Submitted to *IEEE Trans. Inform. Theory*

Stealing Bits from a Quantized Source

Aaron S. Cohen, Stark C. Draper, Emin Martinian, and Gregory W. Wornell

Submitted September 2003

Abstract

We consider "bit stealing" scenarios where the rate of a source code must be reduced without prior planning. We first investigate the efficiency of source requantization to reduce rate, which we term successive degradation. We focus on finite-alphabet sources with arbitrary distortion measures as well as the Gaussian-quadratic and high-resolution scenarios. We show an achievable rate distortion trade-off and prove that this is the best guaranteeable trade-off independent of source code design. This trade-off is in general different from the rate distortion trade-off with successive refinement, where there is prior planning, but we show that with quadratic distortion measures, for all sources with finite differential entropy and at least one finite moment, the gap is at most 1/2-bit or 3 dB in the high-resolution limit. In the Gaussian-quadratic case, the gap is at most 1/2-bit for all resolutions.

We further consider bit stealing in the form of information embedding, whereby an embedder acts on a quantized source and produces an output at the same rate and in the original source codebook. We develop achievable distortion rate trade-offs. Two cases are considered, corresponding to whether or not the source decoder is informed of the embedding rate. In the Gaussian-quadratic case, we show the informed decoder need only augment the regular decoder with simple post-reconstruction distortion compensation in the form of linear scaling, and that in this case such systems can be as efficient as bit stealing via successive degradation. Finally, we show that the penalty for uninformed versus informed decoders is at most 3 dB or 0.21-bit in the Gaussian-quadratic case, and that their performance also lies within the 1/2-bit gap to that of successive refinement.

Index Terms — transcoding, successive refinement, information embedding, digital watermarking, rate distortion theory, coding with side information, successive degradation, quantization

1 Introduction

There are a variety of engineering scenarios where the rate of a source code must be decreased. This paper considers achievable, and guaranteeable, distortion rate trade-offs when no provisions have been made for the rate reduction. Because it is not planned for, we term this "bit stealing".

This work has been supported in part by the National Science Foundation under Grant No. CCR-0073520, by Hewlett-Packard through the MIT/HP Alliance, and by a grant from Microsoft Research. This work was presented in part at the Data Compression Conference, Salt Lake, UT, April 2002, and at the International Symposium on Information Theory, Lausanne, Switzerland, July 2002.

A. S. Cohen was with the Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology. He is now with the Division of Applied Mathematics, Brown University, Providence, RI 02912 (E-mail: acohen@dam.brown.edu).

S. C. Draper was with the Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology. He is now with the Edward S. Rogers Department of Electrical and Computer Engineering, University of Toronto, Toronto, ON M5S 3G4 (E-mail: sdraper@comm.utoronto.ca).

E. Martinian is with the Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA 02139 (E-mail: emin@allegro.mit.edu).

G. W. Wornell is with the Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA 02139 (E-mail: gww@mit.edu).

To illustrate an application of bit stealing, consider transporting an compressed source through a congested multi-hop network. Suppose that the source code is of rate $R_\circ$ and achieves distortion $d_\circ$. We discuss two strategies for alleviating the congestion that occurs when the source packet (which can be further compressed) and a second digital data packet (which cannot be compressed) both arrive at a common link that cannot support their combined rate.

The first is to split the link rate into two independent data streams. With this strategy the source codeword is transcoded into a lower rate codebook of rate R, which increases the distortion to $d > d_\circ$, and the remaining (stolen) rate is used to transmit the second data packet. Transcoding is efficient if the pairs $(R_\circ, d_\circ)$ and (R, d) both lie on the rate distortion curve. If the source was originally encoded in a successively refinable manner (see, e.g., [1] and the references therein), efficient transcoding is sometimes possible by discarding least significant descriptions. In [1], Equitz and Cover give a necessary and sufficient Markov condition for such efficiency. We show, however, that near-efficiency in "successive degradation" can be possible even without such special codebook structure.

An alternative to splitting the rate into two data streams is to inject the data bits into the source bits via information embedding. This scenario differs from other investigations that jointly consider quantization and information embedding (see, e.g., [2] [3] [4]) because the source is quantized before embedding occurs. Therefore, the host signal at the information embedder is not a source vector, but rather a quantization index. If the embedded (stolen) rate is r, then the residual rate is $R = R_\circ - r$. As with successive degradation, the information embedding approach is efficient if $(R_\circ, d_\circ)$ and (R, d) both lie on the rate distortion curve.

Each strategy has its advantages. In rate-splitting the data packet is easily decoded as it is transmitted independently of the reduced-rate source description. However, source decoding is more involved because the decoder must be informed of what lower rate codebook was used during transcoding. In embedding, on the other hand, message decoding is more involved because data bits are now intertwined with source bits. However, while the embedding operation changes which codeword is transmitted, the codebook can be kept the same. In certain applications this may be an advantage, for example, since source decoders do not necessarily have to be informed that any bit stealing has taken place.

From a broader perspective, one can view the original successive refinement problem as one of transcoding with *informed* encoders, i.e., encoders aware of the possibility that the source may subsequently be transcoded, and the rate of that transcoding. In contrast, the bit stealing problem is one of transcoding with *uninformed* encoders, i.e., encoders that either do not know that transcoding may take place, or do but do not know the residual rate. Within this taxonomy, there are two natural bit stealing subproblems, corresponding to whether a source *decoder* is also informed or not about whether transcoding has taken place. We ultimately explore both cases in our development of the successive degradation and information embedding approaches.

An outline of the paper is as follows. Section 3 poses the successive degradation problem and characterizes its solution. The proofs are developed in Section 4, and Section 5 applies them to the case of a binary source and Hamming distortion measure for the purpose of illustration. Section 6 develops the corresponding

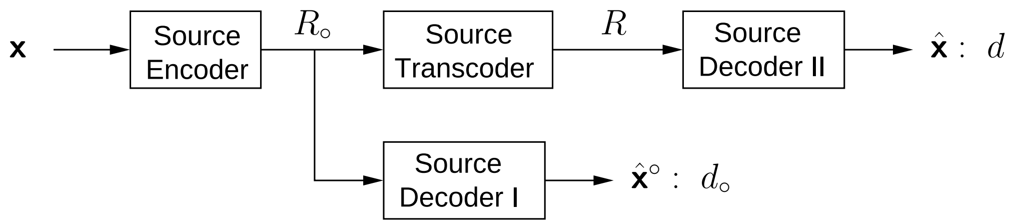


Figure 1: Bit stealing via successive degradation: the rate $R_\circ$ of the code for source $\mathbf{x}$ is reduced to R, incurring an increase in reconstruction distortion from $d_\circ$ to d. Source decoder I is the decoder for the original encoding, which produces reconstruction $\hat{\mathbf{x}}^\circ$ at distortion $d_\circ$. Source decoder II is the decoder for the transcoded source, which produces reconstruction $\hat{\mathbf{x}}$ at distortion d.

results for a Gaussian source and quadratic distortion measure, and Section 7 discusses continuous sources more generally in the high resolution limit. Finally, Section 8 develops aspects of the behavior of information embedding strategies in the corresponding scenarios, for both informed and uninformed decoders, and Section 9 contains some concluding comments.

