boldsymbol{X}$, where $\mathcal{N}(\boldsymbol{m}, \boldsymbol{\Lambda})$ denotes Gaussian random vector with mean $\boldsymbol{m}$ and covariance matrix $\boldsymbol{\Lambda}$.[5] The message M is embedded into $\boldsymbol{X}$, creating a composite signal $\boldsymbol{W}$ such that the mean-square embedding distortion is minimized: $D(x, w) = (x - w)^2$. The capacity $C^{\mathrm{IE}}(d)$ of this system is given by [14]

$$C^{\mathrm{IE}}(d) = \frac{1}{2}\log\left(1 + \frac{d}{\sigma_V^2}\right). \tag{22}$$

Costa proves this result in the context of coding for a channel with a random state known at the encoder. Using a convenient super-channel interpretation of information embedding, Chen and Wornell [7] cite Costa's expression as the information-embedding capacity for the Gaussian case.

Costa first proves that the information-embedding capacity with X known at the encoder and decoder equals the expression in (22). He then proceeds to show that with no host at the decoder, there is a test channel which achieves this capacity.

The test channel used to determine capacity defines the auxiliary random variable $U = \alpha X + E$ for some constant α and with E zero-mean, Gaussian, and independent of X, implying that the encoding function is $W = f(U, X) = U + (1 - \alpha)X$. Solving for $I(Y; U) - I(U; X)$ and maximizing with respect to α yields (22).

B. Gaussian-Quadratic Wyner–Ziv Rate-Distortion Function

The Wyner–Ziv rate-distortion function for a Gaussian source with jointly Gaussian side information at the decoder is a dual to the distortion-constrained information embedding capacity with Gaussian host and Gaussian channel.

For jointly Gaussian $\boldsymbol{X}$ and $\boldsymbol{Y}$ whose element pairs are all drawn i.i.d. from the Gaussian density $f_{X,Y}(x, y) \sim \mathcal{N}(0, \Lambda_{XY})$, the Wyner–Ziv rate distortion function is [33]

$$R_{Y|X}^{\mathrm{WZ}}(d) = \begin{cases} \frac{1}{2}\log\left(\frac{\sigma_{Y|X}^2}{d}\right), & \text{if } 0 \le d < \sigma_{Y|X}^2 \\ 0, & \text{if } d \ge \sigma_{Y|X}^2 \end{cases} \tag{23}$$

where $\sigma_{Y|X}^2$ is the error variance in the minimum mean-square error (MMSE) estimation of Y from X. We can always write the relationship between X and Y in the form $X = \beta Y + V$ for some β, where V is Gaussian with variance σ_V^2 and independent of Y. Without loss of generality, we restrict our attention to the case $\beta = 1$.

Wyner [33] proves (23) by first showing that the conditional rate-distortion function equals the expression in (23), mirroring the approach used by Costa in the corresponding information-embedding problem. He then proceeds to show that with no side information at the encoder, there is a test channel which achieves the same rate-distortion function, thereby finding the Wyner–Ziv rate-distortion function.

In Wyner's formulation, the test channel encoder simply assigns the auxiliary random variable U to be a linear combination of the source and an independent zero-mean Gaussian variable: $U = \alpha Y + E$. The test channel decoder function is also a linear function. For the special case of an additive white Gaussian channel with SNR $\to \infty$, the decoder function is

$$W = f(U, X) = U + (1 - \alpha)X. \tag{24}$$

Note that this special-case decoder is the same as the information-embedding encoding function for the Gaussian case.

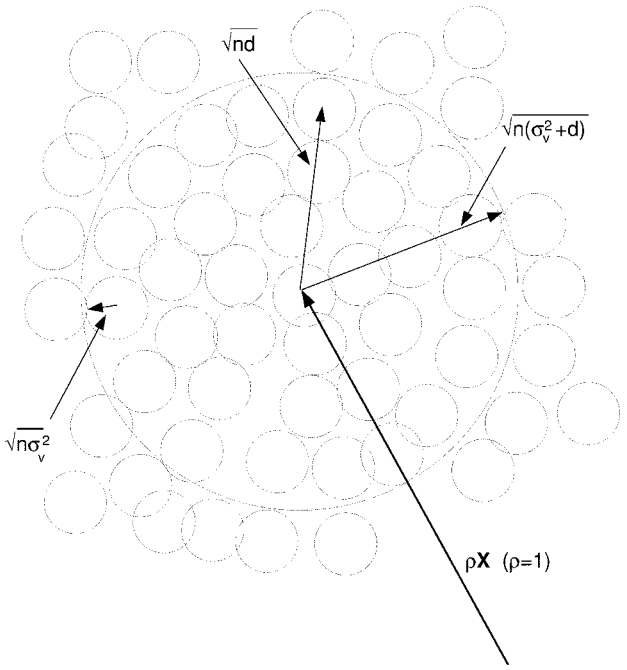

Fig. 5. Geometric interpretation of information embedding as sphere packing in the Gaussian-quadratic case.

C. Geometrical Interpretations

The duality between the information-embedding capacity and Wyner–Ziv rate-distortion function in the Gaussian case has a convenient geometrical interpretation, which we illustrate in this subsection.[6] In particular, we show how information embedding is sphere packing about the host in signal space, while Wyner–Ziv encoding is sphere covering about a source estimate that is a linear function of the side information.

1) Geometry of Information Embedding: Information embedding can be viewed as a sphere-packing problem, as depicted in Fig. 5 in the high distortion-to-noise ratio (DNR) regime. To understand this figure, note that the distortion constraint implies that all composite signals $\boldsymbol{W}$ must be contained in a sphere S_X of radius $\sqrt{nd}$ centered about $\boldsymbol{X}$. In coding for the channel, we use 2^{nR} codewords (signal points) that must be contained within S_X such that smaller spheres of radius $\sqrt{n\sigma_V^2}$ about all of the signal points have negligible overlap—each symbol will be uniquely distinguishable at the decoder. We emphasize that this must be true for all $\boldsymbol{X}$, so that if $\boldsymbol{X}$ changes by some amount, the positions of signal points may change, but the number of signal points will stay the same. Signal design corresponds to filling a sphere of radius $\sqrt{n(d + \sigma_V^2)}$ with smaller spheres of radius $\sqrt{n\sigma_V^2}$.

[5] We use $\boldsymbol{I}$ to denote the identity matrix.

[6] A similar geometrical interpretation is given in [30].

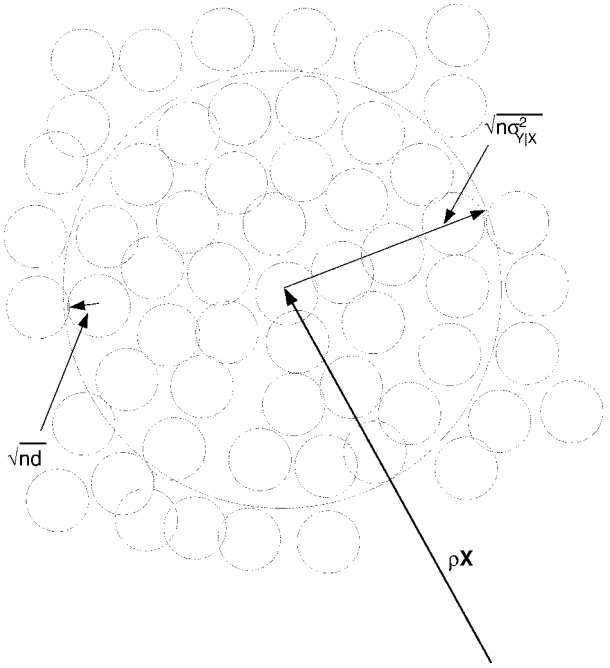

Fig. 6. Geometric interpretation of Wyner–Ziv coding as sphere covering in the Gaussian-quadratic case.

With this geometrical interpretation, clearly the maximum number of spheres that can be used is upper-bounded by the ratio of the volumes of the large to the small spheres. Thus, the number of codewords is bounded

$$2^{nR} \leq \frac{\left(\sqrt{n(d+\sigma_V^2)}\right)^n}{\left(\sqrt{n\sigma_V^2}\right)^n} = \left(\frac{d+\sigma_V^2}{\sigma_V^2}\right)^{n/2}. \quad (25)$$

From (22), we see that a capacity-achieving code will meet this upper bound as

$$2^{nC} = \left(\frac{d+\sigma_V^2}{\sigma_V^2}\right)^{n/2} \quad (26)$$

for large n.

2) Geometry of Wyner–Ziv Encoding: Wyner–Ziv coding can be viewed as a sphere-covering problem, as depicted in Fig. 6 in the low DNR regime. Given a side information vector $\boldsymbol{X}$ at the decoder, an MMSE estimate of the source is $\hat{\boldsymbol{Y}} = \rho\boldsymbol{X}$, where ρ is the associated MMSE estimator gain. The remaining mean-square error about the estimate is $\sigma_{Y|X}^2$, implying that the source must lie in a sphere $S_{Y|X}$ of radius $\sqrt{n\sigma_{Y|X}^2}$ about $\rho\boldsymbol{X}$. Moreover, the noisier the channel from Y to X, the larger this sphere. A Wyner–Ziv codebook for a distortion d will contain $2^{nR(d)}$ code vectors in $\mathbb{R}^n$, and is designed so that most source sequences of length n lying in $S_{Y|X}$ are within a distance $\sqrt{nd}$ of a codeword. Rate-distortion coding for the Gaussian case, therefore, amounts to covering the sphere $S_{Y|X}$ with smaller spheres of radius $\sqrt{nd}$, which we illustrate in Fig. 6. Clearly the number of codewords is lower-bounded by the ratio of the volumes of the large to the small spheres

$$2^{nR(d)} \geq \left(\frac{\sigma_{Y|X}^2}{d}\right)^{n/2} \quad (27)$$

and this lower bound is met by a code that achieves the rate-distortion bound given by (23).

D. Geometrical Duality

The geometric interpretation of the Gaussian case shows that the encoder (respectively, decoder) operation for information embedding is the same as the decoder (respectively, encoder) operation for Wyner–Ziv coding. At the information-embedding encoder, the digital information M specifies a signal point in a sphere about a signal $\boldsymbol{X}$, and similarly, at the Wyner–Ziv decoder the digital information M from the coded source specifies a signal point in a sphere about the signal $\rho\boldsymbol{X}$. A minimum-distance decoder for the information-embedding problem finds the nearest neighbor code vector to the channel observation, which corresponds to a decoded message index. The corresponding Wyner–Ziv encoder finds the nearest neighbor code vector to the source, and transmits the associated index.

Another aspect of the relationship between the information-embedding and Wyner–Ziv problems is the duality between the roles of noise and distortion in the two problems, which is readily seen in our geometric interpretation of the Gaussian case. In particular, from Fig. 6 we see that in the Wyner–Ziv problem the radius of the large sphere is proportional to $\sigma_{Y|X}$, which characterizes the noisiness of the channel in the Wyner–Ziv problem, and the radius of the smaller sphere is proportional to $\sqrt{d}$. In contrast, from Fig. 5, we see that in the case of information embedding the radius of the large sphere is essentially proportional to $\sqrt{d}$, and the radius of the smaller sphere is proportional to the standard deviation of the noise σ_V.

Note that this dual relationship between noise in one problem and distortion in the other is consistent with our observations in Section III-F of the duality in the characterizations of achievable rates between the noise-free and distortion-free scenarios in the two problems with finite alphabets.

E. Nested Lattice Code Constructions

Nested lattices can be used to construct optimum codes for the information-embedding and Wyner–Ziv problems in the Gaussian-quadratic scenario, as we describe in this section in the dual cases of high SDR and high SNR, respectively. The resulting codes are duals of one another.

Our notation is as follows. An (unbounded) n-dimensional lattice $\mathcal{L}$ is a set of codewords $\{\boldsymbol{l}_i\}$ such that

$$\boldsymbol{l}_i \in \mathbb{R}^n, \quad \boldsymbol{l}_0 = \boldsymbol{0}, \quad \boldsymbol{l}_i + \boldsymbol{l}_j \in \mathcal{L}, \qquad \forall i, j. \quad (28)$$

A minimum (Euclidean) distance decoder, which quantizes an arbitrary signal $\boldsymbol{X}$ to the nearest (in a Euclidean sense) codeword, takes the form

$$Q(\boldsymbol{U}) \triangleq \arg\min_{\boldsymbol{l}\in\mathcal{L}} ||\boldsymbol{U} - \boldsymbol{l}||^2 \quad (29)$$

where $||\cdot||$ denotes the (usual) Euclidean norm. The associated quantization error is then

$$\boldsymbol{E} = \boldsymbol{U} - Q(\boldsymbol{U}). \quad (30)$$

The quantizer specifies the characteristic Voronoi region

$$\mathcal{V}_i = \{\boldsymbol{U}\colon Q(\boldsymbol{U}) = \boldsymbol{0}\}$$

for the lattice. A Voronoi region is conveniently described in terms of its volume V, second moment σ^2, and normalized second moment G, which are given by, respectively,

$$V = \int_{\mathcal{V}} \mathrm{d}\boldsymbol{U}, \quad \sigma^2 = \frac{1}{nV}\int_{\mathcal{V}} \|\boldsymbol{U}\|^2 \,\mathrm{d}\boldsymbol{U}, \quad G = \frac{\sigma_i^2}{V^{2/n}}. \tag{31}$$

When $\mathcal{L}$ is a good lattice (i.e., constitutes a good source–channel code, in the sphere-covering/packing sense), n is sufficiently large, and we are operating in the limit of high resolution (high signal-to-quantization-error σ_U^2/σ^2), we have the following properties.[7]

(GQ-1) The quantization error (30) is white and Gaussian with zero mean and variance σ^2, and independent of (29), the codeword to which the vector is quantized [35].

(GQ-2) For every $\epsilon > 0$, the probability of a decoding error, $\Pr\{Q(\boldsymbol{l} + \boldsymbol{Z}) \neq \boldsymbol{l}\} < \epsilon$ when $\boldsymbol{l} \in \mathcal{L}$ and $\boldsymbol{Z}$ is a zero-mean white Gaussian vector independent of $\boldsymbol{l}$ whose elements have variance $\sigma^2 - \epsilon$ [13].

(GQ-3) For all $\epsilon > 0$, $\log(2\pi e G) < \epsilon$ [35].