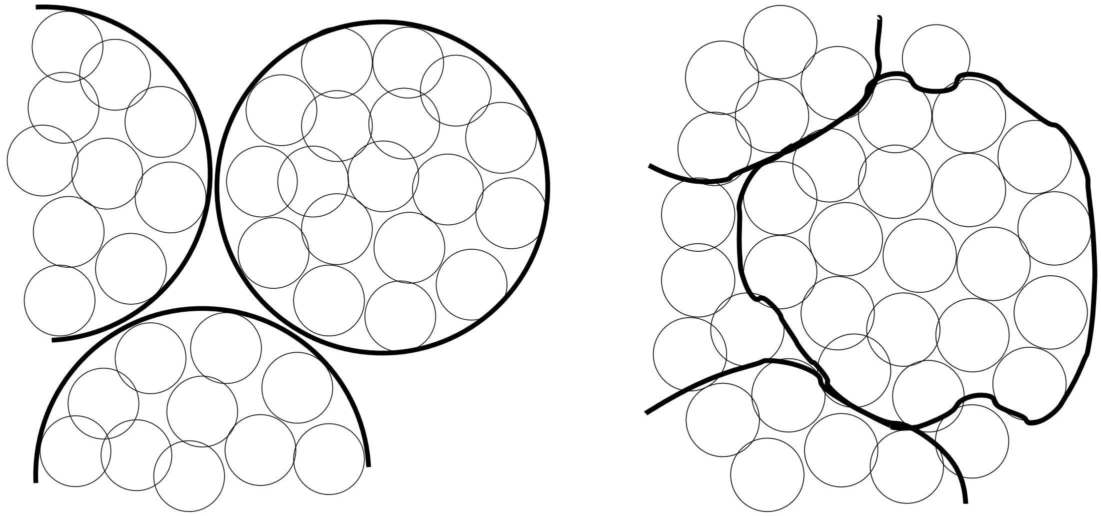
2 Notation

We use $I(\cdot;\cdot)$, $H(\cdot)$, $h(\cdot)$, and $D\left(\cdot\|\cdot\right)$ to denote mutual information, entropy, differential entropy, and relative entropy (divergence), respectively. The argument to $H(\cdot)$ and $h(\cdot)$ can be either a random variable or its distribution; we use both interchangeably. In addition $H_{\mathrm{B}}\left(\cdot\right)$ denotes the entropy of a Bernoulli source with the specified parameter. We use $R(\cdot)$ to generically denote the rate distortion function for a source, and $d(\cdot)$ the corresponding distortion rate function. We further use T to denote the type (i.e., empirical distribution of the elements) of its vector subscript, and $\mathcal{T}(\cdot)$ to denote the type class of its argument, i.e., the set of vectors with empirical distribution given by the argument. Joint types are defined similarly. The superscript $^{\mathrm{c}}$ applied to an event denotes its complement, and $|\cdot|$ applied to a set denotes its cardinality. Finally $\leftrightarrow$ is used to denote Markov chain relationships, and $E\left[\cdot\right]$ denotes expectation.

3 Successive Degradation

Fig. 1 depicts the successive degradation scenario. The i.i.d. source $\mathbf{x}$, to which we restrict our attention to simplify the exposition, is encoded at rate $R_\circ$ giving source reconstruction $\hat{\mathbf{x}}^\circ$ at distortion $d_\circ$. The transcoder re-encodes the source codeword into a second codebook of residual rate $R \leq R_\circ$ giving source reproduction $\hat{\mathbf{x}}$ at distortion d. In our problem model, all codebooks are known to their respective decoders.

Given the possibility of an informed source encoder, this problem is a special case of branching communication systems investigated by Yamamoto in [5]. Yamamoto considered the joint encoding, but sequential decoding, of a pair of correlated sources. Setting the two sources equal gives the same rate distortion region as generalized successive refinement [6]. Therefore, given control over the design of the source encoder, an optimal successive degradation approach is to encode into two codewords, per successive refinement. Decoder



(a) Nested successive refinement code

(b) Non-nested code

Figure 2: Voronoi regions are designed to nest efficiently in a successive refinement code (Fig. 2-a), but such a packing does not come naturally to all codes (Fig. 2-b). To degrade a code, a clustering of Voronoi regions that form a higher-distortion, lower rate, description of the source must be found.

I receives both codewords, while Decoder II receives only the most significant.

In this paper by restricting our attention to uninformed encoders, we preclude the possibility of having control over the design of the original source code. Instead we only assume knowledge of source statistics and that the source code achieves some average distortion. Under these assumptions we determine guaranteeable transcoder performance independent of detailed knowledge of source code design. This scenario is of interest, e.g., when the transcoder must be backward (or future) compatible and function correctly with encoders not designed to anticipate transcoding. Furthermore, since a refinable structure is an additional design requirement, we want to determine how important such structure is. In some prominent cases the performance lost by not imposing such structure is not great — the natural structure of any non-nested source code can be nearly refinable.

An informal sense of the difference between successive refinement and successive degradation is illustrated in Fig. 2. A successive refinement code has a nested structure as indicated in Fig. 2-a. The base index specifies a rough quantization region (the bold circles in Fig. 2-a), and refinement regions (the small circles in Fig. 2-a) are designed conditionally, given a base indices and the source vector. This design is most efficient when fine regions nest in rough ones.[1]

On the other hand, not all high rate codes have a nested structure. In general, to lower the description rate of a source, a clustering of fine quantization regions must be chosen. The clusters describe the source at lower fidelity, and can be enumerated at a lower rate. This is depicted in Fig. 2-b. Degradation performance

[1] The nesting character of successive refinement codes is developed formally by Rimoldi [6].

will therefore be dependent on source code design. We determine guaranteeable rate distortion trade-offs independent of source code design.

3.1 Rate Distortion Functions

We begin by defining an information rate distortion function, which we subsequently show to be operational in a rather natural sense.

Definition 1 *Let $p_{\mathsf{x}}(x)$ be the distribution of the i.i.d. source, and let $d(\cdot,\cdot)$ be the corresponding per-letter distortion measure. Let $p_{\hat{\mathsf{x}}^\circ|\mathsf{x}}(\hat{x}^\circ|x)$ be a conditional distribution that uniquely achieves*[2] *a point on the rate distortion function with*

$$d_\circ = \sum_{x,\hat{x}^\circ} p_{\hat{\mathsf{x}}^\circ|\mathsf{x}}(\hat{x}^\circ|x)\, p_{\mathsf{x}}(x) d(x, \hat{x}^\circ) \tag{1}$$

Then, the information successive degradation rate distortion function $R_{\mathrm{SD}}(d)$ *is defined for distortions $d \geq d_\circ$ as*

$$R_{\mathrm{SD}}(d) = \inf I(\hat{\mathsf{x}}; \hat{\mathsf{x}}^\circ). \tag{2}$$

where the minimization is over all conditional distributions $p_{\hat{\mathsf{x}}|\hat{\mathsf{x}}^\circ}(\hat{x}|\hat{x}^\circ)$ such that the Markov constraint $\mathsf{x} \leftrightarrow \hat{\mathsf{x}}^\circ \leftrightarrow \hat{\mathsf{x}}$ is satisfied and $E\left[d(\mathsf{x}, \hat{\mathsf{x}})\right] \leq d$.

The corresponding information successive degradation distortion rate function is defined as

$$d_{\mathrm{SD}}(R) = \inf\{d \geq 0 \ : \ R_{\mathrm{SD}}(d) \leq R\}. \tag{3}$$

The successive degradation rate distortion function $R_{\mathrm{SD}}(d)$ can be related to some other familiar rate distortion functions. First, we show that it cannot in general coincide with the regular rate distortion function $R_{\mathrm{SC}}(d)$. To see this, note that because of the Markov constraint in its definition, we can rewrite the mutual information being minimized in (2) as

$$I(\hat{\mathsf{x}}; \hat{\mathsf{x}}^\circ) = I(\hat{\mathsf{x}}; \mathsf{x}, \hat{\mathsf{x}}^\circ) = I(\mathsf{x}; \hat{\mathsf{x}}) + I(\hat{\mathsf{x}}^\circ; \hat{\mathsf{x}}|\mathsf{x}). \tag{4}$$

Then since the first term on the right-hand side of (4) corresponds to $R_{\mathrm{SC}}(d)$, the successive degradation rate distortion function is $R_{\mathrm{SC}}(d)$ when the second term is zero, i.e., when the additional Markov constraint $\hat{\mathsf{x}}^\circ \leftrightarrow \mathsf{x} \leftrightarrow \hat{\mathsf{x}}$ is satisfied. Under fairly general conditions,[3] this Markov constraint and that in Definition 1 can only hold simultaneously if $\hat{\mathsf{x}}$ is independent of x and $\hat{\mathsf{x}}^\circ$. From this we conclude that, in most cases, $R_{\mathrm{SD}}(d) \neq R_{\mathrm{SC}}(d)$ unless $R_{\mathrm{SC}}(d) = 0$.

[2] Achievability means that (1) implies $I(\mathsf{x}; \hat{\mathsf{x}}^\circ) = R_{\mathrm{SC}}(d_\circ) \triangleq R_\circ$, where $R_{\mathrm{SC}}(\cdot)$ is the rate distortion function for regular source coding. Uniqueness means that no other conditional distribution $q_{\hat{\mathsf{x}}^\circ|\mathsf{x}}(\hat{x}^\circ|x)$ can achieve the same point on the rate distortion function. We assume the sufficient conditions for such uniqueness (see, e.g., [7, Lemma 7]) are met.

[3] A sufficient condition is that $p_{\mathsf{x},\hat{\mathsf{x}}^\circ}(x, \hat{x}^\circ) > 0$ for all $x, \hat{x}^\circ$. This condition is satisfied, for example, in the binary-Hamming and Gaussian-quadratic cases unless $d_\circ = 0$.

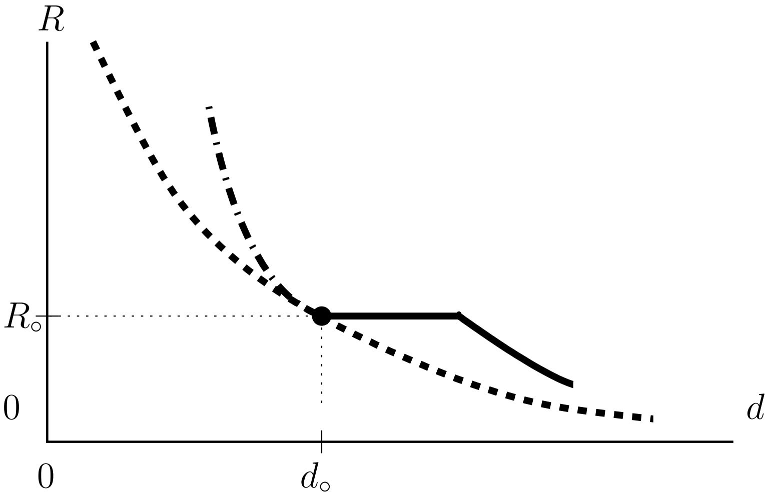


Figure 3: Typical rate distortion functions corresponding to an original operating point $(R_\circ, d_\circ)$. The successive degradation rate distortion function is indicated by the solid curve, and is defined for distortions $d \geq d_\circ$. The successive refinement rate distortion function is indicated by the dash-dotted curve, and is defined for distortions $d \leq d_\circ$. Finally, the regular rate distortion function is indicated by the lower dashed curve.

Similarly, $R_{\mathrm{SD}}(d)$ generally differs from the successive refinement rate distortion function $R_{\mathrm{SR}}(d)$ as well. In particular, [6] shows that a source description at distortion $d_\circ$ using rate $R_\circ = R_{\mathrm{SC}}(d_\circ)$ is *refinable* to a distortion $d \leq d_\circ$ if and only if the total rate of the refined description is at least

$$R_{\mathrm{SR}}(d) = \inf I(\mathsf{x}; \hat{\mathsf{x}}, \hat{\mathsf{x}}^\circ), \tag{5}$$

where the minimization is over all conditional distributions $p_{\hat{\mathsf{x}}|\hat{\mathsf{x}}^\circ,\mathsf{x}}(\hat{x}|\hat{x}^\circ, x)$ such that $E\left[d\left(\mathsf{x}, \hat{\mathsf{x}}\right)\right] \leq d$. Expanding the mutual information in (5) as

$$I(\mathsf{x}; \hat{\mathsf{x}}, \hat{\mathsf{x}}^\circ) = I(\mathsf{x}; \hat{\mathsf{x}}) + I(\mathsf{x}; \hat{\mathsf{x}}^\circ | \hat{\mathsf{x}}) \tag{6}$$

we see, as established in [1], that $R_{\mathrm{SR}}(d)$ differs from $R_{\mathrm{SC}}(d)$ unless a still different Markov constraint $\mathsf{x} \leftrightarrow \hat{\mathsf{x}} \leftrightarrow \hat{\mathsf{x}}^\circ$ is satisfied.

For reference, typical successive degradation and successive refinement rate distortion functions are depicted in Fig. 3.

3.2 The Successive Degradation Game

To determine the best guaranteeable rate distortion trade-off for the successive degradation problem, consider the following zero-sum game. The first player picks a transcoder anticipating the worst possible encoder. The second player picks an encoder designed to be as difficult as possible to transcode while meeting given rate and distortion constraints. In Section 3.3 we show that the information successive degradation rate distortion function $R_{\mathrm{SD}}(d)$ defined in Section 3.1 is the operational rate distortion function for this game.

Specifically, any of a broad class of rate-$R_\circ$ source codes for $\mathbf{x}$ that achieve average distortion $d_\circ$ with $R_\circ$ arbitrarily close to $R_{\mathrm{SC}}(d_\circ)$ can be *degraded* to distortion arbitrarily close to $d > d_\circ$ if and only if the transcoder rate satisfies $R > R_{\mathrm{SD}}(d)$. To develop this result, we first introduce the formal problem setting.

An instance of the successive degradation game consists of the tuple

$$\{\mathcal{X}, p(x), d\left(\cdot,\cdot\right), R_\circ, R\}. \tag{7}$$

The source (and reconstruction) alphabet $\mathcal{X}$ is finite unless otherwise indicated, $p(x)$, $p(\hat{x}^\circ|x)$ and $d:\ \mathcal{X}\times\mathcal{X} \mapsto \mathbb{R}^+$ are as given in Definition 1, and $R_\circ$ and R are rate constraints on the encoder and transcoder that are the inputs to the game. Note, too, that by the uniqueness in Definition 1, $R_\circ$ specifies $p(\hat{x}^\circ|x)$ in the game.

An encoder $\Phi_n^\circ = (\mathcal{C}_n^\circ, W^\circ(\cdot|\cdot))$ consists of a codebook $\mathcal{C}_n^\circ$ of cardinality $|\mathcal{C}_n^\circ| = 2^{nR_\circ}$ and a potentially randomized encoding rule $W^\circ(i|\mathbf{x})$. The latter denotes the probability that the n-length input vector $\mathbf{x}$ is mapped to codebook index $i \in \{1, 2, \ldots, |\mathcal{C}_n^\circ|\}$. The reconstruction corresponding to this encoding is simply the codeword corresponding to the chosen index, i.e., $\hat{\mathbf{x}}^\circ = \mathcal{C}_n^\circ(i)$.

A transcoder $\Phi_n = (\mathcal{C}_n, W(\cdot|\cdot,\cdot))$ is a codebook $\mathcal{C}_n$ of cardinality $|\mathcal{C}_n| = 2^{nR}$ and a potentially randomized rule $W(j|i, \Phi_n^\circ)$. The latter denotes the probability that a given index i produced by the (known) original encoder Φ_n° is mapped to index $j \in \{1, 2, \ldots, |\mathcal{C}_n|\}$. The reconstruction corresponding to this transcoding is simply the codeword corresponding to the chosen index, i.e., $\hat{\mathbf{x}} = \mathcal{C}_n(j)$.

Notationally, we use $d_n\left(\cdot,\cdot\right)$ for the average distortion between length-n sequences, i.e., for any a^n and b^n,

$$d_n\left(a^n, b^n\right) = \frac{1}{n}\sum_{i=1}^{n} d\left(a_i, b_i\right). \tag{8}$$

where $d\left(\cdot,\cdot\right)$ is the per-letter distortion measure, which is bounded unless otherwise indicated. Hence, the distortion in the reconstruction $\hat{\mathbf{x}}^\circ$ associated with encoding is $d_n\left(\hat{\mathbf{x}}^\circ, \mathbf{x}\right)$, and likewise for the reconstruction $\hat{\mathbf{x}}$ associated with transcoding it is $d_n\left(\hat{\mathbf{x}}, \mathbf{x}\right)$. We omit the subscript n when there is no risk of confusion.

We now describe the pay-offs in the successive degradation game for a fixed n. The encoder and transcoder players choose randomized strategies P_{Φ° and P_Φ for generating encoder/decoder and transcoder/decoder pairs, respectively. Given these strategies and a reference distortion level d, the pay-off to the transcoder is given by

$$\pi(P_{\Phi^\circ}, P_\Phi, d) = \Pr\{d\left(\mathbf{x}, \hat{\mathbf{x}}\right) \le d\}. \tag{9}$$

The pay-off to the encoder is simply $-\pi(P_{\Phi^\circ}, P_\Phi, d)$. The probability in (9) is evaluated as follows. First, the source $\mathbf{x}$ is drawn according to $p_{\mathbf{x}}(\mathbf{x}) = \prod_{i=1}^n p(x_i)$. Then, an encoder Φ° is chosen according to P_{Φ° and used to generate $\hat{\mathbf{x}}^\circ$ from $\mathbf{x}$. Finally, a transcoder Φ is chosen according to P_Φ and used to generate $\hat{\mathbf{x}}$ from $\hat{\mathbf{x}}^\circ$ and Φ°. The random choices in these steps are mutually independent.