We make use of two good lattices $\mathcal{L}_1$ and $\mathcal{L}_2$, where $\mathcal{L}_2$ is nested in $\mathcal{L}_1$, i.e., $\mathcal{L}_2 \subset \mathcal{L}_1$.[8] The associated quantizer, Voronoi cell, volume, second moment, and normalized second moment for the lattice $\mathcal{L}_i$ are denoted $Q_i(\cdot)$, $\mathcal{V}_i$, V_i, σ_i^2, and G_i.

The lattice $\mathcal{L}_1$ can be partitioned into $2^{nR} = V_2/V_1$ cosets corresponding to $\mathcal{L}_2$ and its translates. As in [36], for $\boldsymbol{l} \in \mathcal{L}_1$, we refer to the quantity $\boldsymbol{S} = \boldsymbol{l} - Q_2(\boldsymbol{l})$ as the coset shift of $\boldsymbol{L}$ with respect to the lattice $\mathcal{L}_2$. The function $k(\boldsymbol{L})$: $\mathcal{L}_1 \mapsto \{1, 2, \ldots, V_2/V_1\}$ indexes the coset shifts, and the inverse function is $\boldsymbol{g}(\cdot)$, i.e., $\boldsymbol{g}(k(\boldsymbol{l})) = \boldsymbol{l} - Q_2(\boldsymbol{l})$.

We let $\mathcal{L}_2^{\boldsymbol{S}}$ denote the coset corresponding to coset shift $\boldsymbol{S}$, and we note that the quantizer for this coset, $Q_2^{\boldsymbol{S}}$, takes the form

$$Q_2^{\boldsymbol{S}}(\boldsymbol{U}) = Q_2(\boldsymbol{U} - \boldsymbol{S}) + \boldsymbol{S}. \tag{32}$$

1) Nested Lattice Codes for Information Embedding: In this subsection, we construct a nested lattice implementation of distortion-compensated quantization index modulation (DC-QIM) [7]. The codes achieve information embedding capacity in the limit of high SDR (σ_Y^2/d). For this version we avoid the use of dither; versions that exploit dither are given in [1] and [18].

We choose our nested lattices such that

$$\sigma_1^2 = (1-b)^2\sigma_2^2 + \sigma_V^2 + \epsilon \quad \text{and} \quad \sigma_2^2 = \frac{d}{b^2} \tag{33}$$

where

$$b = \frac{d}{d + \sigma_V^2}. \tag{34}$$

Our information-embedding encoder using these lattices takes the form of DC-QIM [7], i.e., the composite signal $\boldsymbol{W}$ is constructed from the host $\boldsymbol{X}$ and the (unique) coset shift $\boldsymbol{S} = g(m)$ of the message m according to

$$\boldsymbol{W} = a\boldsymbol{X} + b\tilde{\boldsymbol{W}} \tag{35}$$

with

$$\tilde{\boldsymbol{W}} = Q_2^{\boldsymbol{S}}(\boldsymbol{X}) \tag{36}$$

where a is another parameter. The associated decoder produces the message estimate as the index of the closest coset to its observation $\boldsymbol{Y}$, i.e., $\hat{M} = k(Q_1(\boldsymbol{Y}))$.

We first verify that the embedding rate R is arbitrarily close to capacity, which follows from the lattice properties. Indeed, with the message M drawn uniformly from the indexes $\{1, 2, \ldots, 2^{nR}\}$, using Property (GQ-3) and (33), the rate of the system is within $1/n$ bits of

$$\begin{aligned} R &= \frac{1}{n}\log\left(\frac{V_2}{V_1}\right) = \frac{1}{2}\log\left(\frac{\sigma_2^2}{\sigma_1^2}\frac{G_1}{G_2}\right) \\ &\geq \frac{1}{2}\log\left(\frac{\sigma_V^2 + d}{\sigma_V^2}\right) - O(\epsilon) = C^{\mathrm{IE}}(d) - O(\epsilon) \end{aligned} \tag{37}$$

where the last equality follows from (22).

Furthermore, with the right choice of the parameter a, we can ensure the encoder meets the distortion constraint in the regime of interest. Indeed, defining the quantization error

$$\boldsymbol{E}_2 = \boldsymbol{X} - Q_2^{\boldsymbol{S}}(\boldsymbol{X}) \tag{38}$$

and letting $a = 1 - b$ we have that

$$\boldsymbol{W} = \boldsymbol{X} - b\boldsymbol{E}_2. \tag{39}$$

Applying Property (GQ-1) in the context of the lattice $\mathcal{L}_2$, we obtain that the embedding distortion is, using (39), as desired

$$\frac{1}{n}E[\|\boldsymbol{X} - \boldsymbol{W}\|^2] = b^2\sigma_2^2 = d. \tag{40}$$

Finally, it is straightforward to verify that the decoder achieves arbitrarily low error probability. Indeed

$$\begin{aligned} \boldsymbol{Y} &= \boldsymbol{W} + \boldsymbol{V} && (41)\\ &= (\boldsymbol{X} - \boldsymbol{E}_2) + ((1-b)\boldsymbol{E} + \boldsymbol{V}) && (42)\\ &= Q_2^{\boldsymbol{S}}(\boldsymbol{X}) + ((1-b)\boldsymbol{E}_2 + \boldsymbol{V}) && (43)\\ &= \tilde{\boldsymbol{W}} + \boldsymbol{Z} && (44) \end{aligned}$$

where (42) follows from (39), where (43) follows from (38), where $\tilde{\boldsymbol{W}}$ is as defined in (36), and where $\boldsymbol{Z} = (1-b)\boldsymbol{E}_2 + \boldsymbol{V}$. Now, using Property (GQ-1) in the context of lattice $\mathcal{L}_2$, we know that $\boldsymbol{E}_2$, and hence $\boldsymbol{Z}$, is Gaussian and independent of $\tilde{\boldsymbol{W}}$. Thus using Property (GQ-2) in the context of lattice $\mathcal{L}_1$ we have from (44) that $Q_1(\boldsymbol{Y}) = \tilde{\boldsymbol{W}}$ with probability at least $1 - \epsilon$ since $\mathrm{var}Z = \sigma_1^2 - \epsilon$.[9] But $k(\tilde{\boldsymbol{W}}) = k(\boldsymbol{S}) = M$, so the decoder estimates $k(Q_1(\boldsymbol{Y}))$ as M with probability at least $1 - \epsilon$.

2) Nested Lattice Code for Wyner–Ziv Encoding: Analogously, nested lattices can be used to build Wyner–Ziv codes for the Gaussian-quadratic case, as Zamir and Shamai develop with a dithered construction in [36] in the limit of high SNR (σ_Y^2/σ_V^2).[10] A generalization of this construction that achieves the rate-distortion limit for all SNRs is outlined in Appendix III and also appears in [1] and [37]. A version of this construction that avoids dither, which we summarize here, is the dual of that we consider in Section IV-E1. As the solution will reveal, the Wyner–Ziv encoder (respectively, decoder) has the same form

[7]These properties are true only in the asymptotic sense, which makes them somewhat hypothetical for any n. See [18] for a more rigorous treatment.

[8]The existence of pairs of good nested lattices is shown in [19], [37].

[9]That the overall $\boldsymbol{Z}$ can, for good lattices, be effectively treated as Gaussian with the indicated variance is also justified by more formal treatment of the underlying limits, as shown in [19].

[10]Such constructions are explored further by Servetto [27].

as the information-embedding decoder (respectively, encoder). It suffices to restrict our attention to the case $d < \sigma_{Y|X}^2$.

For this problem, the nested lattices are chosen such that

$$\sigma_1^2 = \frac{d\sigma_V^2}{(\sigma_V^2 - d)} \quad \text{and} \quad \sigma_2^2 = \sigma_1^2 + \sigma_V^2 + \epsilon. \tag{45}$$

A suitable encoder using these lattices transmits the index of the closet coset to the source $\boldsymbol{Y}$, i.e., it transmits $M = k(Q_1(\boldsymbol{Y}))$. The decoder observes $\boldsymbol{X}$ and M, calculates the coset shift $\boldsymbol{S} = g(M)$, then produces a source estimate of the form

$$\boldsymbol{W} = a\boldsymbol{X} + b\tilde{\boldsymbol{W}} \tag{46}$$

where

$$\tilde{\boldsymbol{W}} = Q_2^{\boldsymbol{S}}(\boldsymbol{X}). \tag{47}$$

That the system operates at the target rate follows from the lattice properties. Indeed, Property (GQ-3) and (45) prescribe the rate of the code to be within $1/n$ bits of

$$\begin{aligned} R &= \frac{1}{n}\log\left(\frac{V_1}{V_2}\right) = \frac{1}{2}\log\left(\frac{\sigma_2^2 G_1}{\sigma_1^2 G_2}\right) \\ &\le \frac{1}{2}\log\left(\frac{\sigma_V^2}{d}\right) + O(\epsilon) \\ &= \lim_{\mathrm{SNR}\to\infty} R_{Y|X}^{\mathrm{WZ}}(d) + O(\epsilon) \end{aligned} \tag{48}$$

where the last equality follows from (23) together with the fact that $\sigma_{Y|X}^2 = \sigma_Y^2\sigma_V^2/(\sigma_Y^2 + \sigma_V^2) \to \sigma_V^2$ as $\sigma_Y^2/\sigma_V^2 \to \infty$.

Next, to verify that the decoder reconstructs the source $\boldsymbol{Y}$ to within distortion d, we first define the quantization error

$$\boldsymbol{E}_1 = \boldsymbol{Y} - Q_1(\boldsymbol{Y}) \tag{49}$$

and express the received data $\boldsymbol{X}$ in the form

$$\boldsymbol{X} = Q_1(\boldsymbol{Y}) + \boldsymbol{E}_1 + \boldsymbol{V} = Q_1(\boldsymbol{Y}) + \boldsymbol{Z} \tag{50}$$

where we have defined $\boldsymbol{Z} = \boldsymbol{E}_1 + \boldsymbol{V}$. Now applying (GQ-1) in the context of[11] $\mathcal{L}_1$, and exploiting that $\boldsymbol{V}$ is independent of $\boldsymbol{Y}$ and, therefore, $Q_1(\boldsymbol{Y})$, we have that $\boldsymbol{Z}$ is (effectively) Gaussian and independent of $Q_1(\boldsymbol{Y})$. In turn, since $Q_1(\boldsymbol{Y}) \in \mathcal{L}_2^{\boldsymbol{S}}$, we can use (GQ-2) in the context of $\mathcal{L}_2^{\boldsymbol{S}}$ to obtain that, with probability at least $1 - \epsilon$

$$\tilde{\boldsymbol{W}} = Q_2^{\boldsymbol{S}}(\boldsymbol{X}) = Q_1(\boldsymbol{Y}) \tag{51}$$

since $\mathrm{var}Z = \sigma_2^2 - \epsilon$. In turn, substituting (51) into (46), we have that with probability $1 - \epsilon$

$$\boldsymbol{W} = a\boldsymbol{X} + bQ_1(\boldsymbol{Y}). \tag{52}$$

Choosing a and b so as to minimize the mean-square distortion between $\boldsymbol{W}$ and $\boldsymbol{Y}$, we obtain, using basic linear MMSE estimation theory, that the optimum a and b yield a mean-square estimation error of

$$\frac{1}{n}E\left[||\boldsymbol{W} - \boldsymbol{Y}||^2\right] = d + O(\epsilon) \tag{53}$$

which confirms the distortion constraint is met.

[11]Since Wyner–Ziv coding is nontrivial only when $d < \sigma_{Y|X}^2$, and since $\sigma_{Y|X}^2 < \sigma_V^2$, then our operating in the high-SNR regime implies we are also operating in the high-SDR regime.

V. Binary-Hamming Case

In this section, we consider the scenario where the signals of interest—the host in the information-embedding problem and the source in the Wyner–Ziv problem—are Bernoulli$(1/2)$ sequences, where Bernoulli(p) denotes a sequence of i.i.d. binary $(\in\{0, 1\})$ random variables, each of which is takes on the value 1 with probability p. In both problems, the associated channel of interest is the binary-symmetric channel with crossover probability p. The distortion metric $D(\cdot, \cdot)$ is Hamming metric, corresponding to bit-error rate. In this section, we use $h(\alpha)$ to denote the entropy of a Bernoulli(α) source, i.e.,

$$h(q) = -q\log(q) - (1-q)\log(1-q)$$

and $p * q$ to denote binary convolution, i.e.,

$$p * q = p(1-q) + q(1-p).$$

A. Binary-Hamming Information-Embedding Capacity

The information-embedding capacities in the binary-Hamming case are as follows.

Claim 3: For the binary-Hamming case, the distortion-constrained information-embedding capacity $C^{\mathrm{IE}}(d)$ is the upper concave envelope of the function

$$g_p^{\mathrm{IE}}(d) = \begin{cases} 0, & \text{if } 0 \le d < p, \\ h(d) - h(p), & \text{if } p \le d \le 1/2, \end{cases} \tag{54}$$

i.e.,

$$C^{\mathrm{IE}}(d) = \begin{cases} \dfrac{g_p^{\mathrm{IE}}(d_p)}{d_p}\, d, & \text{if } 0 \le d \le d_p, \\ g_p^{\mathrm{IE}}(d), & \text{if } d_p < d \le 1/2, \end{cases} \tag{55}$$

where $d_p = 1 - 2^{-h(p)}$.

Claim 4: For the binary-Hamming case, the distortion-constrained information-embedding capacity $C_{\mathrm{priv}}^{\mathrm{IE}}(d)$ is given by

$$C_{\mathrm{priv}}^{\mathrm{IE}}(d) = h(p * d) - h(p), \qquad 0 \le d \le 1/2. \tag{56}$$

Proofs of Claims 3 and 4 are developed in Appendixes IV-A and IV-B, respectively.[12] Fig. 7 illustrates $C^{\mathrm{IE}}(d)$ and $C_{\mathrm{priv}}^{\mathrm{IE}}(d)$ as a function of the distortion constraint for a channel transition probability of $p = 0.1$. Note that $C_{\mathrm{priv}}^{\mathrm{IE}}(d) > C^{\mathrm{IE}}(d)$ for all $0 < d < 1/2$. This is not surprising: it is easy to verify that (6) is not satisfied for d in this range.