We consider the asymptotic value of $\pi(P_{\Phi_n^\circ}, P_{\Phi_n}, d)$ for sequences of strategies $\{P_{\Phi_n^\circ}\}$ and $\{P_{\Phi_n}\}$ where the sequence of encoders must achieve the point $(R_\circ, d_\circ)$ in a sense to be defined below and the sequence of transcoders must use rate R. We demonstrate a saddle point for this asymptotic game so that the order of play does not effect the equilibrium pay-offs. Note, however, that by restricting our attention to the case in which the random encoder and transcoder choices are made independently, our results do not address scenarios in which the encoder codebook $\mathcal{C}^\circ$ is a function of the transcoder codebook $\mathcal{C}$, nor vice-versa.

3.3 A Coding Theorem

Intuitively, we expect that just as the familiar source coding rate distortion function depends on the source distribution, the successive degradation rate distortion function should depend on properties of the original source code. We focus on efficient encoders with performance close to the rate distortion bound.

To formalize this notion, we define a class of admissible encoders. In the sequel, we use standard definitions of (and notation for) empirical distributions or types [8] [9]. For example, if $T_{\hat{\mathbf{x}}^\circ,\mathbf{x}}$ denotes the joint type of $(\hat{\mathbf{x}}^\circ, \mathbf{x})$ then $T_{\hat{\mathbf{x}}^\circ,\mathbf{x}}(\hat{x}^\circ, x)$ is the relative frequency of occurrences of the sample-pair $(\hat{x}^\circ, x)$ in the sequences $(\hat{\mathbf{x}}^\circ, \mathbf{x})$.

Definition 2 *Let $T_{\hat{\mathbf{x}}^\circ,\mathbf{x}}$ be the joint type of encoder output $\hat{\mathbf{x}}^\circ$ and source $\mathbf{x}$. A sequence of encoders $\{\Phi_n^\circ\}$ is said to be* admissible *if $D\left(T_{\hat{\mathbf{x}}^\circ,\mathbf{x}} \| p_{\hat{x}^\circ,x}\right) \to 0$ in probability as $n \to \infty$, where $D\left(\cdot\|\cdot\right)$ denotes relative entropy, and $p_{\hat{x}^\circ,x} = p_x(x)\, p_{\hat{x}^\circ|x}(\hat{x}^\circ|x)$.*

An admissible sequence of rate-$R_\circ$ encoders achieves the point $(R_\circ, d_\circ)$ on the rate distortion function. This is because the probability that $\hat{\mathbf{x}}^\circ$ and $\mathbf{x}$ are asymptotically strongly typical according to $p_{\hat{x}^\circ,x}$ approaches one, whence $d\left(\mathbf{x}, \hat{\mathbf{x}}^\circ\right) \to E\left[d\left(x, \hat{x}^\circ\right)\right] = d_\circ$ in probability. To verify the strong typicality claim, it suffices to note that via [9, Lemma 12.6.1, p. 300] we have, for all $(\hat{x}^\circ, x)$,

$$\left|T_{\hat{\mathbf{x}}^\circ,\mathbf{x}}(\hat{x}^\circ, x) - p_{\hat{x}^\circ,x}(\hat{x}^\circ, x)\right| \leq \sqrt{2 \ln 2 \cdot D\left(T_{\hat{\mathbf{x}}^\circ,\mathbf{x}} \| p_{\hat{x}^\circ,x}\right)}. \tag{10}$$

and, moreover, that for all $(\hat{x}^\circ, x)$ with $p_{\hat{x}^\circ,x}(\hat{x}^\circ, x) = 0$ we have $T_{\hat{\mathbf{x}}^\circ,\mathbf{x}}(\hat{x}^\circ, x) = 0$ (otherwise $D\left(T_{\hat{\mathbf{x}}^\circ,\mathbf{x}} \| p_{\hat{x}^\circ,x}\right)$ would be infinite).

Note that the set of admissible encoders is reasonably broad. For example, it includes the familiar strong typicality encoders.[4] Furthermore, extensions to source coding (see [10], and particularly [11, Chap. 2.6]) of analogous results in channel coding [12] tell us that the probability that source sequences are jointly non-typical with their codewords decays exponentially in block length. In this statement, joint typicality is with respect to the joint distribution $p_{\hat{x}^\circ,x}$ induced by the source probability mass function and the (assumed unique) rate distortion achieving channel.

Our main theorem is as follows.

[4] To see that strong typicality implies small divergence, consider a vector $\mathbf{x}$ that is strongly typical with respect to the distribution $p_x(x)$. That is, $|T_{\mathbf{x}}(x) - p_x(x)| < \epsilon$ for all x and some ϵ. Without loss of generality, $p_x(x) \geq p_{\min}$ for all x. Thus,

$$D(T_{\mathbf{x}}||p_x) = \sum_x T_{\mathbf{x}}(x) \log \frac{T_{\mathbf{x}}(x)}{p_x(x)} \leq \sum_x T_{\mathbf{x}}(x) \frac{T_{\mathbf{x}}(x) - p_x(x)}{p_x(x)} \leq \frac{\epsilon}{p_{\min}},$$

where the first inequality follows since $\log x \leq x - 1$ for $x > 0$ and the second inequality follows by strong typicality and the above assumption. With the choice of ϵ, the divergence can be made as small as desired.

Theorem 1 *For the successive degradation game,*

$$\inf_{\{P_{\Phi_n^\circ}\}} \sup_{\{P_{\Phi_n}\}} \lim_{n\to\infty} E\left[\pi(P_{\Phi_n^\circ}, P_{\Phi_n}, d)\right] = \sup_{\{P_{\Phi_n}\}} \inf_{\{P_{\Phi_n^\circ}\}} \lim_{n\to\infty} E\left[\pi(P_{\Phi_n^\circ}, P_{\Phi_n}, d)\right] \tag{11}$$

$$= \begin{cases} 1 & \textit{if } R > R_{\mathrm{SD}}(d) \\ 0 & \textit{if } R \le R_{\mathrm{SD}}(d) \end{cases}, \tag{12}$$

where the minimizations are over admissible sequences of rate-$R_\circ$ encoders and the maximizations are over sequences of rate-R transcoders.

Theorem 1 implies that the information rate distortion function of (2) gives the best possible worst-case successive degradation trade-off. If the transcoder's rate is below (2) then there exists at least one encoder that causes the transcoder to fail. If a rate higher than (2) is used, then there exists a transcoder that almost always wins. The achievability argument is developed in Section 4.3; the converse is developed in Section 4.4.

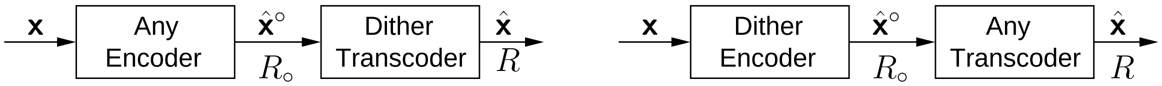
4 Proof of Successive Degradation Rate Distortion Theorem

Intuitively, one can view the encoder output $\hat{\mathbf{x}}^\circ$ as a noisy source observation. If the quantization noise were i.i.d. then results on encoding noisy sources (see, e.g., [13] and the references therein) would give the transcoding rate distortion region. However, the joint distribution of $(\mathbf{x}, \hat{\mathbf{x}}^\circ)$ for good vector quantizers is generally not i.i.d..

Accordingly, we prove Theorem 1 by using a special form of dithered quantization. The joint input/output distribution of quantizers in this class is essentially indistinguishable from an i.i.d. relationship, and yet as we will see their performance approaches the rate distortion bound. The forward part of the theorem is proved using this dithered quantization at the transcoder. This induces an i.i.d.-like joint distribution on the transcoder inputs and outputs, allowing us to use the Markov lemma [14] to guarantee that as long as the the source $\mathbf{x}$ and encoder output $\hat{\mathbf{x}}^\circ$ are strongly typical, then the transcoder output $\hat{\mathbf{x}}$ and the source $\mathbf{x}$ will be also. The converse is shown in a complementary manner. In this case, our dithered quantization is used at the source encoder. No transcoder can do better in this situation than the rate distortion results for quantizing noisy source. The position of dithered quantization in the achievability and converse halves of the proof is indicated in Fig. 4-a and 4-b, respectively.

4.1 Dithered Quantization

The design of a dithered quantizer is governed by an input distribution $p_{\mathsf{u}}(u)$, a quantization distribution $p_{\mathsf{v}|\mathsf{u}}(v|u)$, a quantization rate R^δ, and a parameter $\delta > 0$ that can be arbitrarily small. The resulting quantizer $\Phi_n^\delta = (\mathcal{C}_n^\delta, W_n^\delta(\cdot|\cdot))$ consists of a codebook $\mathcal{C}_n^\delta$ of cardinality $|\mathcal{C}_n^\delta| = 2^{nR^\delta}$ and a quantization rule $W_n^\delta(\cdot|\cdot)$ that avoids joint typicality encoding. We let $\mathbf{u}$ and $\mathbf{v}$ denote the quantizer input and output, respectively.



(a) Achievability

(b) Converse

Figure 4: The role of dithered quantization in the successive degradation rate distortion coding theorem. Dithering is used at the encoder and transcoder in the converse and achievability arguments, respectively.

Dithered Quantization Codebook $\mathcal{C}_n^\delta$ Construction:

1. Generate 2^{nR^δ} sequences of length n in an i.i.d. manner according to $p_{\mathsf{v}}(v) = \sum_u p_{\mathsf{u}}(u) p_{\mathsf{v}|\mathsf{u}}(v|u)$. These are the codewords in the codebook.
2. Label these codewords $\mathbf{v}(1), \mathbf{v}(2), \ldots, \mathbf{v}(2^{nR^\delta})$.

Dithered Quantization Rule $W_n^\delta(\cdot|\cdot)$:

1. Generate a noisy observation $\mathbf{w}$ from the input $\mathbf{u}$ according to the conditional distribution

$$p_{\mathbf{w}|\mathbf{u}}(\mathbf{w}|\mathbf{u}) = \prod_{i=1}^{n} p_{\mathsf{v}|\mathsf{u}}(w_i|u_i). \tag{13}$$

   Denote by $T_{\mathbf{w},\mathbf{u}}$ the joint type of $(\mathbf{w},\mathbf{u})$. Denote by $p_{\mathsf{v},\mathsf{u}}$ the joint probability distribution $p_{\mathsf{u}}(u) p_{\mathsf{v}|\mathsf{u}}(v|u)$.
2. If $D\left(T_{\mathbf{w},\mathbf{u}} \| p_{\mathsf{v},\mathsf{u}}\right) > \delta$ the quantization fails, so choose a codeword at random from the codebook.
3. Otherwise, list all sequences $\mathbf{v}$ in the codebook such that $T_{\mathbf{v},\mathbf{u}} = T_{\mathbf{w},\mathbf{u}}$. If no such sequences exist, the quantization fails, so choose a codeword at random from the codebook.
4. Otherwise, choose a codeword at random from this list. In this case the quantization succeeds.

When they succeed, dithered quantizers have the property that their outputs "look" like the output of a memoryless noisy channel with a channel law given by the quantization distribution. This property is formalized by the following lemma, which establishes that any theorem regarding a noisy observation of the quantizer input also holds for the output of dithered quantizers, even when there is encoder side information. A proof is provided in Appendix A.

Lemma 1 *Consider any binary valued test* $\theta[\cdot] \in \{0, 1\}$*, and a random map* $\boldsymbol{g}(\cdot)$ *with domain* $\mathcal{X}$ *and a finite but arbitrary range. Furthermore, let* $\mathbf{u}$ *and* $\mathbf{v}$ *be the input and output, respectively, of a dithered quantizer, and let* $\mathbf{w}$ *be generated from* $\mathbf{u}$ *in an i.i.d. manner according to the distribution* (13)*. Then*

$$|E\{\theta[\mathbf{v}, \mathbf{u}, \boldsymbol{g}(\mathbf{v})]\} - E\{\theta[\mathbf{w}, \mathbf{u}, \boldsymbol{g}(\mathbf{w})]\}| < 4 \Pr[\mathcal{E}] \tag{14}$$

provided that $\Pr[\mathcal{E}] < 1/2$ *and that the map* $\boldsymbol{g}(\cdot)$ *is independent of* $\mathbf{u}$*,* $\mathbf{v}$*, and* $\mathbf{w}$*, where* $\mathcal{E}$ *denotes the event that the dithered quantization fails.*

4.2 Design and Properties of Dithered Encoders and Transcoders

We now develop specific dithered encoders and dithered transcoders, and some of their key properties that will be useful in the sequel.

In the design of our dithered encoder, we choose the input distribution to be $p_{\mathsf{x}}(x)$, that of the i.i.d. source. Furthermore, we choose the quantization rate to be $R_\circ$ and the quantization distribution to be the conditional distribution $p_{\hat{\mathsf{x}}^\circ|\mathsf{x}}(\hat{x}^\circ|x)$ in Definition 1 that achieves the distortion rate function at the target rate, i.e., $d_{\mathrm{SC}}(R_\circ) = d_\circ$ for the distortion measure $d\left(\cdot,\cdot\right)$ of interest.[5]

Propositions 1 and 2, whose proofs are provided in Appendices B and C, respectively, establish that dithered encoders designed in this way are good, i.e., can perform within ϵ of the rate distortion function, and are admissible in the sense of Definition 2.

Proposition 1 *The distortion for the dithered encoder satisfies*

$$\Pr\left[d\left(\mathbf{x},\hat{\mathbf{x}}^\circ\right) > E\left[d\left(\mathsf{x},\hat{\mathsf{x}}^\circ\right)\right] + d_{\max}\sqrt{2\delta\ln 2}\right] \leq \Pr[\mathcal{E}], \tag{15}$$

where δ is the encoder parameter and $d_{\max} = \max_{a,b\in\mathcal{X}} d\left(a,b\right)$.

Proposition 2 *There exists a $\delta_{\max} > 0$ such that for every $R_\circ > I(\hat{\mathsf{x}}^\circ;\mathsf{x})$ and $\delta \in (0,\delta_{\max})$, $\Pr[\mathcal{E}] \to 0$ exponentially as $n \to \infty$ when the input is an i.i.d. sequence $\mathbf{x}$. Furthermore $\delta_{\max}$ depends only on $R_\circ$ and $p_{\hat{\mathsf{x}}^\circ,\mathsf{x}}(\hat{x}^\circ,x)$.*

We turn next to our dithered transcoders. For their design, we choose the input distribution to be

$$p_{\hat{\mathsf{x}}^\circ}(\hat{x}^\circ) = \sum_x p_{\hat{\mathsf{x}}^\circ|\mathsf{x}}(\hat{x}^\circ|x)\, p_{\mathsf{x}}(x), \tag{16}$$

where $p_{\mathsf{x}}(x)$ is the original source distribution and $p_{\hat{\mathsf{x}}^\circ|\mathsf{x}}(\hat{x}^\circ|x)$ is again the conditional distribution associated with the original source code in Definition 1 corresponding to the distortion rate function operating point $d_{\mathrm{SC}}(R_\circ) = d_\circ$. Furthermore, we choose the quantization rate to be R, and the quantization distribution to be the optimizing $p_{\hat{\mathsf{x}}|\hat{\mathsf{x}}^\circ}$ in (2).

The following proposition, whose proof is provided in Appendix D establishes that dithered transcoders designed in this way are successful.