B. Binary-Hamming Wyner–Ziv Rate-Distortion Function

The Wyner–Ziv rate-distortion function for this scenario is determined in [34] to be the lower convex envelope of the function

$$g_p^{\mathrm{WZ}}(d) = \begin{cases} h(p * d) - h(d), & \text{if } 0 \le d < p \\ 0, & \text{if } d = p \end{cases} \tag{57}$$

i.e.,

$$R_{Y|X}^{\mathrm{WZ}}(d) = \begin{cases} g_p^{\mathrm{WZ}}(d), & \text{if } 0 \le d \le d_p \\ g_p^{\mathrm{WZ}}(d_p)\left(1 - \dfrac{(d - d_p)}{(p - d_p)}\right), & \text{if } d_p < d \le p \end{cases} \tag{58}$$

[12]The proof of Claim 3 mirrors that for the corresponding Wyner–Ziv rate-distortion function in [34].

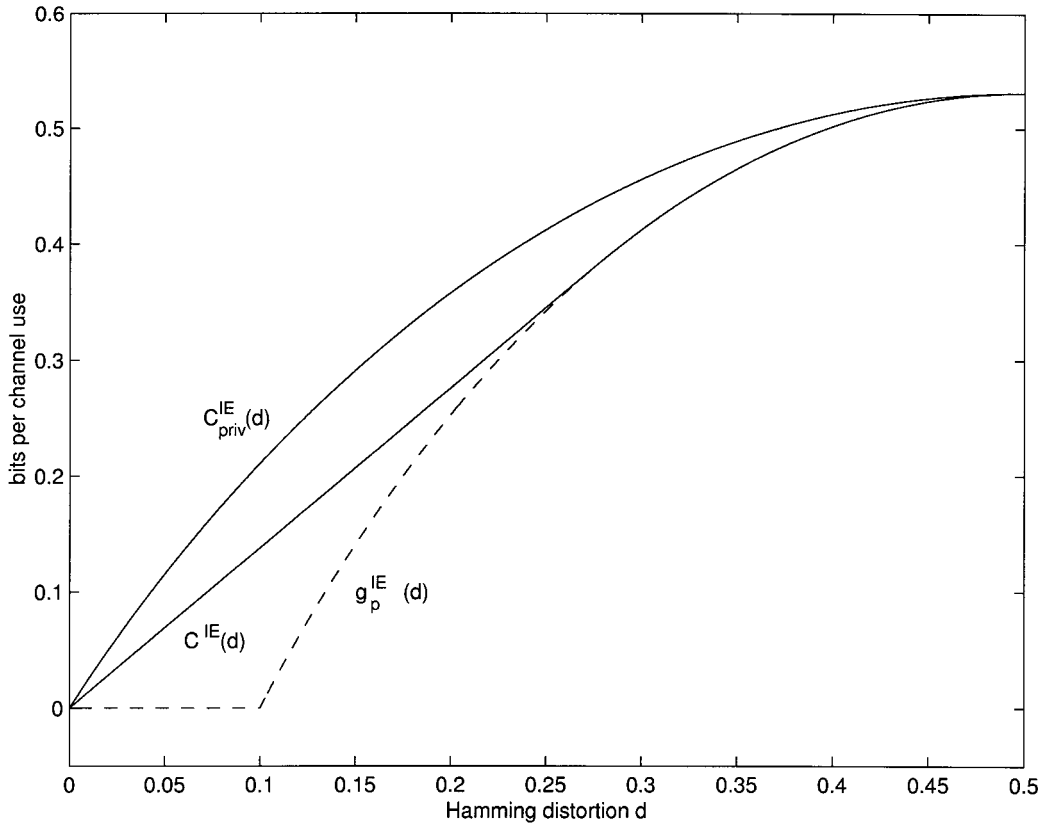

Fig. 7. The information-embedding capacities for the binary-Hamming case with channel transition probability $p = 0.1$. The dashed line is the function $g_p^{\mathrm{IE}}(d)$ from (54). The successively lower solid lines are $C_{\mathrm{priv}}^{\mathrm{IE}}(d)$ and $C^{\mathrm{IE}}(d)$, the information-embedding capacities with and without $\boldsymbol{X}$ known at the decoder, respectively.

where d_p is the solution to the equation

$$\frac{g_p^{\mathrm{WZ}}(d_p)}{d_p - p} = \dot{g}_p^{\mathrm{WZ}}(d_p) \tag{59}$$

with $\dot{}$ denoting the differentiation operator. For comparison, we show the conditional rate-distortion function ($\boldsymbol{X}$ known at the encoder and decoder) for the binary symmetric case [5]

$$R_{Y|X}(d) = \begin{cases} h(p) - h(d), & \text{if } 0 \le d \le p \\ 0, & \text{if } d \ge p. \end{cases} \tag{60}$$

Fig. 8 shows an example of $R_{Y|X}^{\mathrm{WZ}}(d)$ and $R_{Y|X}(d)$ for channel transition probability $p = 0.25$, which can be compared to Fig. 7.

C. Nested Binary Linear Codes

Optimum information embedding and Wyner–Ziv coding in the binary-Hamming case can be realized using a pair of nested binary linear codes, as we develop in this subsection.

Our code notation is as follows. A binary linear code $\mathcal{C}$ of 2^m codewords having length n is defined by a parity-check matrix of $\boldsymbol{H}$ dimension $m \times n$ with the property that

$$\boldsymbol{H}\boldsymbol{C}^{\mathrm{T}} = \boldsymbol{0}, \qquad \forall \boldsymbol{C} \in \mathcal{C} \tag{61}$$

where $^{\mathrm{T}}$ denotes the transpose operator. The syndrome of an arbitrary vector $\boldsymbol{X}$ is $\boldsymbol{H}\boldsymbol{X}^{\mathrm{T}}$. A minimum (Hamming) distance decoder, which quantizes an arbitrary signal $\boldsymbol{U}$ to the nearest (in a Hamming sense) codeword, takes the form

$$Q(\boldsymbol{U}) = \boldsymbol{U} \oplus f\left(\boldsymbol{H}\boldsymbol{U}^{\mathrm{T}}\right) \tag{62}$$

where $\oplus$ denotes modulo-2 addition, and where $f(\cdot)$ is the associated decoding function. The resulting quantization error is, therefore,

$$\boldsymbol{E} = \boldsymbol{U} \oplus Q(\boldsymbol{U}) = f\left(\boldsymbol{H}\boldsymbol{U}^{\mathrm{T}}\right). \tag{63}$$

Let $0 \le q < 1/2$ be determined from the code rate via $m/n = h(q)$. Then when $\mathcal{C}$ is a good code and n is sufficiently large, we have the following properties.

(BH-1) The quantization error (63) is Bernoulli(q) distributed and independent of (62), the codeword to which it is quantized.

(BH-2) For all codewords $\boldsymbol{C} \in \mathcal{C}$, the probability of a decoding error $\Pr\{Q(\boldsymbol{C}+\boldsymbol{Z}) \neq \boldsymbol{C}\}$ is small when $\boldsymbol{Z}$ is Bernoulli(q) distributed and independent of $\boldsymbol{C}$.

We make use of two good binary linear codes $\mathcal{C}_1$ and $\mathcal{C}_2$, where $\mathcal{C}_2$ is nested in $\mathcal{C}_1$, i.e., $\mathcal{C}_2 \subset \mathcal{C}_1$.[13] The associated code rate, parity-check matrix, quantizer, and decoding function for code $\mathcal{C}_i$ are denoted $m_i/n = h(q_i)$, $\boldsymbol{H}_i$, $Q_i(\cdot)$, and $f_i(\cdot)$, respectively. Note that because of the nesting, we can write

$$\boldsymbol{H}_2 = \begin{bmatrix} \boldsymbol{H}_1 \\ \boldsymbol{H}_a \end{bmatrix} \tag{64}$$

where $\boldsymbol{H}_a$ has dimension $(m_2 - m_1) \times n$. Furthermore, $\mathcal{C}_1$ can be partitioned into $2^{m_2 - m_1}$ cosets corresponding to $\mathcal{C}_2$ and its shifts.

[13]The existence of pairs of good nested linear codes is shown in [19], [37].

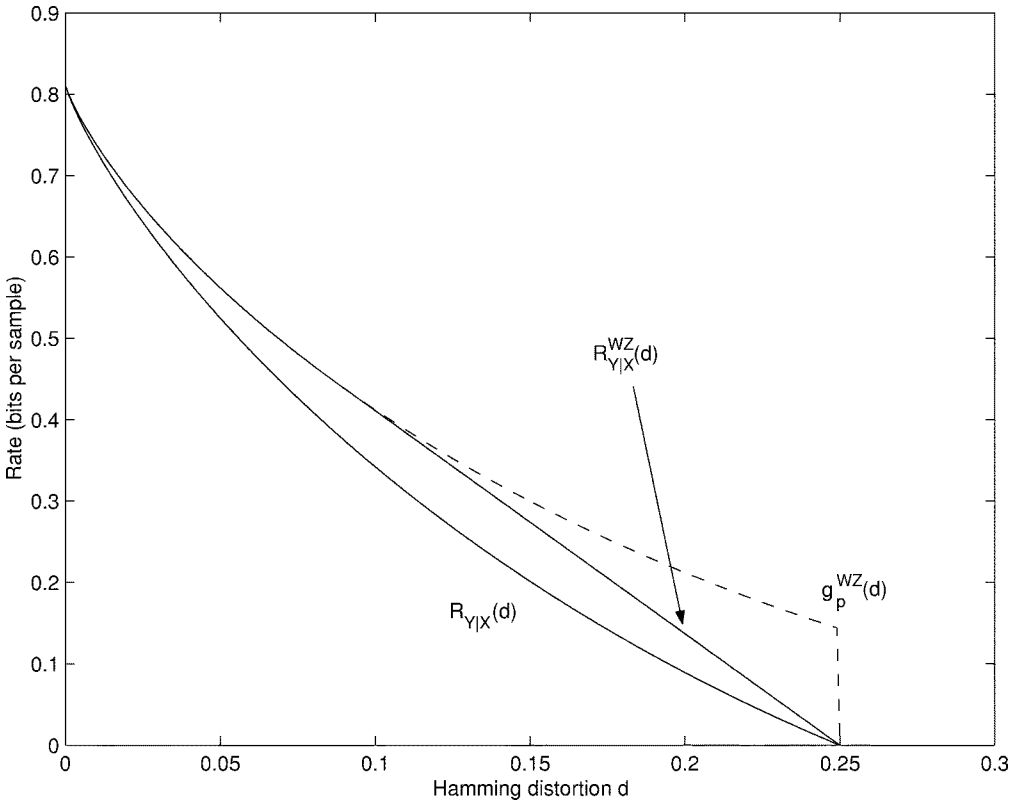

Fig. 8. The Wyner–Ziv rate-distortion functions for the binary-Hamming case with channel transition probability $p = 0.25$. The dashed line is the function $g_p^{\mathrm{WZ}}(d) = h(p * d) - h(d)$ from (57). The successively lower solid lines are $R_{Y|X}^{\mathrm{WZ}}(d)$ and $R_{Y|X}(d)$, the rate-distortion functions with and without $\boldsymbol{X}$ known at the encoder, respectively.

1) Nested Binary Codes for Information Embedding: Information embedding for the binary-Hamming case again takes the form of QIM [7]. However, in this binary-Hamming case, no distortion compensation is involved.

To develop the appropriate QIM scheme, we choose $q_1 = p$ and $q_2 = d$, so our code rates are $m_1/n = h(p)$ and $m_2/n = h(d)$. It suffices to restrict our attention to the region $d \geq d_p$, since lower distortions can be achieved through time sharing.

We let the rate of the information signal M be

$$R = C^{\mathrm{IE}}(d) = h(d) - h(p) = \frac{m_2 - m_1}{n} \tag{65}$$

and associate with each message M a (unique) coset shift $\boldsymbol{S} \in \mathcal{C}_1$ via the relation

$$\boldsymbol{H}_a \boldsymbol{S}^{\mathrm{T}} = \mathrm{Bin}(M) \tag{66}$$

where $\mathrm{Bin}(M)$ denotes a vector of length n whose elements are the binary expansion of M. The encoder then generates the composite signal $\boldsymbol{W} \in \mathcal{C}_1$ according to QIM

$$\boldsymbol{W} = Q_2(\boldsymbol{X} \oplus \boldsymbol{S}) \oplus \boldsymbol{S} \tag{67}$$

$$= \boldsymbol{X} \oplus f_2\left(\boldsymbol{H}_2 \boldsymbol{X}^{\mathrm{T}} \oplus \boldsymbol{H}_2 \boldsymbol{S}^{\mathrm{T}}\right) \tag{68}$$

where (68) follows from applying (62), and where

$$\boldsymbol{H}_2 \boldsymbol{S}^{\mathrm{T}} = \begin{bmatrix} \boldsymbol{0} \\ \mathrm{Bin}(M) \end{bmatrix} \tag{69}$$

using (66) and the fact that $\boldsymbol{S} \in \mathcal{C}_1$.

To confirm that the encoder meets the distortion constraint it suffices to note that by using Property (BH-1) in the context of the codebook $\mathcal{C}_2$, we obtain that the (quantization) error

$$\boldsymbol{E} = \boldsymbol{W} \oplus \boldsymbol{X} = (\boldsymbol{X} \oplus \boldsymbol{S}) \oplus Q_2(\boldsymbol{X} \oplus \boldsymbol{S}) \tag{70}$$

is Bernoulli(d).

The associated decoder operates as follows. The received signal is $\boldsymbol{Y} = \boldsymbol{W} \oplus \boldsymbol{V}$, where $\boldsymbol{V}$ is Bernoulli(p). Using Property (BH-2) in the context of codebook $\mathcal{C}_1$, we obtain that $\boldsymbol{W}$ can be recovered via

$$\boldsymbol{W} = Q_1(\boldsymbol{Y}) \tag{71}$$

with high probability. In turn, we use $\boldsymbol{W}$ to recover $\mathrm{Bin}(M)$ (and thus M) via

$$\mathrm{Bin}(M) = \boldsymbol{H}_a \boldsymbol{S}^{\mathrm{T}} = \boldsymbol{H}_a \boldsymbol{W}^{\mathrm{T}} \tag{72}$$

where the first equality is due to (66), and where the second is a consequence of (67), since $Q_2(\cdot)$ produces codewords in $\mathcal{C}_2$.

a) Noise-free case: Using (20), we easily determine that under the constraint that the composite signal Y be within Hamming distance d of the host X, the binary-Hamming embedding capacity is

$$C_{\text{noise-free}} = \max_{p_{Y|X}(y|x)} H(Y|X) = H(d). \tag{73}$$

To achieve rates arbitrarily close to the capacity (73), it therefore suffices to use the nested linear coding method for information embedding described in Section V-C1 with $p = 0$.