Proposition 3 *Let $\hat{\mathbf{x}}^\circ$ be the reconstruction corresponding to an admissible rate-$R_\circ$ encoder. Then there exists a $\delta_{\max} > 0$ such that for every $R > I(\hat{\mathsf{x}}^\circ;\hat{\mathsf{x}})$ and $\delta \in (0,\delta_{\max})$, $\Pr[\mathcal{E}] \to 0$ as $n \to \infty$ when the input is $\hat{\mathbf{x}}^\circ$. Furthermore $\delta_{\max}$ only depends on R and $p_{\hat{\mathsf{x}}^\circ,\hat{\mathsf{x}}}(\hat{x}^\circ,\hat{x})$.*

[5] The distortion rate function is defined in terms of the rate distortion function in the usual way: $d_{\mathrm{SC}}(R_\circ) = \inf\{d_\circ \geq 0 : R_{\mathrm{SC}}(d_\circ) \leq R_\circ\}$.

4.3 Successive Degradation Converse

In this section we show that $R_{\mathrm{SD}}(d)$ from (2) gives a lower bound on the best rates that can be guaranteed. To develop this result, it is convenient to express the pay-off (9) for the successive degradation game in the form

$$\Pr\{d(\mathbf{x},\hat{\mathbf{x}}) \le d \mid \Phi^\circ, \Phi\} = E\left\{\theta_{d,\Phi}\left[\hat{\mathbf{x}}^\circ, \mathbf{x}, \Phi^\circ\right]\right\} \tag{17}$$

where

$$\theta_{d,\Phi}\left[\hat{\mathbf{x}}^\circ, \mathbf{x}, \Phi^\circ\right] \triangleq \begin{cases} 1 & \text{if } d(\mathbf{x},\hat{\mathbf{x}}) \le d, \\ 0 & \text{otherwise.} \end{cases} \tag{18}$$

Then since a min-max upper bounds a max-min, to establish the converse of Theorem 1, it suffices to show the following.

Proposition 4 *For every $\epsilon > 0$ and rate $R \le R_{\mathrm{SD}}(d)$ there exists an admissible sequence of randomized encoders (in particular, a sequence of dithered encoders, $\{\Phi_n^\delta\}$) with*

$$\sup_{\{\Phi_n\}} \lim_{n\to\infty} E\left\{\theta_{d_{\mathrm{SD}}(R),\Phi_n}\left[\hat{\mathbf{x}}^\circ, \mathbf{x}, \Phi_n^\delta\right]\right\} \le \epsilon. \tag{19}$$

The expectation is over the source, codebooks, and encoding rules, and the maximization is over all sequences of rate-R transcoders.

Proof:

The key to our proof is showing that the transcoder input "looks" like a noisy observation of the source, and therefore the results of noisy source coding can be applied directly.

First, let $\hat{\mathbf{x}}^\circ$ be an encoding of $\mathbf{x}$ created by the dithered encoder $\Phi_n^\delta = (\mathcal{C}_n^\delta, W_n^\delta(\cdot|\cdot))$. Then we can express the dithered encoder in the form $\Phi_n^\delta = \boldsymbol{g}(\hat{\mathbf{x}}^\circ)$ where $\boldsymbol{g}(\cdot)$ is a random map. In particular, for $\mathbf{v} \in \mathcal{X}$, $\boldsymbol{g}(\mathbf{v})$ consists of a codebook $\mathcal{C}_n^\delta(\mathbf{v})$ with $\mathbf{v}$ at a random entry and the remaining $(2^{nR_\circ} - 1)$ entries generated in a i.i.d. manner independently of $\mathbf{v}$, and a rule $W_n^\delta(\cdot|\cdot)(\mathbf{v})$ that is independent of $\mathbf{v}$. Hence, $\boldsymbol{g}(\cdot)$ is independent of $\mathbf{x}$ and $\hat{\mathbf{x}}^\circ$.

Next, let $\mathbf{y}$ be a random vector generated from $\mathbf{x}$ in an i.i.d. manner according to the distribution $p_{\hat{\mathsf{x}}^\circ|\mathsf{x}}$, independently of $\boldsymbol{g}(\cdot)$. Then, we can use Lemma 1 with $d = d_{\mathrm{SD}}(R)$ along with Proposition 2 to infer that for all n sufficiently large

$$E\left\{\theta_{d_{\mathrm{SD}}(R),\Phi}\left[\hat{\mathbf{x}}^\circ, \mathbf{x}, \boldsymbol{g}(\hat{\mathbf{x}}^\circ)\right]\right\} \le E\left\{\theta_{d_{\mathrm{SD}}(R),\Phi}\left[\mathbf{y}, \mathbf{x}, \boldsymbol{g}(\mathbf{y})\right]\right\} + \epsilon/2 \tag{20}$$

for any choice of transcoder Φ.

Now the expectation on the right-hand side of (20) is the probability that the distortion in quantizing a noisy observation of $\mathbf{x}$ given the associated side information is smaller than $d_{\mathrm{SD}}(R)$. But such encoder side information does not help in source coding problems (see, e.g., [15] [16]), and the rate distortion function for quantizing noisy sources is, in fact, given by $d_{\mathrm{SD}}(R)$ in (2) when $p_{\hat{\mathsf{x}}^\circ|\mathsf{x}}(\hat{x}^\circ|x)$ represents the noisy channel law [17] [13]. This means that no matter how Φ is chosen,

$$E\left\{\theta_{d_{\mathrm{SD}}(R),\Phi}\left[\mathbf{y}, \mathbf{x}, \boldsymbol{g}(\mathbf{y})\right]\right\} \le \epsilon/2. \tag{21}$$

Substituting (21) into the right-hand side of (20) yields

$$E\left\{\theta_{d_{\mathrm{SD}}(R),\Phi}\left[\hat{\mathbf{x}}^\circ, \mathbf{x}, \boldsymbol{g}(\hat{\mathbf{x}}^\circ)\right]\right\} \le \epsilon, \tag{22}$$

from which we obtain (19) as desired. ■

4.4 Successive Degradation Direct Part

In this section we establish that if the transcoding rate is larger than $R_{\mathrm{SD}}(d)$ from (2), then the output of any admissible encoder can be transcoded successfully.

Since a max-min lower bounds a min-max, to establish this forward part of Theorem 1 it suffices to show the following.

Proposition 5 *For every $\epsilon > 0$ and rate $R > R_{\mathrm{SD}}(d)$ there exists a sequence of dithered transcoders $\{\Phi_n^\delta\}$ with rates $R_{\mathrm{SD}}(d)$ such that*

$$\inf_{\{\Phi_n^\circ\}} \lim_{n\to\infty} E\left\{\theta_{d,\Phi_n^\delta}\left[\hat{\mathbf{x}}^\circ, \mathbf{x}, \Phi_n^\circ\right]\right\} \geq 1-\epsilon. \tag{23}$$

The expectation is over the source, codebooks, and encoding and transcoding rules, and the minimization is over all admissible sequences of encoders with rate $R_\circ$.

Proof:

Let the transcoder in this case be a dithered transcoder as designed in Sections 4.1 and 4.2. Fix $\epsilon > 0$, then according to Proposition 3, this dithered transcoder succeeds in transcoding $\hat{\mathbf{x}}^\circ$ to $\hat{\mathbf{x}}$ with probability at least $1-\epsilon$.

Therefore $\mathbf{x}$ and $\hat{\mathbf{x}}^\circ$ are strongly typical by the definition of admissible encoders and (10). Furthermore the distribution of $(\hat{\mathbf{x}}^\circ, \hat{\mathbf{x}})$ is indistinguishable from an i.i.d. distribution according to Lemma 1, so $\hat{\mathbf{x}}^\circ$ and $\hat{\mathbf{x}}$ are also strongly typical. Thus, according to the Markov lemma [9], $\mathbf{x}$ and $\hat{\mathbf{x}}$ are strongly typical according to the distribution

$$p(\hat{x}, x) = \sum_{\hat{x}^\circ} p(\hat{x}|\hat{x}^\circ) p(\hat{x}^\circ|x) p(x).$$

Hence the distortion will be close to $E\left[d\left(x, \hat{x}\right)\right]$. ■

5 Binary-Hamming Case

As we now show by example, the successive degradation rate distortion function is generally different from the regular rate distortion function.