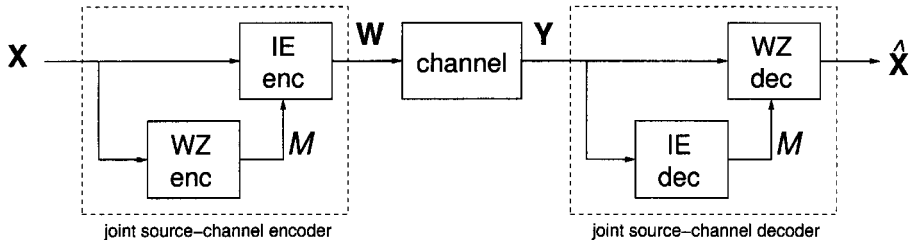

Fig. 9. A layered joint source–channel coding system.

2) Nested Binary Codes for Wyner–Ziv Coding: The corresponding nested codes for the Wyner–Ziv problem are developed by Shamai, Verdú, and Zamir [28] by setting $q_1 = d$ and $q_2 = h * d$, so the code rates are $m_1/n = h(d)$ and $m_2/n = h(p * d)$. To illustrate the duality with the information-embedding solution, we summarize the salient features of the construction in [28] here, again, restricting our attention to bit-error rates $0 \leq d \leq d_p$, as time sharing with no coding can achieve all other operating points on the capacity curve.

The encoder computes

$$\boldsymbol{S} = Q_1(\boldsymbol{Y}) \in \mathcal{C}_1 \tag{74}$$

and sends the length $m_2 - m_1$ vector (syndrome) $\boldsymbol{H}_a\boldsymbol{S}$, which describes the nearest coset of $\mathcal{C}_1$ to $\boldsymbol{Y}$. The rate of the encoder is thus

$$\frac{m_2 - m_1}{n} = h(p * d) - h(d) = R_{Y|X}^{\mathrm{WZ}}(d). \tag{75}$$

The associated decoder observes $\boldsymbol{H}_a\boldsymbol{S}^{\mathrm{T}}$ and $\boldsymbol{X}$, and reconstructs an estimate of the source as (c.f. (67), (68))

$$\boldsymbol{W} = Q_2(\boldsymbol{X} \oplus \boldsymbol{S}) \oplus \boldsymbol{S} \tag{76}$$

$$= \boldsymbol{X} \oplus f_2\left(\boldsymbol{H}_2\boldsymbol{X}^{\mathrm{T}} \oplus \boldsymbol{H}_2\boldsymbol{S}^{\mathrm{T}}\right) \tag{77}$$

where $\boldsymbol{H}_2\boldsymbol{S}^{\mathrm{T}}$ is constructed from the received side information via

$$\boldsymbol{H}_2\boldsymbol{S}^{\mathrm{T}} = \begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{H}_a\boldsymbol{S}^{\mathrm{T}} \end{bmatrix}. \tag{78}$$

Following Shamai *et al.* [28], the reconstruction $\boldsymbol{W}$ can be shown to meet the distortion constraint as follows. First, using Property (BH-1) in the context of the codebook $\mathcal{C}_1$, we obtain that the quantization error in $\boldsymbol{S}$, i.e.,

$$\boldsymbol{E} = \boldsymbol{S} \oplus \boldsymbol{Y} = Q_1(\boldsymbol{Y}) \oplus \boldsymbol{Y} \tag{79}$$

is Bernoulli(d). Next, expressing the channel output in the form

$$\boldsymbol{X} = \boldsymbol{Y} \oplus \boldsymbol{V} \tag{80}$$

where $\boldsymbol{V}$ is Bernoulli(p), we obtain, combining (79) with (80)

$$\boldsymbol{X} \oplus \boldsymbol{S} = \boldsymbol{E} \oplus \boldsymbol{V} \tag{81}$$

which is, therefore, Bernoulli$(p*d)$. But then applying Property (BH-2) to (81) in the context of the codebook $\mathcal{C}_2$, we have that with high probability

$$\boldsymbol{W} = Q_2(\boldsymbol{X} \oplus \boldsymbol{S}) \oplus \boldsymbol{S} = Q_2(\boldsymbol{E} \oplus \boldsymbol{V}) \oplus \boldsymbol{S} = \boldsymbol{0} \oplus \boldsymbol{S} = \boldsymbol{S}. \tag{82}$$

Thus, we obtain that the reconstruction error is with high probability

$$\boldsymbol{W} \oplus \boldsymbol{Y} = \boldsymbol{S} \oplus \boldsymbol{Y} = \boldsymbol{E}, \tag{83}$$

which is Bernoulli(d) as required.

b) Distortion-free case: When we set $d = 0$ in the nested code construction of Section V-C2, the code not surprisingly specializes to the well-known practical Slepian–Wolf code developed by Wyner [32].[14] From this perspective, we can see that the nested linear code for information embedding in the noise-free case described at the end of Section V-C1 is the dual of Wyner's Slepian–Wolf code, i.e., the encoder in one case is the decoder for the other, and *vice versa*.

VI. Layered Joint Source–Channel Coding

The relationship between information embedding and Wyner–Ziv coding developed in this paper can be exploited in the development of a variety of novel systems. As one illustration, in this section we introduce a layered joint source–channel coding system.

Such a system can be formed from the interconnection of Wyner–Ziv and information-embedding subsystems. A simple two-layer implementation is depicted in Fig. 9. As this figure reflects, in this system the bits comprising the Wyner–Ziv representation M of the source $\boldsymbol{X}$ are embedded into the source using information embedding to produce a transmitted signal $\boldsymbol{W}$, where the Wyner–Ziv encoding takes into account the additional degradation of the source (beyond that introduced in the channel) that will result from the embedding.[15] The associated decoder operates on both layers of the received signal $\boldsymbol{Y}$ as also shown in Fig. 9. It extracts the bits of the Wyner–Ziv representation M using the information-embedding decoder, and uses them in the Wyner–Ziv decoder to reconstruct the estimate $\hat{\boldsymbol{X}}$ of the source. Note the interesting property that encoder and decoder for this system have identical structure, which follows from the fact that the structure of the information-embedding encoder is the same as that for the Wyner–Ziv decoder, and *vice versa*.

Such a system has the feature that can be used in a broadcast setting involving two classes of receivers: private receivers, to which the Wyner–Ziv and information-embedding codebooks are revealed, and public receivers, which have no codebook information. Thus, public receivers construct an estimate $\hat{\boldsymbol{X}}_{\mathrm{pub}}$ of the base layer, without decoding the embedded information, while private receivers construct the estimate $\hat{\boldsymbol{X}}$ from both layers.

By varying the Wyner–Ziv bit rate within the encoder (and adjusting the private decoder parameters accordingly), one

[14]There has been renewed interest in implementations of Wyner's construction lately; see, e.g., [26].

[15]Note that implicit in our assumption of a discrete-time source, all such codes we develop use the same bandwidth as the source.

can control the quality of the public estimate $\boldsymbol{Y}$: the higher the Wyner–Ziv bit rate, the lower the quality of the public estimate. In the sequel, we examine how the quality of the private estimate varies as this bit rate is varied. We refer to a system as "efficient" when the quality of the private estimate is independent of the chosen public estimate quality. Two systems will be examined: one for the Gaussian-quadratic scenario, and one for the binary-Hamming scenario.

A. The Gaussian-Quadratic Case

In this subsection, we construct a layered joint source–channel code that is efficient for the Gaussian-quadratic case. Let our i.i.d. Gaussian source $\boldsymbol{X}$ have elements distributed according to $\mathcal{N}(0, \sigma_X^2)$, let the independent additive white Gaussian noise $\boldsymbol{V} = \boldsymbol{Y} - \boldsymbol{W}$ in the channel have elements distributed according to $\mathcal{N}(0, \sigma_V^2)$. Our implementation uses the information-embedding and Wyner–Ziv subsystems in precisely the forms developed in Sections IV-A and IV-B.

When we embed under an embedding distortion constraint σ_E^2 using a capacity-achieving code, the embedding adds noise $\boldsymbol{E} \sim \mathcal{N}(\boldsymbol{0}, \sigma_E^2 \boldsymbol{I})$ that is independent of $\boldsymbol{X}$. In order to normalize the overall transmitted power to σ_X^2, the host $\boldsymbol{X}$ must be scaled by

$$\mu = \sqrt{\frac{\sigma_X^2 - \sigma_E^2}{\sigma_X^2}} \tag{84}$$

prior to embedding.

At the receiver, the observed signal is

$$\boldsymbol{Y} = \boldsymbol{W} + \boldsymbol{V} = (\mu \boldsymbol{X} + \boldsymbol{E}) + \boldsymbol{V} \tag{85}$$

from which we see using MMSE estimation theory that the best public receiver estimate is

$$\hat{\boldsymbol{X}}_{\text{pub}} = \frac{\mu \sigma_X^2}{\sigma_X^2 + \sigma_V^2} \cdot \boldsymbol{Y} \tag{86}$$

and yields distortion

$$d_{\text{pub}} = \sigma_{X|Y}^2 = \sigma_X^2 \cdot \frac{\sigma_V^2 + \sigma_E^2}{\sigma_V^2 + \sigma_X^2}. \tag{87}$$

Thus, as σ_E^2 is varied from 0 to σ_X^2, d_{pub} varies from

$$d_{\min} = \frac{\sigma_X^2 \sigma_V^2}{\sigma_X^2 + \sigma_V^2} \tag{88}$$

to σ_X^2, corresponding to the observation carrying no useful (public) information about $\boldsymbol{X}$.

To obtain the distortion d of the private decoder, we note that with σ_E^2 fixed, the maximum achievable embedding rate is $C^{\text{IE}}(\sigma_E^2)$. Given this supplied data rate, we Wyner–Ziv-encode for minimum distortion at the decoder subject to the available embedding rate, i.e., the resulting distortion d is the solution to

$$\frac{1}{2} \log\left(1 + \frac{\sigma_E^2}{\sigma_V^2}\right) = C^{\text{IE}}(\sigma_E^2) = R_{X|Y}^{\text{WZ}}(d) = \frac{1}{2} \log\left(\frac{\sigma_{X|Y}^2}{d}\right) \tag{89}$$

which upon substitution of (87) for $\sigma_{X|Y}^2$ is easily verified to be given by[16] (88) independent of the embedding level σ_E^2 for all $0 \le \sigma_E^2 \le \sigma_X^2$.

Efficiency follows immediately. In particular, note that the choice $\sigma_E^2 = \sigma_X^2$ corresponds to a single-layer, fully private separate source and channel coding system, which by the source–channel separation theorem we know is the lowest possible distortion achievable by any system. That (88) is independent of σ_E^2 means, therefore, is that this layered joint source–channel coding system is efficient for *all* choices of σ_E^2. Consistent with our analysis, this includes the other extreme case ($\sigma_E^2 = 0$), which corresponds to single-layer uncoded (fully public) transmission, whose efficiency in the Gaussian-quadratic scenario is well-known [5, Sec. 5.2].

Multilayer Joint Source–Channel Codes: The two-layer joint source–channel coding scheme just described generalizes naturally to a multiple-layer scheme involving successive embeddings at the encoder. Such a system can be used to support nested classes of private users, each able to recover a progressively better estimate of the source.

The encoding for a $(t+1)$-layer system is generated from t successive embeddings at distortion levels σ_i^2, producing the sequence of composite signals $\boldsymbol{W}_i$, $i = 1, 2, \ldots, t$. In particular, at each layer i, the composite signal $\boldsymbol{W}_i$ is generated by embedding the bits of the associated Wyner–Ziv encoding of the preceding composite signal $\boldsymbol{W}_{i-1}$ into itself. The final composite signal $\boldsymbol{W} = \boldsymbol{W}_t$ is transmitted over the channel. In each embedding, the amplitude is renormalized to keep each composite signal at power σ_X^2. The composite signals thus created can, therefore, be expressed in the form

$$\begin{aligned} \boldsymbol{W}_0 &= \boldsymbol{X} \\ \boldsymbol{W}_i &= \mu_i \boldsymbol{W}_{i-1} + \boldsymbol{E}_i, \qquad i = 1, 2, \ldots, t \end{aligned} \tag{90}$$

where

$$\mu_i = \sqrt{\frac{\sigma_X^2 - \sigma_i^2}{\sigma_X^2}} \tag{91}$$

and $\boldsymbol{E}_i \sim \mathcal{N}(\boldsymbol{0}, \sigma_i^2 \boldsymbol{I})$, independent of $\boldsymbol{W}_{i-1}$, for $i = 1, 2, \ldots, t$.

The received signal is decoded as follows. There are t codebooks $\mathcal{C}_i$, $i = 1, 2, \ldots, t$, of which the last r are available to the rth class of (private) decoders. The tth embedding is extracted from the channel output $\boldsymbol{Y} = \hat{\boldsymbol{W}}_t$ by the information-embedding decoder using codebook $\mathcal{C}_t$, and the bits are used to form an estimate $\hat{\boldsymbol{W}}_{t-1}$ of $\boldsymbol{W}_{t-1}$ via the associated Wyner–Ziv decoder. By the analysis in Section VI-A, the distortion in the estimate so produced is given by (88). We proceed to form an estimate $\hat{\boldsymbol{W}}_{t-2}$ from the preceding composite signal estimate $\hat{\boldsymbol{W}}_{t-1}$, where the distortion in this estimate remains (88). This process is continued until $\hat{\boldsymbol{W}}_{t-r}$ is formed by decoding with codebook $\mathcal{C}_{t-r}$.

If all t codebooks are available to the decoder, i.e., the decoder is in the tth class, it follows that the source reconstruction $\hat{\boldsymbol{X}} = \hat{\boldsymbol{W}}_0$ achieves the best possible fidelity, i.e., (88). Thus,

[16]Note that this result is consistent with the broadcast channel result in [7] showing that the layered digital coding method involving the embedding of bits into a host that is itself a coded bit stream achieves capacity for the Gaussian channel.

the multilayer embedding continues to be efficient in the sense that no other alternative coding scheme for the channel could do better.