Consider a binary (i.e., Bernoulli-p) source and a Hamming distortion measure:

$$d\left(u, v\right) = \begin{cases} 1 & u \neq v \\ 0 & u = v \end{cases}, \tag{24}$$

so the (average) distortion range of interest is the interval $[0, 1/2]$. In this case, the optimal reverse test channel for the original quantization is a binary symmetric channel with crossover probability $d_\circ$. It is straightforward to verify that the optimizing successive degradation reverse test channel takes the form of a binary symmetric channel as well, but with crossover probability $0 \leq \gamma \leq 1/2$. The corresponding rate and overall distortion are, respectively, $R = H_{\mathrm{B}}\left(q\right) - H_{\mathrm{B}}\left(\gamma\right)$, with $H_{\mathrm{B}}\left(\cdot\right)$ denoting the binary entropy function,

$q = \Pr[\hat{x}^\circ = 1]$, and $d = d_\circ + \gamma(1-2d_\circ)$, so the rate-distortion trade-off is

$$R_{\mathrm{SD}}(d) = H_{\mathrm{B}}\left(\frac{p-d_\circ}{1-2d_\circ}\right) - H_{\mathrm{B}}\left(\frac{d-d_\circ}{1-2d_\circ}\right). \tag{25}$$

By comparison, the corresponding regular rate distortion bound for this source and distortion measure is

$$R_{\mathrm{SC}}(d) = H_{\mathrm{B}}(p) - H_{\mathrm{B}}(d). \tag{26}$$

Since, without loss of generality we may assume $0 \le d \le p \le 1/2$, $R_{\mathrm{SD}}(d) \ge H_{\mathrm{B}}(p) - H_{\mathrm{B}}((d-d_\circ)/(1-2d_\circ))$, and since $H_{\mathrm{B}}(\cdot)$ is strictly monotonic on $[0,1/2]$, it follows that $R_{\mathrm{SD}}(d) > R_{\mathrm{SC}}(d)$ whenever $d_\circ \neq 0$.

For some range of distortions above $d_\circ$, requantization via (25) does not yield rate savings and transcoding is better avoided. Thus, we can conclude from (25) that

$$R_{\mathrm{SD}}(d) \ge \min\left\{H_{\mathrm{B}}\left(\frac{p-d_\circ}{1-2d_\circ}\right) - H_{\mathrm{B}}\left(\frac{d-d_\circ}{1-2d_\circ}\right), R_\circ\right\} \tag{27}$$

with $R_\circ = R_{\mathrm{SC}}(d_\circ)$ given by (26). By equating the two terms in (27) and solving for d, we obtain the distortion threshold d_* below which transcoding should be avoided. A simple lower bound on d_* is found by lower bounding the first term of (27) with $H_{\mathrm{B}}(p)$, and setting the result equal to $R_\circ$. This gives

$$d_* \ge 2d_\circ(1-d_\circ). \tag{28}$$

It remains only to show that there is no better successive degradation rate distortion trade-off than (27) that can be achieved. To see this, first note that a dithered encoder will produce an output that is indistinguishable from a noisy version of the source where the noise is generated by a binary asymmetric forward test channel. The crossover probabilities for this channel can be determined from Bayes rule, the prior $\Pr[x = 1] = p$, and the crossover probability γ of the symmetric reverse test channel. The results of [13] confirm for the symmetric case ($p = 1/2$) that the rate distortion function for this source corrupted by such noise is indeed given by (25).

We next outline the generalization of the approach in [13] needed for general binary sources where $p \neq 1/2$. We consider asymmetric sources and binary asymmetric channel observations. The generalization follows the approach in [13], except that the two types of bit error events (cf. eq. (38) of [13]) are now weighted by the (non-equal) prior probabilities of the symbols to be encoded. However, one can check that this weighting corresponds to the same uneven likelihood of bit errors that we induce by using a symmetric reverse test channel with crossover probability γ for successive degradation. This is exactly the effect that when encoding binary asymmetric sources source symbols with higher apriori likelihood are more likely to be flipped. Thus, again (25) is the rate distortion function for this source and (27) is the best rate that can be achieved.[6]

[6] While time-sharing can in general expand a rate distortion function to include its lower convex envelope, the resulting distortion cannot be achieved within a block, but only averaged over multiple blocks with codebooks of different rates. While

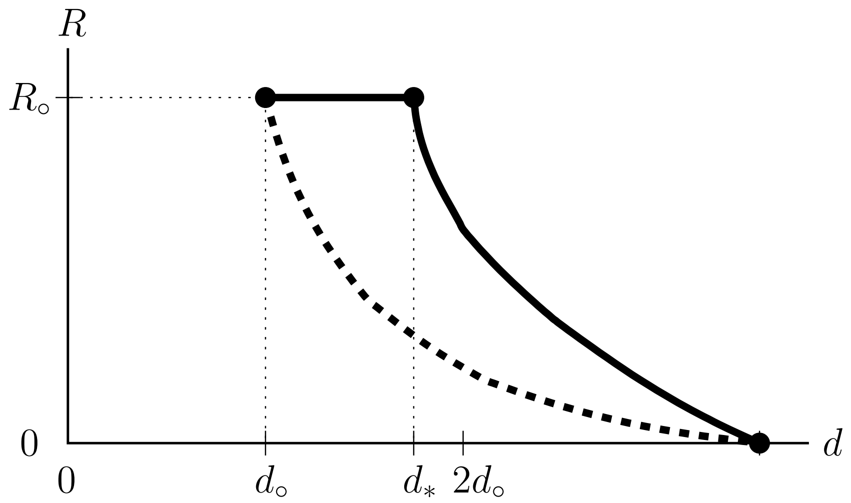


Figure 5: Common form of the rate distortion function for the symmetric ($p = 0$) binary-Hamming and Gaussian-quadratic cases. The successive degradation rate distortion function is indicated by the solid curve, and d_* is the threshold above which transcoding should be performed. The dashed curve indicates the lower bound corresponding to the regular rate distortion function. The distortion at which both the dashed and solid curves intersect the axis ($R = 0$) is 1/2 in the binary-Hamming case, and σ_x^2 in the Gaussian-quadratic case.

The associated rate gap for this example takes the form depicted in Fig. 5 for the symmetric case ($p = 1/2$). The successive degradation and regular rate distortion functions are depicted with the solid and dashed curves, respectively, and intersect at distortion $d = 1/2$.

6 Gaussian-Quadratic Case

In this section we develop the rate distortion expression (2) for Gaussian sources under a quadratic distortion measure: $d(a, b) = (a - b)^2$. We also extend the class of admissible encoders for achievability results in the Gaussian quadratic case to all encoders that achieve $d_\circ$ and not just those encoders that satisfy the divergence condition of Definition 2. This shows that Gaussian-quadratic transcoding is robust in the sense that any good source code for the Gaussian-quadratic case can be successfully transcoded.

The proofs in Section 4 assumed that the signals were drawn from finite alphabets and that all distortion measures were bounded. This simplified the development, but the results can be generalized to continuous alphabets with unbounded distortion. Among other subtleties, care must be taken to preserve the Markov relationship $\mathsf{x} \leftrightarrow \hat{\mathsf{x}}^\circ \leftrightarrow \hat{\mathsf{x}}$. Such techniques are developed in, e.g., [18]. However, in the sequel we establish the achievable transcoding trade-off for the Gaussian-quadratic scenario more directly.

we do not consider such variable rate systems in this paper, the associated expanded rate distortion functions are readily computed.

6.1 Rate Distortion Function

For an i.i.d. zero-mean Gaussian source x of variance σ_{x}^2, and an original source code with average mean-square reconstruction distortion $d_\circ$, the information successive degradation rate distortion function given by (2) is, for $d \geq d_\circ$,

$$R_{\mathrm{SD}}(d) = \min\left\{\frac{1}{2}\log\left(\frac{\sigma_{\mathsf{x}}^2 - d_\circ}{d - d_\circ}\right), R_\circ\right\}. \tag{29}$$

where, according to the familiar Gaussian-quadratic rate distortion function,

$$R_\circ = \frac{1}{2}\log\left(\frac{\sigma_{\mathsf{x}}^2}{d_\circ}\right). \tag{30}$$

This successive degradation rate distortion function also takes the form shown in Fig. 5. Again, the successive degradation and regular rate distortion functions are depicted with the solid and dashed curves, respectively, and intersect at distortion $d = \sigma_{\mathsf{x}}^2$.