It remains only to analyze the performance experienced by the other classes of decoders. To simplify the exposition, we restrict our attention to the case of equal embedding distortion at each layer, i.e., $\sigma_i^2 = \sigma_E^2$ for $i = 1, 2, \ldots, t$, so that, via (84), we have

$$\mu_i = \mu = \sqrt{(\sigma_X^2 - \sigma_E^2)/\sigma_X^2}, \qquad i = 1, 2, \ldots, t. \tag{92}$$

The rth class of decoders, which can decode down to the $(t-r)$th layer, obtain $\hat{\boldsymbol{W}}_{t-r}$, which can be expressed using the results of Section VI-A as

$$\lambda^{-r}\boldsymbol{Y}_{t-r} = \lambda^{-r}\hat{\boldsymbol{W}}_{t-r} = \boldsymbol{W}_{t-r} + \tilde{\boldsymbol{V}}_{t-r} \tag{93}$$

where $\tilde{\boldsymbol{V}}_{t-r} \sim \mathcal{N}(\boldsymbol{0}, \sigma_V \boldsymbol{I})$ is independent of $\boldsymbol{W}_{t-r}$, and where

$$\lambda = \frac{\sigma_X^2}{\sigma_X^2 + \sigma_V^2}. \tag{94}$$

Expanding $\boldsymbol{W}_{t-r}$ according to the iteration (90), we have

$$\lambda^{-r}\boldsymbol{Y}_{t-r} = \mu^{t-r}\boldsymbol{X} + \sum_{i=0}^{t-r-1} \mu^i \boldsymbol{E}_{t-r-i} + \tilde{\boldsymbol{V}}_{t-r}, \tag{95}$$

where $\boldsymbol{X}$, $\tilde{\boldsymbol{V}}_{t-r}$, and $\boldsymbol{E}_1, \boldsymbol{E}_2, \ldots \boldsymbol{E}_{t-r}$ are mutually independent and Gaussian. Thus, the rth class of decoders estimate $\boldsymbol{X}$ as

$$\hat{\boldsymbol{X}} = \mu^{t-r}\lambda^{1-r}\boldsymbol{Y}_{t-r} \tag{96}$$

and the associated distortion these users experience corresponds to the error in the associated MMSE estimate of $\boldsymbol{X}$ from $\boldsymbol{Y}_{t-r}$, i.e.,

$$d_r = \sigma_X^2 \cdot \frac{\sigma_V^2 + \sigma_X^2\left(1 - \mu^{2(t-r)}\right)}{\sigma_V^2 + \sigma_X^2} \tag{97}$$

which decays exponentially with r, the number of codebooks available to the receiver. The time constant of the decay increases linearly with $\log \mu^2$. In turn, μ^2 decreases linearly with σ_E^2, the embedding distortion for an individual layer. When $(t-r) \to \infty$ (which requires that $t \to \infty$), $d_r \to \sigma_X^2$ for all r. More generally, different d_r versus r profiles can be obtained by choosing the σ_i^2 to vary with i.

B. Binary-Hamming Case

A layered joint source-channel coding system of the form of Fig. 9 can also be developed for the binary-Hamming case. We consider an implementation, analogous to that for the Gaussian-quadratic case, in which we use the information-embedding and Wyner–Ziv subsystems in precisely the forms developed in Sections V-A and V-B. In the sequel, we use p to denote the crossover probability of the binary-symmetric channel.

Let us evaluate the achievable distortions. As shown in Appendix IV-A, the information-embedding capacity of $C^{\mathrm{IE}}(q)$ is achieved with a distortion of q that acts on the source $\boldsymbol{X}$ as a binary-symmetric channel with crossover probability q. Thus, the combined effect of the embedding and the physical channel will be a binary-symmetric channel with crossover probability $q * p$, so that for the Wyner–Ziv encoding, the side information is the source corrupted by a Bernoulli$(q * p)$ process. Thus, the best (public) estimate of the source in this case is $\hat{\boldsymbol{X}}_{\mathrm{pub}} = \boldsymbol{Y}$, and the associated distortion is

$$d_{\mathrm{pub}} = p * q \tag{98}$$

so that as q is varied from 0 to $1/2$, d_{pub} varies from

$$d_{\min} = p \tag{99}$$

to $1/2$, corresponding to the observation carrying no useful (public) information about $\boldsymbol{X}$.

Meanwhile, the achieved private distortion d is the solution to

$$C^{\mathrm{IE}}(q) = R_{X|Y}^{\mathrm{WZ}}(d) \tag{100}$$

where the left-hand side of (100) is the upper concave envelope of the function $g_p^{\mathrm{IE}}(q)$ as defined in (54), and the right-hand side of (100) is the lower convex envelope of the function $g_{p*q}^{\mathrm{WZ}}(d)$ as defined in (57).

The distortion in two limiting cases can be evaluated in closed form. In the case $q = 1/2$, which corresponds to single-layer, fully private, separate source and channel coding system, (100) specializes to

$$1 - h(p) = C^{\mathrm{IE}}(1/2) = R_{X|\emptyset}^{\mathrm{WZ}}(d) = 1 - h(d) \tag{101}$$

yielding $d = p$. By the source–channel separation theorem, this is the best distortion one can achieve using any system on this channel. The other limiting case for which $q = 0$, corresponding to a single-layer, fully public uncoded system, the distortion $d = p$ is obviously also achievable simply using the received data as the source estimate.

More generally, the resulting (normalized) distortion d is plotted in Fig. 10 as a function of q for various values of p. Note that while the system is efficient for the limiting cases $q = 0$ and $q = 1/2$, it is not in between: the distortion is strictly greater than (99) for all $0 < q < 1/2$ and all $p < 1/2$. This fact is proven analytically in [1].

Part of the reason for the inefficiency may lie in the fact that in the encoder of our system: 1) the chosen information-embedding encoder subsystem does not take into account the correlation between the source $\boldsymbol{X}$ and the message M; and 2) the chosen Wyner–Ziv encoder does not take into account partial knowledge it has of the ultimate channel output $\boldsymbol{Y}$ in the form of $\boldsymbol{W}$. Clearly, in the Gaussian-quadratic case nothing can be gained by exploiting such partial side information, since the system was efficient. However, in the binary-Hamming case, taking them into account could lead to a system that is efficient for all q and p, though that remains to be investigated.

VII. Concluding Remarks

In this paper, we identified and developed the inherent duality between information-embedding and Wyner–Ziv coding, and used this relationship to establish a variety of new results on the performance limits of information-embedding and deterministic nested codes for achieving them. As an illustration of other applications of these results, a layered joint source–channel coding system was developed with a symmetric encoder–decoder structure, and evaluated in the context of a broadcast setting in which

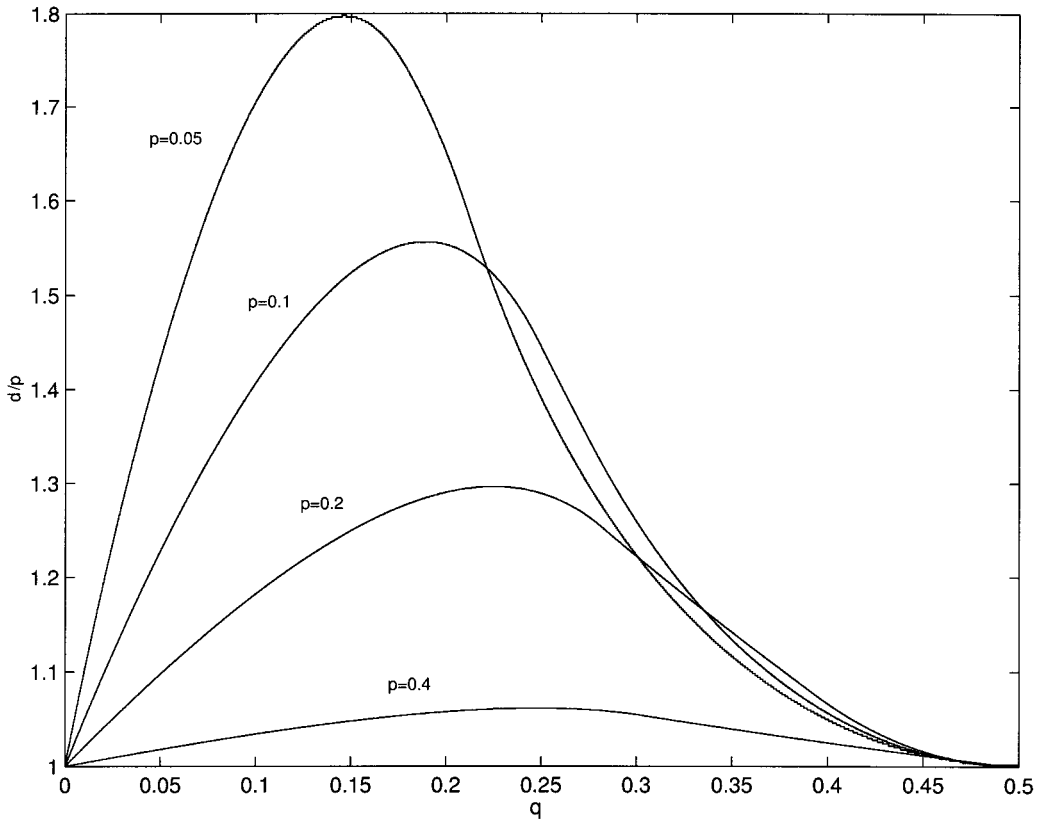

Fig. 10. Performance of layered source–channel coding in the binary-symmetric case. Plotted is the reconstruction distortion (normalized by p) as a function of embedding distortion for $p = 0.05, 0.1, 0.2, 0.4$.

there is a need to control the fidelity available to different receivers. Efficiency was evaluated in the context of channels for which the source–channel separation theorem holds, but still more interesting results may follow from examining its behavior in context where it does not.

More generally, in many respects, our results are simply representative examples of a considerably broader set of results that may ultimately evolve from the relationship between the information-embedding and Wyner–Ziv problems and exploring such directions is a rich area for future research.

Appendix I
Proof of Converse in Claim 1 (Public Embedding Capacity)

We show that for any rate $R \geq C^{\mathrm{IE}}(d)$, the maximal probability of error for a length n code, $P_{\mathrm{e}}^{(n)}$, is bounded away from zero. We begin with two useful lemmas.

Lemma 1: The capacity $C^{\mathrm{IE}}(d)$ is a nondecreasing concave function of d.

Proof: First, that $C^{\mathrm{IE}}(d)$ is a nondecreasing function follows from the fact that increasing d increases the domain over which the maximization is performed.

To establish concavity, consider any two distortions d_1 and d_2 and the corresponding arguments, U_1, f_1 and U_2, f_2, respectively, which maximize the argument of (1) for the given distortion. Let Q be a random variable independent of X, Y, U_1, and U_2, that takes on the value 1 with probability λ and the value 2 with probability $1-\lambda$. Define $Z = (Q, U_q)$ and let $f(Z, X) = f_Q(U_Q, X)$, implying a distortion

$$d = E[D(X, W)] \tag{102}$$
$$= \lambda E[D(X, f_1(U_1, X)) + (1-\lambda)E[D(X, f_2(U_2, X))] \tag{103}$$
$$= \lambda d_1 + (1-\lambda)d_2 \tag{104}$$

and

$$I(Z; Y) - I(Z; X) \tag{105}$$
$$= H(Y) - H(Y|Z) - H(X) + H(Y|Z) \tag{106}$$
$$= H(Y) - H(Y|U_Q, Q) - H(X) + H(X|U_Q, Q) \tag{107}$$
$$= H(Y) - \lambda H(Y|U_1) - (1-\lambda)H(Y|U_2) - H(X) + \lambda H(X|U_1) + (1-\lambda)H(X|U_2) \tag{108}$$
$$= \lambda(I(U_1; Y) - I(U_1; X)) + (1-\lambda)(I(U_2; Y) - I(U_2; X)). \tag{109}$$

Thus,

$$C^{\mathrm{IE}}(d) = \max_{U,\, f:\, E[D(X, f(U, X))] \leq d} (I(Y; U) - I(U; X)) \tag{110}$$
$$\geq I(Z; Y) - I(Z; X) \tag{111}$$
$$= \lambda(I(U_1; Y) - I(U_1; X)) + (1-\lambda)(I(U_2; Y) - I(U_2; X)) \tag{112}$$
$$= \lambda C^{\mathrm{IE}}(d_1) + (1-\lambda)C^{\mathrm{IE}}(d_2) \tag{113}$$

proving the concavity of $C^{\mathrm{IE}}(d)$. □

Gel'fand and Pinsker [20] show that in the absence of a distortion constraint, one cannot do better than (1) with a nonsingular distribution $p_{W|U,X}$. The same is true with a distortion constraint present. The following proof is due to Cohen [11].

Lemma 2 (Cohen): For a fixed p_X and $p_{Y|W,X}$,

$$\sup_{p_{U|X},\, p_{W|U,X},\, E[D(X,W)]\le d} I(Y;U) - I(U;X) = \sup_{p_{U|X},\, f:\mathcal{U}\times\mathcal{X}\mapsto\mathcal{W}, E[D(X,W)]\le d} I(Y;U) - I(U;X) \quad (114)$$

where $p_{W|U,X}(w|u,x) = 1_{\{w=f(u,x)\}}$ on the right-hand side.

Proof: To show that any nondeterministic $p_{W|U,X}(w|u,x)$ has at best the performance of a deterministic distribution, consider any such $p_{W|U,X}(w|u,x)$. Then there exists $u_0 \in \mathcal{U}$ such that $0 < p_{W|U,X}(w|u_0,x) < 1$ for some x and w. Define n and functions $f_1, \ldots, f_n : \mathcal{X} \mapsto \mathcal{W}$, and positive constants $c_1, \ldots, c_n$ with $\sum_i c_i = 1$ such that

$$p_{W|UX}(w|u_0,x) = \sum_{i=1}^{n} c_i 1_{\{w=f_i(x)\}}, \qquad \forall\, w\in\mathcal{W},\, x\in\mathcal{X}. \quad (115)$$

We show by a simple construction that a sufficient size for n is $|\mathcal{X}||\mathcal{W}|$. We let $\mathcal{W} = \{w_1, \ldots, w_{|\mathcal{W}|}\}$ and $\mathcal{X} = \{x_1, \ldots, x_{|\mathcal{X}|}\}$. We define the variables

$$b_{jk} = \sum_{m=1}^{j} p_{W|UX}(w_m|u_0,x_k), \qquad j = 1,\ldots,|\mathcal{W}|,\ k = 1,\ldots,|\mathcal{X}|$$

and place them, along with $q_0 = 0$, in an ordered set of nondecreasing (possibly repeating) values, $\mathcal{Q} = \{q_0, \ldots, q_n\}$, where $n = |\mathcal{X}||\mathcal{W}|$. We let $c_i = q_i - q_{i-1}$, $i = 1, \ldots, n$. Corresponding to each q_i is a b_{jk}, from which we define $f_i(x_r) = w_m$, $r = 1, \ldots |\mathcal{X}|$, where m is the smallest index for which $b_{mr} \ge b_{jk}$. These definitions satisfy (115) and $n = |\mathcal{X}||\mathcal{W}|$, which is finite.

Continuing with the proof of the lemma, we define a new alphabet $\tilde{\mathcal{U}} = \mathcal{U}' \cup \mathcal{U}\backslash\{u_0\}$ where $\mathcal{U}' = \{u_1', \ldots, u_n'\}$, and let a new auxiliary random variable $\tilde{U}$ take values in $\mathcal{U}'$ and have joint distributions

$$p_{W|\tilde{U},X}(w|\tilde{u},x) = \begin{cases} p_{W|U,X}(w|\tilde{u},x), & \text{if } \tilde{u}\in\mathcal{U}\backslash\{u_0\} \\ 1_{\{w=f_i(x)\}}, & \text{if } \tilde{u} = u_i'\in\mathcal{U}' \end{cases} \quad (116)$$

and

$$p_{\tilde{U}|X}(\tilde{u}|x) = \begin{cases} p_{U|X}(\tilde{u}|x), & \text{if } \tilde{u}\in\mathcal{U}\backslash\{u_0\} \\ c_i p_{U|X}(u_0|x), & \text{if } \tilde{u} = u_i'\in\mathcal{U}'. \end{cases} \quad (117)$$

It is straightforward to verify that the joint distribution on Y, W, and X is the same under the original and new auxiliary random variable choices, i.e.,

$$\sum_{\tilde{u}\in\mathcal{U}'} p_{Y|W,X}(y|w,x) p_{W|\tilde{U},X}(w|\tilde{u},x) p_{\tilde{U}|X}(\tilde{u},x) p_X(x) = \sum_{u\in\mathcal{U}} p_{Y|W,X}(y|w,x) p_{W|U,X}(w|u,x) p_{U|X}(u,x) p_X(x). \quad (118)$$

Thus, both $H(Y)$ and $E[D(X,W)]$ are unchanged by switching to the new auxiliary random variable.

If, in addition, the following joint distribution between U, $\tilde{U}$, and X is defined via

$$p_{\tilde{U}|U,X}(\tilde{u}|u,x) = \begin{cases} 1_{\{\tilde{u}=u\}}, & \text{if } \tilde{u}\in\mathcal{U}\backslash\{u_0\} \\ c_i 1_{\{u=u_0\}}, & \text{if } \tilde{u} = u_i'\in\mathcal{U}' \end{cases} \quad (119)$$

which is consistent with (117), then $X \leftrightarrow U \leftrightarrow \tilde{U}$ form a Markov chain. Thus, by the data-processing inequality

$$I(U;X) \ge I(\tilde{U};X). \quad (120)$$

Moreover, since from (115)–(117), we have

$$p_{Y|U}(y|u_0) = \sum_{i=1}^{n} c_i p_{Y|\tilde{U}}(y|u_i'). \quad (121)$$

Thus, by the concavity of entropy we have $H(Y|\tilde{U}) \le H(Y|U)$, which together with the fact that $H(Y)$ is unchanged yields

$$I(Y;\tilde{U}) \ge I(U;Y). \quad (122)$$

Combining (120) with (122) we see that

$$I(Y;\tilde{U}) - I(\tilde{U};X) \ge I(Y;U) - I(U;X). \quad (123)$$

Thus, $\tilde{U}$ is an optimal choice of random variable, whose alphabet $\tilde{\mathcal{U}}$ has one less element than $\mathcal{U}$ for which $p_{W|U,X}$ is nondeterministic. Recursive application of this logic for all $u \in \mathcal{U}$ such that $0 < p_{W|U,X}(w|u,x) < 1$ yields an auxiliary random variable $\tilde{U}$ that is optimal and for which $p_{W|U,X}$ is deterministic. □

Despite the fact that the repeated application of the logic in Lemma 2 will increase the cardinality of the auxiliary random variable, the final $|\tilde{\mathcal{U}}|$ is bounded above. Straightforward application of Caratheodory's theorem tells us that for the original U with nondeterministic $p_{W|U,X}$, we have $|\mathcal{U}| \le |\mathcal{X}| + |\mathcal{W}| + 2$. Since we apply the recursive argument at most $|\mathcal{U}|$ times, we have

$$|\tilde{\mathcal{U}}| \le |\mathcal{U}| n \le (|\mathcal{X}| + |\mathcal{W}| + 2)|\mathcal{X}||\mathcal{W}|$$

which is a finite upper bound on the cardinality.

Returning to our proof of the converse, consider an information-embedding code, with an encoding function f_n: $\mathcal{X}^n \times \{1, 2, \ldots, 2^{nR}\} \mapsto \mathcal{W}^n$ and a decoding function g_n: $\mathcal{Y}^n \mapsto \{1, 2, \ldots, 2^{nR}\}$. Let $f_{n,i}$: $\mathcal{X}^n \times \{1, 2, \ldots, 2^{nR}\} \mapsto \mathcal{W}$ denote the ith symbol produced by the encoding function. The distortion constraint is

$$\frac{1}{n} E\left[\sum_{i=1}^{n} D(X_i, f_{n,i}(X^n, M))\right] \le d. \quad (124)$$

We have the following chain of inequalities:

$$nR = H(M) = I(M;Y^n) + H(M|Y^n) \quad (125)$$
$$= I(M;Y^n) - I(M;X^n) + H(M|Y^n) \quad (126)$$
$$\le \sum_{i=1}^{n} [I(Z_i;Y_i) - I(Z_i;X_i)] + H(M|Y^n) \quad (127)$$

$$\le \sum_{i=1}^{n} C^{\mathrm{IE}}(E[D(X_i, f'_{n,i}(X_i, Z_i))]) + H(M|Y^n) \quad (128)$$

$$\le nC^{\mathrm{IE}}\left(E\left[\frac{1}{n}\sum_{i=1}^{n} D(X_i, f'_{n,i}(X_i, Z_i))\right]\right) + H(M|Y^n) \quad (129)$$

$$\le nC^{\mathrm{IE}}(d) + H(M|Y^n) \quad (130)$$

$$\le nC^{\mathrm{IE}}(d) + P_{\mathrm{e}}^{(n)} nR + 1 \quad (131)$$

where

(125) follows from the fact that M is distributed uniformly on $\{1, 2, \ldots, 2^{nR}\}$ from our formulation;

(126) follows from the fact that $I(M; X^n) = 0$ by the independence of M and X^n in our problem formulation;

(127) follows from [20, Lemma 4], where Z_i is defined as $Z_i = (M, Y^{i-1}, X_{i+1}^n)$;

(128) follows from (1);

(129) follows from Jensen's inequality and the concavity of $C^{\mathrm{IE}}(d)$ from Lemma 1;

(130) follows from (124) and the nondecreasing property of $C^{\mathrm{IE}}(d)$ from Lemma 1; and

(131) follows from the Fano inequality.

Rearranging terms in (131) we have

$$P_{\mathrm{e}}^{(n)} \ge 1 - \frac{C^{\mathrm{IE}}(d)}{R} - \frac{1}{nR} \quad (132)$$

which shows for $R > C$, the probability or error is bounded away from 0.

Appendix II
Proof Claim 2 (Private Embedding Capacity)

In this appendix, we prove that the private information-embedding capacity is given by (4), where the supremum is over the set

$$\mathcal{P}_{W|X}^{\mathrm{IE}} = \{p_{W|X}(w|x)\colon E[D(X, W)] \le d\}. \quad (133)$$

A. Converse

The proof of the converse uses a technique very similar to that used in Appendix I, exploiting the concavity of $C_{\mathrm{priv}}^{\mathrm{IE}}(d)$, a fact which is established through the following lemma.

Lemma 3: The information-embedding capacity given in (4) is a nondecreasing, concave function of the distortion constraint d.

Proof: With increasing d, the domain over which the mutual information is maximized increases, which implies $C_{\mathrm{priv}}^{\mathrm{IE}}(d)$, is nondecreasing.

We prove concavity by considering two capacity–distortion pairs (C_1, d_1) and (C_2, d_2), which are points on the information-embedding capacity function. These points are achieved with the distributions $p_1(w, x, y) = p_X(x)p_{Y|W}(y|w)p_1(w|x)$ and $p_2(w, x, y) = p_X(x)p_{Y|W}(y|w)p_2(w|x)$, respectively. We define

$$p_\lambda(w, x, y) = \lambda p_1(w, x, y) + (1 - \lambda)p_2(w, x, y). \quad (134)$$

Because distortion is a linear function of the transition probabilities, the distortion for p_λ is

$$d_\lambda = \lambda d_1 + (1 - \lambda)d_2. \quad (135)$$

It is easily verified that the mutual information $I(W; Y|X = x)$ is a concave function of the distribution $p_{W|X}(w|x)$. Therefore,

$$I_{p_\lambda}(W; Y|X = x) \ge \lambda I_{p_1}(W; Y|X = x) + (1 - \lambda)I_{p_2}(W; Y|X = x) \quad (136)$$

where we subscript the mutual informations with their respective distributions. Thus, we have the following chain of inequalities:

$$C_{\mathrm{priv}}^{\mathrm{IE}}(d_\lambda) \ge I_{p_\lambda}(W; Y|X) \quad (137)$$

$$= \sum_{x\in\mathcal{X}} I_{p_\lambda}(W; Y|X = x)p_X(x) \quad (138)$$

$$\ge \sum_{x\in\mathcal{X}} \lambda I_{p_1}(W; Y|X = x)p_X(x) + \sum_{x\in\mathcal{X}} (1 - \lambda)I_{p_2}(W; Y|X = x)p_X(x) \quad (139)$$

$$\ge \lambda C_{\mathrm{priv}}^{\mathrm{IE}}(d_1) + (1 - \lambda)C_{\mathrm{priv}}^{\mathrm{IE}}(d_2) \quad (140)$$

where (139) follows from (136), which proves the concavity of $C_{\mathrm{priv}}^{\mathrm{IE}}(d)$. □

Returning to the proof of our main result, recall the input to the channel is the composite signal W^n, which is an encoded function of the host X^n and the message M. The distortion between X^n and W^n is constrained by

$$\frac{1}{n} E\left[\sum_{i=1}^{n} D(X_i, W_i)\right] \le d. \quad (141)$$

The converse is proven by the following chain of inequalities:

$$nR = H(M) \quad (142)$$

$$= H(M|X^n) = I(M; Y^n|X^n) + H(M|Y^n, X^n) \quad (143)$$

$$\le \sum_{i=1}^{n} I(M; Y_i|X^n, Y^{i-1}) + H(M|X^n, Y^n) \quad (144)$$

$$= \sum_{i=1}^{n} \left[H(Y_i|X^n, Y^{i-1}) - H(Y_i|M, X^n, Y^{i-1})\right] + H(M|X^n, Y^n) \quad (145)$$

$$\le \sum_{i=1}^{n} \left[H(Y_i|X_i) - H(Y_i|M, X^n, Y^{i-1})\right] + H(M|X^n, Y^n) \quad (146)$$

$$= \sum_{i=1}^{n} [H(Y_i|X_i) - H(Y_i|M, X^n, X_i)] + H(M|X^n, Y^n) \quad (147)$$

$$= \sum_{i=1}^{n} I(Y_i; M, X^n|X_i) + H(M|X^n, Y^n) \quad (148)$$

$$\le \sum_{i=1}^{n} I(Y_i; W_i|X_i) + H(M|X^n, Y^n) \quad (149)$$

$$\leq \sum_{i=1}^{n} C_{\text{priv}}^{\text{IE}}(E[D(X_i, W_i)]) + H(M|X^n, Y^n) \tag{150}$$

$$\leq nC_{\text{priv}}^{\text{IE}}\left(E\left[\frac{1}{n}\sum_{i=1}^{n} D(X_i, W_i)\right]\right) + H(M|X^n, Y^n) \tag{151}$$

$$\leq nC_{\text{priv}}^{\text{IE}}(d) + H(M|X^n, Y^n) \tag{152}$$

$$\leq nC_{\text{priv}}^{\text{IE}}(d) + P_{\text{e}}^{(n)}nR + 1 \tag{153}$$

where

(142) follows from our formulation that M is uniformly distributed on $\{1, 2, \ldots, 2^{nR}\}$;
(143) follows from our formulation that M and X^n are independent;
(144) follows from the chain rule for mutual information;
(146) follows from the fact that conditioning reduces entropy;
(149) follows from the data processing inequality, using the fact that $(M, X^n) \rightarrow W_i \rightarrow Y_i$ is a Markov chain;
(150) follows from (4);
(151) follows from Jensen's inequality and the concavity of $C_{\text{priv}}^{\text{IE}}(d)$ from Lemma 3;
(152) follows from (141) and that $C_{\text{priv}}^{\text{IE}}(d)$ is nondecreasing from Lemma 3;
(153) follows from the Fano inequality.

Rearranging terms in (153) we have

$$P_{\text{e}}^{(n)} \geq 1 - \frac{C_{\text{priv}}^{\text{IE}}(d)}{R} - \frac{1}{nR} \tag{154}$$

which shows for $R > C$, the probability or error is bounded away from 0.

B. Achievability

For our proof, it is convenient to express the capacity (4) in terms of $C_x^{\text{IE}}(d_x)$, the capacity of a channel when the host X is some constant value x known at the encoder and decoder, as developed in the following lemma.

Lemma 4: The information-embedding capacity with host known at the encoder and decoder satisfies

$$C_{\text{priv}}^{\text{IE}}(d) = \sup_{\{d_x:\, E[d_X]=d\}} \sum_{x\in\mathcal{X}} C_x^{\text{IE}}(d_x)p_X(x) \tag{155}$$

where, by the conventional channel capacity theorem

$$C_x^{\text{IE}}(d_x) = \sup_{p_{W|X}(w|x)\in\mathcal{P}_{W|x}} I(Y; W|X = x) \tag{156}$$

with

$$\mathcal{P}_{W|x} = \{p_{W|X}(w|x)\colon E[D(X, W|X = x)] \leq d_x\} \tag{157}$$

denoting the constraint set for the embedding.

Using this lemma, consider the set of d_x^* that achieves the maximum on the right-hand side of (155). By the conventional channel-coding theorem, we can achieve the rate $C_x^{\text{IE}}(d_x^*)$ with embedding distortion d_x^* and negligible probability of error if $X = x$ for all samples of data. Thus, the following coding scheme suffices: we embed data in X^n, a length-n block of host samples, using a different codebook for each x which achieves the rate $C_x^{\text{IE}}(d_x^*)$ at embedding distortion d_x^*. For each x, we collect all of the samples X_i for each i such that $X_i = x$ and code using the codebook corresponding to x. The total rate is thus

$$\sum_{x\in\mathcal{X}} C_x^{\text{IE}}(d_x^*)p_X(x) = \sup_{\{d_x:\, E[d_X]=d\}} \sum_{x\in\mathcal{X}} C_x^{\text{IE}}(d_x)p_X(x), \tag{158}$$

which by the lemma equals capacity.

It remains only to prove Lemma 4.

Proof of Lemma 4: We first prove that $C_{\text{priv}}^{\text{IE}}(d)$ is lower-bounded by the right-hand side of (155). To see this, choose a fixed d_x for each x such that $E[d_X] = d$ and a test channel $p_{W|X}(w|x) \in \mathcal{P}_{W|x}$. It is easily confirmed from (157) that

$$E[D(X, W)] \leq E[d_X] = d \tag{159}$$

which implies $p_{W|X}(w|x) \in \mathcal{P}_{W|X}^{\text{IE}}$ as defined in (133). For any test channel

$$\sum_{x\in\mathcal{X}} I(W; Y|X = x)p_X(x) = I(W; Y|X) \leq C_{\text{priv}}^{\text{IE}}(d) \tag{160}$$

so that choosing $p_{W|X}(w|x)$ to satisfy the maximization in (156) yields

$$\sup_{p_{W|X}(w|x)\in\mathcal{P}_{W|X}^{\text{IE}}} \sum_{x\in\mathcal{X}} C_x^{\text{IE}}(d_x)p_X(x) \leq C_{\text{priv}}^{\text{IE}}(d) \tag{161}$$

for any set of d_X satisfying $E[d_X] = d$.

It remains only to show that $C_{\text{priv}}^{\text{IE}}(d)$ is upper-bounded by the right-hand side of (155). To see this, we choose a test channel $p_{W|X}(w|x) \in \mathcal{P}_{W|X}^{\text{IE}}$, which results in a set of conditional distortions $d_x' = E[D(X, W)|X = x]$ that satisfy $E[d_x'] \leq d$. For any such test channel

$$I(W; Y|X) = \sum_{x\in\mathcal{X}} I(W; Y|X = x)p_X(x) \tag{162}$$

$$\leq \sum_{x\in\mathcal{X}} C_x^{\text{IE}}(d_x')p_X(x) \tag{163}$$

$$\leq \sup_{\{d_x:\, E[d_X]\leq d\}} \sum_{x\in\mathcal{X}} C_x^{\text{IE}}(d_x)p_X(x). \tag{164}$$

Choosing $p_{W|X}(w|x)$ to achieve the maximum in (4) yields

$$C_{\text{priv}}^{\text{IE}}(d) \leq \sup_{\{d_x:\, E[d_X]=d\}} \sum_{x\in\mathcal{X}} C_x^{\text{IE}}(d_x)p_X(x) \tag{165}$$

which completes the proof of the lemma. □

Appendix III
Dithered Nested Lattice Code for Wyner–Ziv Encoding

Nested lattices can also be used to build Wyner–Ziv codes that are capacity achieving at all SNRs. Our construction exploits dithered quantizers, and can be viewed as a generalization of the result in [36]. As before, it suffices to restrict our attention to the case $d < \sigma_{Y|X}^2$.

Our dithered quantizers are defined via

$$Q(\boldsymbol{U}) = \underset{\boldsymbol{l}+\boldsymbol{T}\in\mathcal{L}+\boldsymbol{T}}{\arg\min} \|\boldsymbol{U}-(\boldsymbol{l}+\boldsymbol{T})\|^2 \tag{166}$$

where the dither $\boldsymbol{T}$ is uniform over the characteristic Voronoi cell and generated independently. By the properties of subtractive dithered quantization, we must change the property (GQ-1) for the new quantizer $Q(\cdot)$ as follows.

(GQ-1′) The quantization error (30) is white and Gaussian with zero-mean and variance σ^2, and independent of the input to the quantizer [35].

The other properties (GQ-2) and (GQ-3) remain valid with $Q(\cdot)$.

With these quantizers, the nested lattices are chosen such that

$$\sigma_1^2 = \frac{d\sigma_{Y|X}^2}{\left(\sigma_{Y|X}^2 - d\right)} \quad \text{and} \quad \sigma_2^2 = \sigma_1^2 + \sigma_{Y|X}^2 + \epsilon. \tag{167}$$

A suitable encoder using these lattices transmits the index of the closet coset to the source $\boldsymbol{Y}$, i.e., it transmits $M = k(Q_1(\boldsymbol{Y}))$. The decoder observes $\boldsymbol{X}$ and M, calculates the coset shift $\boldsymbol{S} = g(M)$ and an MMSE estimate $\tilde{\boldsymbol{Y}} = \rho\boldsymbol{X}$, where $\rho = \sigma_Y^2/(\sigma_Y^2+\sigma_V^2)$. The decoder then produces a source estimate of the form

$$\boldsymbol{W} = a\boldsymbol{X} + b\tilde{\boldsymbol{W}} \tag{168}$$

where

$$\tilde{\boldsymbol{W}} = Q_2^{\boldsymbol{S}}\left(\tilde{\boldsymbol{Y}}\right). \tag{169}$$

That the system operates at the target rate follows from the lattice properties. Indeed, Property (GQ-3) and (167) prescribe the rate of the code to be within $1/n$ bits of

$$\begin{aligned} R &= \frac{1}{n}\log\left(\frac{V_1}{V_2}\right) = \frac{1}{2}\log\left(\frac{\sigma_2^2 G_1}{\sigma_1^2 G_2}\right) \\ &\le \frac{1}{2}\log\left(\frac{\sigma_{Y|X}^2}{d}\right) + O(\epsilon) = R_{Y|X}^{\mathrm{WZ}}(d) + O(\epsilon) \end{aligned} \tag{170}$$

where the last equality follows from (23).

Next, to verify that the decoder reconstructs the source $\boldsymbol{Y}$ to within distortion d, we first define the quantization error

$$\boldsymbol{E}_1 = \boldsymbol{Y} - Q_1(\boldsymbol{Y}) \tag{171}$$

and the estimation error

$$\boldsymbol{E}_{Y|X} = \boldsymbol{Y} - \tilde{\boldsymbol{Y}} \tag{172}$$

so that

$$Q_1(\boldsymbol{Y}) = \tilde{\boldsymbol{Y}} + \boldsymbol{Z} \tag{173}$$

with

$$\boldsymbol{Z} = \boldsymbol{E}_{Y|X} - \boldsymbol{E}_1. \tag{174}$$

To establish that $\boldsymbol{Z}$ in (174) is independent of $\tilde{\boldsymbol{Y}}$ in (173), we first note that $\boldsymbol{E}_{Y|X}$ is independent of $\tilde{\boldsymbol{Y}}$ by the orthogonality principle. It only remains to show that $\boldsymbol{E}_1$ is independent of $\tilde{\boldsymbol{Y}}$ and $\boldsymbol{E}_{Y|X}$. To see this, note that

$$\begin{aligned} E\left[\boldsymbol{E}_1\boldsymbol{E}_{Y|X}\middle|\tilde{\boldsymbol{Y}}=\tilde{\boldsymbol{y}}\right] &= E\left[\boldsymbol{E}_1\boldsymbol{Y}\middle|\tilde{\boldsymbol{Y}}=\tilde{\boldsymbol{y}}\right] - \tilde{\boldsymbol{Y}}E\left[\boldsymbol{E}_1\middle|\tilde{\boldsymbol{Y}}=\tilde{\boldsymbol{y}}\right] \\ &= E[\boldsymbol{E}_1'\boldsymbol{Y}'] - \tilde{\boldsymbol{y}}E[\boldsymbol{E}_1'] = 0 \end{aligned} \tag{175}$$

where the first equality follows from (172), the second equality follows from the definition

$$\boldsymbol{Y}' = \tilde{\boldsymbol{y}} + \boldsymbol{E}_{Y|X} \tag{176}$$

and the third equality follows from the fact that by (GQ-1′) the quantization error $\boldsymbol{E}_1' = \boldsymbol{Y}' - Q_1(\boldsymbol{Y}')$ is zero mean and independent of $\boldsymbol{Y}'$. Hence, it follows that $\boldsymbol{E}_1$ is independent of both $\tilde{\boldsymbol{Y}}$ and $\tilde{\boldsymbol{E}}_{Y|X}$

$$E\left[\boldsymbol{E}_1\tilde{\boldsymbol{Y}}\right] = E[\boldsymbol{E}_1\boldsymbol{E}_{Y|X}] = 0 \tag{177}$$

where the first equality follows from an application of (GQ-1′), and the second by averaging over $\tilde{\boldsymbol{y}}$ in (175).

Now, since $\boldsymbol{Z}$ is effectively Gaussian with zero mean and variance $\sigma_{Y|X}^2+\sigma_1^2$, we know from (GQ-2) that $\Pr[Q_2(\boldsymbol{Z})\neq\boldsymbol{0}]<\epsilon$. Furthermore, if $Q_2(\boldsymbol{Z})=\boldsymbol{0}$, then by the translational invariance of lattice geometry

$$Q_2^{\boldsymbol{s}}(\boldsymbol{l}+\boldsymbol{Z}) = \boldsymbol{l}, \quad \text{for any coset shift } \boldsymbol{s} \text{ and any } \boldsymbol{l}\in\mathcal{L}_2^{\boldsymbol{s}}. \tag{178}$$

So with $\tilde{\boldsymbol{W}}$ as defined in (169) we have that, using (173) and exploiting the independence of $\boldsymbol{Z}$ and $\tilde{\boldsymbol{Y}}$

$$\begin{aligned} &\Pr\left[\tilde{\boldsymbol{W}} \neq Q_1(\boldsymbol{Y})\middle|\tilde{\boldsymbol{Y}}=\tilde{\boldsymbol{y}}\right] \\ &\quad= \Pr\left[Q_2^{\boldsymbol{S}}(\tilde{\boldsymbol{Y}}) \neq Q_1(\boldsymbol{Y})\middle|\tilde{\boldsymbol{Y}}=\tilde{\boldsymbol{y}}\right] \\ &\quad= \Pr\left[Q_2^{\boldsymbol{S}}(Q_1(\boldsymbol{Y})-\boldsymbol{Z}) \neq Q_1(\boldsymbol{Y})\middle|\tilde{\boldsymbol{Y}}=\tilde{\boldsymbol{y}}\right] \\ &\quad= \Pr[Q_2(\boldsymbol{Z})=0] \\ &\qquad\cdot \Pr\left[Q_2^{\boldsymbol{S}}(Q_1(\boldsymbol{Y})-\boldsymbol{Z}) \neq Q_1(\boldsymbol{Y})\middle|Q_2(\boldsymbol{Z})=0,\,\tilde{\boldsymbol{Y}}=\tilde{\boldsymbol{y}}\right] \\ &\qquad+ \Pr[Q_2(\boldsymbol{Z})\neq 0] \\ &\qquad\cdot \Pr\left[Q_2^{\boldsymbol{S}}(Q_1(\boldsymbol{Y})-\boldsymbol{Z}) \neq Q_1(\boldsymbol{Y})\middle|Q_2(\boldsymbol{Z})\neq 0,\,\tilde{\boldsymbol{Y}}=\tilde{\boldsymbol{y}}\right] \\ &\quad\le \epsilon \end{aligned} \tag{179}$$

since the term on the fifth line is zero by using $\boldsymbol{l} = Q_1(\boldsymbol{Y}) \in \mathcal{L}_2^{\boldsymbol{S}}$ in (178), and since $\Pr[Q_2(\boldsymbol{Z})\neq 0]$ is bounded by ϵ. Thus, using (179) in (168), we have that with probability $1-\epsilon$

$$\boldsymbol{W} = a\boldsymbol{X} + bQ_1(\boldsymbol{Y}). \tag{180}$$

Choosing a and b so as to minimize the mean-square distortion between $\boldsymbol{W}$ and $\boldsymbol{Y}$, we obtain, using basic linear MMSE estimation theory, that the optimum a and b yield a mean-square estimation error of

$$\frac{1}{n}E\left[\|\boldsymbol{W}-\boldsymbol{Y}\|^2\right] = d + O(\epsilon) \tag{181}$$

which confirms the distortion constraint is met.

APPENDIX IV
CAPACITIES OF INFORMATION EMBEDDING FOR THE BINARY-HAMMING CASE

A. Proof of Claim 3 (Public Case)

The upper concave envelope of $g_p^{\mathrm{IE}}(d)$ in (54) is given by

$$g^*(d) = \sup_{\theta,\beta_1,\beta_2} \left[\theta g_p^{\mathrm{IE}}(\beta_1) + (1-\theta) g_p^{\mathrm{IE}}(\beta_2)\right] \tag{182}$$

where the supremum is taken with respect to all $\theta \in [0, 1]$ and $\beta_1, \beta_2 \in [0, \frac{1}{2}]$ such that $d = \theta\beta_1 + (1-\theta)\beta_2$. By the concavity of $h(\cdot)$, it is clear that $g_p^{\mathrm{IE}}(d)$ is concave over $p \le d \le \frac{1}{2}$. Thus, the maximization in (182) can be simplified by letting $\beta_2 = 0$

$$g^*(d) = \sup_{\theta,\beta}[\theta(h(\beta) - h(p))], \qquad 0 \le d \le \frac{1}{2} \tag{183}$$

where the supremum is taken with respect to all $\theta \in [0, 1]$ and $\beta \in [0, \frac{1}{2}]$ such that

$$d = \theta\beta. \tag{184}$$

We establish that $C^{\mathrm{IE}}(d) = g^*(d)$ by separately proving that $C^{\mathrm{IE}}(d)$ is lower- and upper-bounded by $g^*(d)$.

The lower bound is developed by considering a special case. Let the auxiliary random variable U be the output of a binary-symmetric channel with crossover probability β which has X as input. Furthermore, we choose f such that $W = f(U, X) = U$, which makes the distortion equal β. We evaluate

$$\begin{aligned} I(Y; U) - I(U; X) &= I(Y; W) - I(U; X) \\ &= (1 - h(p)) - (1 - h(\beta)) = h(\beta) - h(p) \end{aligned} \tag{185}$$

and conclude from (1) that

$$C^{\mathrm{IE}}(d) \ge h(\beta) - h(p) \tag{186}$$

when we choose the values $\theta \in [0, 1]$ and $\beta \in [0, \frac{1}{2}]$ such that (184) holds for some given $d \in [0, \frac{1}{2}]$.

By the concavity of $C^{\mathrm{IE}}(d)$ from Lemma 1, we have

$$C^{\mathrm{IE}}(d) = C^{\mathrm{IE}}(\theta\beta) \ge \theta C^{\mathrm{IE}}(\beta) \ge \theta(h(\beta) - h(p)) \tag{187}$$

which is true for all θ and β satisfying (184), whence $C^{\mathrm{IE}}(d) \ge g^*(d)$.

It remains only to show the upper bound $C^{\mathrm{IE}}(d) \le g^*(d)$, for which it suffices to show that

$$I(Y; U) - I(U; X) \le g^*(d) \tag{188}$$

for any $p_{W,U|X}(w, u|x)$ such that $E[D(X, W)] = d$.

Defining the set

$$\mathcal{A} = \{u\colon f(u, 0) = f(u, 1)\} \tag{189}$$

we have

$$\begin{aligned} d &\ge E[D(X, W)] && (190)\\ &= \Pr(U \in \mathcal{A}) E[D(X, W)|U \in \mathcal{A}] \\ &\quad + \Pr(U \in \mathcal{A}^{\mathrm{C}}) E[D(X, W)|U \in \mathcal{A}^{\mathrm{C}}] && (191)\\ &\ge \Pr(U \in \mathcal{A}) E[D(X, W)|U \in \mathcal{A}]. && (192) \end{aligned}$$

Using

$$E[D(X,W)|U \in \mathcal{A}] = \sum_{u\in\mathcal{A}} \frac{p_U(u)}{\Pr(U \in \mathcal{A})} E[D(X,W)|U = u] \tag{193}$$

with (192) yields

$$d' = \theta \sum_{u\in\mathcal{A}} \lambda_u d_u \le d \tag{194}$$

where $\theta = \Pr(U \in \mathcal{A})$, $\lambda_u = p_U(u)/\Pr(U \in \mathcal{A})$, and

$$d_u = E[D(X, W)|U = u]. \tag{195}$$

We observe that, because $H(X) = 1$, we have $H(X) - H(Y) = \epsilon$, $\epsilon \ge 0$, and thus,

$$\begin{aligned} &I(Y; U) - I(U; X) \\ &\quad = H(X|U) - H(Y|U) - \epsilon && (196)\\ &\quad \le \sum_{u\in\mathcal{A}} [H(X|U = u) - H(Y|U = u)] p_U(u) \\ &\qquad + \sum_{u\in\mathcal{A}^{\mathrm{C}}} [H(X|U = u) - H(Y|U = u)] p_U(u) && (197)\\ &\quad \le \sum_{u\in\mathcal{A}} [H(X|U = u) - H(Y|U = u)] p_U(u) && (198)\\ &\quad = \theta \sum_{u\in\mathcal{A}} \lambda_u [H(X|U = u) - H(Y|U = u)] && (199) \end{aligned}$$

where for (198) we have used

$$H(X|U = u) < H(Y|U = u), \qquad \forall U \in \mathcal{A}^{\mathrm{C}}$$

which is true because for any $U \in \mathcal{A}^{\mathrm{C}}$, the channel input W is either X or the complement of X. Because the channel is binary symmetric, the entropy of Y is thus greater than or equal to that of X.

We proceed to evaluate the right-hand side of (199). Consider any $u \in \mathcal{A}$. Defining $\gamma(u) = f(u, 0) = f(u, 1)$, we obtain, using (195)

$$d_u = E[D(X, W)|U = u] = \Pr(X \ne \gamma(u)|U = u). \tag{200}$$

So

$$H(X|U = u) = h(d_u). \tag{201}$$

Next, given $U = u$, the channel input is uniquely specified by $W = \gamma(u)$, and thus,

$$H(Y|U = u) = h(p). \tag{202}$$

Thus,

$$\begin{aligned} I(Y; U) - I(U; X) &\le \theta \sum_{u\in\mathcal{A}} \lambda_u (h(d_u) - h(p)) && (203)\\ &= \theta \sum_{u\in\mathcal{A}} \lambda_u G(d_u), && (204)\\ &\le \theta G\left(\sum_{u\in\mathcal{A}} \lambda_u d_u\right) && (205) \end{aligned}$$

$$= \theta(h(\beta) - h(p)) \tag{206}$$
$$= g^*(d') \tag{207}$$
$$\leq g^*(d) \tag{208}$$

where

(203) is obtained by substituting (201) and (202) into (199);
(204) is obtained by defining $G(v) \stackrel{\Delta}{=} h(v) - h(p)$;
(205) follows from the facts that G is concave for $0 \leq v \leq \frac{1}{2}$ and $\sum_{u \in \mathcal{A}} \lambda_u = 1$;
(206) follows from defining $\beta = \sum_{u \in \mathcal{A}} \lambda_u d_u$;
(207) follows from the definition of $g^*(d)$ in (183) with $\theta\beta = d'$; and
(208) follows from the fact that $d' \leq d$ and g^* is a nondecreasing function.

Hence, we have shown that for any distribution $p_{W,U|X}(w, u|x)$ there exists a $\theta \in [0, 1]$ and $\beta \in [0, \frac{1}{2}]$ such that (188) holds.

B. Proof of Claim 4 (Private Case)

Since adding (modulo-2) a known symbol to both W and Y in (4) does not affect their mutual information, we have

$$C_{\text{priv}}^{\text{IE}}(d) = \sup_{p_{E|X}(e|x)} I(Y \oplus X; E|X) \tag{209}$$

where $E = W \oplus X$ is the distortion due to embedding, which is constrained to have $p_E(1) \leq d$. Note that $Y \oplus X = E \oplus V$, where V is a Bernoulli(p) source representing the noise of the binary-symmetric channel. Under the constraint that $p_E(1) \leq d$, we have the following chain of inequalities:

$$I(Y \oplus X; E|X)$$
$$= H(E \oplus V|X) - H(E \oplus V|E, X) \tag{210}$$
$$\leq H(E \oplus V) - \sum_{e \in \{0,1\}} p_E(e) H(E \oplus V|E = e, X) \tag{211}$$
$$= H(E \oplus V) - \sum_{e \in \{0,1\}} p_E(e) h(p) \tag{212}$$
$$\leq h(p * d) - h(p). \tag{213}$$

The inequalities are met with equality if E is Bernoulli(d), independent of X, and V, which proves the claim.

ACKNOWLEDGMENT

The authors wish to thank Prof. Amos Lapidoth, Dr. Aaron Cohen, Dr. Ram Zamir, and E. Martinian for helpful interactions, and the anonymous reviewers and Associate Editor Prof. Imre Csiszár for their thoughtful feedback and careful reading of the manuscript, which led to many improvements.

REFERENCES

[1] R. J. Barron, "Systematic hybrid analog/digital signal coding," Ph.D. dissertation, MIT, Cambridge, MA, June 2000.
[2] R. J. Barron, B. Chen, and G. W. Wornell, "The duality between information embedding and source coding with side information and some applications," in *Proc. IEEE Int. Symp. Information Theory*, Washington, DC, June 2001, p. 300.
[3] R. J. Barron and A. V. Oppenheim, "Signal processing for hybrid channels," in *Proc. ARL Fedlabs Symp.*, Feb. 1999, pp. 481–484.
[4] ——, "A systematic hybrid analog/digital audio coder," in *Proc. Workshop Applications of Signal Processing to Audio, Acoustics*, Mohonk, NY, Oct. 1999.
[5] T. Berger, *Rate Distortion Theory*. Englewood Cliffs, NJ: Prentice-Hall, 1971.
[6] B. Chen, "Design and analysis of digital watermarking, information embedding, and data hiding systems," Ph.D. dissertation, MIT, Cambridge, MA, June 2000.
[7] B. Chen and G. W. Wornell, "Quantization index modulation: A class of provably good methods for digital watermarking and information embedding," *IEEE Trans. Inform. Theory*, vol. 47, pp. 1423–1443, May 2001.
[8] ——, "Quantization index modulation methods for digital watermarking and information embedding of multimedia," *J. VLSI Signal Processing Syst. for Signal, Image, and Video Technol.*, vol. 27, pp. 7–33, Feb. 2001.
[9] M. Chiang and T. M. Cover, "Duality of channel capacity and rate distortion with side information," in *Proc. Int. Symp. Information Theory and Its Applications*, Nov. 2000.
[10] J. Chou, S. S. Pradhan, and K. Ramchandran, "On the duality between distributed source coding and data hiding," in *Proc. Asilomar Conf. Signals, Systems, Computers*, Pacific Grove, CA, Oct. 1999, pp. 1503–1507.
[11] A. S. Cohen, private communication.
[12] A. S. Cohen and A. Lapidoth, "The Gaussian watermarking game: Parts I and II," *IEEE Trans. Inform. Theory*, vol. 48, pp. 1639–1667, June 2002.
[13] J. H. Conway and N. J. A. Sloane, *Sphere Packings, Lattices, and Groups*. New York: Spring-Verlag, 1988.
[14] M. H. Costa, "Writing on dirty paper," *IEEE Trans. Inform. Theory*, vol. IT-29, pp. 439–441, May 1983.
[15] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.
[16] T. M. Cover and M. Chiang, "Duality between channel capacity and rate distortion with two-sided state information," *IEEE Trans. Inform. Theory*, vol. 48, pp. 1629–1638, June 2002.
[17] M. Chiang and T. M. Cover, "Unified duality between channel capacity and rate distortion with state information," in *Proc. IEEE Int. Symp. Information Theory*, Washington, DC, June 2001, p. 301.
[18] U. Erez, S. Shamai, and R. Zamir, "Capacity and lattice-strategies for cancelling known interference," in *Proc. Int. Symp. Information Theory and Its Applications*, Honolulu, HI, Nov. 2000, pp. 681–684.
[19] U. Erez and R. Zamir, "Lattice decoding can achieve $\frac{1}{2}\log(1 + \text{SNR})$ on the AWGN channel using nested codes," in *Proc. IEEE Int. Symp. Information Theory*, Washington, DC, June 2001, p. 125. Also, submitted to *IEEE Trans. Inform. Theory*.
[20] S. I. Gel'fand and M. S. Pinsker, "Coding for channel with random parameters," *Probl. Contr. Inform. Theory*, vol. 9, no. 1, pp. 19–31, 1980.
[21] R. M. Gray, "Conditional rate-distortion theory," Stanford Univ., Stanford, CA, Electronics Laboratories Tech. Rep. 6502-2, Oct. 1972.
[22] C. Heegard and A. El Gamal, "On the capacity of computer memory with defects," *IEEE Trans. Inform. Theory*, vol. IT-29, pp. 731–739, Sept. 1983.
[23] P. Moulin and J. A. O'Sullivan, "Information-theoretic analysis of information hiding," , vol. 49, pp. 563–593, Mar. 2003.
[24] S. S. Pradhan, J. Chou, and K. Ramchandran, "Duality between channel and source coding with side information," Univ. California, Berkeley, UCB/ERL Tech. Memo. M01/34, Dec. 2001.
[25] ——, "A characterization of functional duality between source and channel coding," in *Proc. IEEE Int. Symp. Information Theory*, Lausanne, Switzerland, June 2002, p. 224.
[26] S. S. Pradhan and K. Ramchandran, "Distributed source coding using syndromes (DISCUS): Design and construction," in *Proc. Data Compression Conf.*, Snowbird, UT, Mar. 1999, pp. 158–167.
[27] S. Servetto, "Lattice quantization with side information," in *Proc. Data Compression Conf.*, Snowbird, UT, Mar. 2000, pp. 510–522.
[28] S. Shamai (Shitz), S. Verdú, and R. Zamir, "Systematic lossy source/channel coding," *IEEE Trans. Inform. Theory*, vol. 44, pp. 564–579, Mar. 1998.
[29] D. Slepian and J. K. Wolf, "Noiseless coding of correlated information sources," *IEEE Trans. Inform. Theory*, vol. IT-19, pp. 471–480, July 1973.
[30] J. K. Su, J. J. Eggers, and B. Girod, "Illustration of the duality between channel coding and rate distortion with side information," in *Proc. Asilomar Conf. Signals, Systems, Computers*, Pacific Grove, CA, Nov. 2000.

[31] J. Wolfowitz, *Coding Theorems of Information Theory*, 2nd ed. New York: Springer-Verlag, 1964.
[32] A. D. Wyner, "Recent results in Shannon theory," *IEEE Trans. Inform. Theory*, vol. IT-20, pp. 2–9, Jan. 1974.
[33] ——, "The rate-distortion function for source coding with side information at the decoder—II: General sources," *Inform. Contr.*, vol. 38, pp. 60–80, 1978.
[34] A. D. Wyner and J. Ziv, "The rate-distortion function for source coding with side information at the decoder," *IEEE Trans. Inform. Theory*, vol. IT-22, pp. 1–10, Jan. 1976.
[35] R. Zamir and M. Feder, "On lattice quantization noise," *IEEE Trans. Inform. Theory*, vol. 42, pp. 1152–1159, July 1996.
[36] R. Zamir and S. Shamai, "Nested linear/lattice codes for Wyner–Ziv encoding," in *Proc. Inform. Theory Workshop*, Kilarney, Ireland, June 1998, pp. 92–93.
[37] R. Zamir, S. Shamai (Shitz), and U. Erez, "Nested linear/lattice codes for multiterminal binning," *IEEE Trans. Inform. Theory*, vol. 48, pp. 1250–1276, June 2002.