Eq. (29) is obtained as follows. The usual conditional distribution for achieving the rate distortion bound for Gaussian sources, which we assume to be unique, corresponds to $\hat{\mathsf{x}}^\circ = \alpha\mathsf{x} + \mathsf{e}_\circ$, where

$$\alpha = (1 - d_\circ/\sigma_{\mathsf{x}}^2), \tag{31}$$

and where $\mathsf{e}_\circ$ is a zero-mean Gaussian random variable with variance $\alpha d_\circ$ that is independent of x. Since $\hat{\mathsf{x}}^\circ$ is Gaussian, we know from [19, Lemma A.3] that the optimum $\hat{\mathsf{x}}$ and $\hat{\mathsf{x}}^\circ$ must also be jointly Gaussian. Optimizing the specific moments we conclude that second conditional distribution is of the form $\hat{\mathsf{x}} = \gamma\hat{\mathsf{x}}^\circ + \mathsf{e}$, where $\gamma = (1 - \Delta/\sigma_{\hat{\mathsf{x}}^\circ}^2)$, and where e is a zero-mean Gaussian random variable with variance $\gamma\Delta$ that is independent of $\hat{\mathsf{x}}^\circ$. Combining the two conditional distributions results in a rate

$$I(\hat{\mathsf{x}}; \hat{\mathsf{x}}^\circ) = \begin{cases} \frac{1}{2}\log\left(\frac{\sigma_{\mathsf{x}}^2 - d_\circ}{\Delta}\right) & \Delta > 0 \\ \frac{1}{2}\log\left(\frac{\sigma_{\mathsf{x}}^2}{d_\circ}\right) & \Delta = 0 \end{cases} \tag{32}$$

with overall mean-square distortion

$$d = E\left[(\hat{\mathsf{x}} - \mathsf{x})^2\right] = d = d_\circ + \Delta. \tag{33}$$

Substituting for Δ in (32) using (33), and we obtain (29).

The minimization in (29) suggests that there is a range of distortions above $d_\circ$ for which requantization will not save any rate, i.e., it is better to leave the original source code as is. This is inherently a result of the discrete nature of the input to the transcoder. To identify the distortion threshold d_* at which successive degradation becomes effective, we equate the two terms in (29) and solve for d, yielding

$$d_* = d_\circ\left(2 - \frac{d_\circ}{\sigma_{\mathsf{x}}^2}\right). \tag{34}$$

In the Gaussian-quadratic scenario, the successive degradation rate loss is at most 1/2 bit, which we see as follows. To begin, note that $1 \leq d_*/d_\circ \leq 2$, with $d_*/d_\circ \to 2$ as $R_\circ \to \infty$. Thus, in the high resolution limit, each time the source is requantized with an independently generated source code of the same rate, the overall distortion increases by roughly $d_\circ$. Since the rate loss in successive degradation is largest in this high resolution limit, an upper bound on the loss occurs at $d = d_*$:

$$R_{\mathrm{SD}}(d_*) - R_{\mathrm{SC}}(d_*) \leq \frac{1}{2}\log\left(\frac{\sigma_x^2}{d_\circ}\right) - \frac{1}{2}\log\left(\frac{\sigma_x^2}{d_\circ(2 - d_\circ/\sigma_x^2)}\right) = \frac{1}{2}\log\left(2 - \frac{d_\circ}{\sigma_x^2}\right) < \frac{1}{2}. \tag{35}$$

where the last inequality follows from the fact that $d_\circ \geq 0$. Recall that, by comparison, there is no rate loss with successive refinement in this Gaussian-quadratic scenario [1]. This is because the the full original source signal — not just a quantized version of it — is available when selecting the second codeword. Finally, from Fig. 5 and (34), we see that the corresponding distortion gap is at most 3 dB:

$$\frac{d_*}{d_\circ} = 2 - \frac{d_\circ}{\sigma_x^2} < 2. \tag{36}$$

6.2 Converse

The operational successive degradation rate distortion function is lower bounded by (29). To see this, it suffices to use entropy coded dithered quantization (ECDQ) [20] for the original source encoder. This is the natural counterpart in the Gaussian-quadratic case of the finite-alphabet dithered quantization strategy introduced in Section 4.1. With this source encoder, the quantizer becomes asymptotically indistinguishable in an appropriate sense from an additive white Gaussian noise channel [21], and has output variance $\sigma_x^2 - d_\circ$. In particular, the quantized output is indistinguishable from this channel to the transcoder, and thus the best transcoder is a source encoder for noisy sources. Thus, in the Gaussian-quadratic case the resulting rate-distortion trade-off is given by the first term in (29).

6.3 Achievability

We now show that (29) is also an upper bound on the operational successive degradation rate distortion function. Moreover, we allow the original source encoder to be *any* good one. In particular, we don't restrict our attention to some admissible subset as we did earlier. In this sense, our results in this case are inherently more robust.

Specifically, we show that for any rate-$R_\circ$ original encoder operating close to its optimal distortion level

$$d_{\mathrm{SC}}(R_\circ) = \sigma_x^2 2^{-2R_\circ} \tag{37}$$

a rate-R transcoder can be designed so that it operates close to the information successive degradation distortion rate function

$$d_{\mathrm{SD}}(R) = d_{\mathrm{SC}}(R_\circ) + (\sigma_x^2 - d_{\mathrm{SC}}(R_\circ))2^{-2R}. \tag{38}$$

obtained from (3) with (29).

We quantify this argument in the following theorem, which makes use of a basic result from [22].

Theorem 2 *Let $\mathbf{x}$ be a length-n i.i.d. sequence of Gaussian random variables with variance σ_x^2. For any $\epsilon > 0$ and any rate-$R_\circ$ original source code with*

$$\frac{1}{n}E\left[\|\mathbf{x}-\hat{\mathbf{x}}^\circ\|^2\right] \leq d_{\mathrm{SC}}(R_\circ)+\epsilon, \tag{39}$$

there exists a rate-R transcoder with

$$\frac{1}{n}E\left[\|\mathbf{x}-\hat{\mathbf{x}}\|^2\right] \leq d_{\mathrm{SD}}(R)+\epsilon+\delta(n), \tag{40}$$

where $\delta(n) \to 0$ as $n \to \infty$.

Proof:

To obtain this result, we first bound the variance of a processed version of the output of the original coder. We then use a result from [22] on quantizing sources given only second order statistics.

Let $\hat{\mathbf{x}}'$ be the minimum mean-square error (MMSE) estimate of $\mathbf{x}$ given $\hat{\mathbf{x}}^\circ$. Since $\hat{\mathbf{x}}'$ is a reconstruction of $\mathbf{x}$ based on $nR_\circ$ bits and since $\hat{\mathbf{x}}'$ estimates $\mathbf{x}$ (in the mean-square sense) at least as well as $\hat{\mathbf{x}}^\circ$, we know that

$$d_{\mathrm{SC}}(R_\circ) \leq \frac{1}{n}E\left[\|\mathbf{x}-\hat{\mathbf{x}}'\|^2\right] \leq \frac{1}{n}E\left[\|\mathbf{x}-\hat{\mathbf{x}}^\circ\|^2\right]. \tag{41}$$

Furthermore, the error in the MMSE estimate is uncorrelated with the reconstruction itself, i.e.,

$$E\left[\langle(\mathbf{x}-\hat{\mathbf{x}}'),\hat{\mathbf{x}}'\rangle\right]=0, \tag{42}$$

whence, by Pythagoras' Theorem,

$$\begin{aligned}\frac{1}{n}E\left[\|\hat{\mathbf{x}}'\|^2\right] &= \frac{1}{n}E\left[\|\mathbf{x}\|^2\right]-\frac{1}{n}E\left[\|\mathbf{x}-\hat{\mathbf{x}}'\|^2\right] && (43)\\ &\leq \sigma_x^2-d_{\mathrm{SC}}(R_\circ), && (44)\end{aligned}$$

where (43) follows from (42), and (44) follows from the left-hand inequality in (41) and the fact that $\mathbf{x}$ is an i.i.d. variance-σ_x^2 Gaussian sequence.

Consider a random transcoder which encodes $\hat{\mathbf{x}}'$ by mapping it to the element $\hat{\mathbf{x}}$ of a rate-R random Gaussian codebook that is closest in Euclidean distance. Regardless of the distribution of $\hat{\mathbf{x}}'$, via [22, Theorem 3] we know there exists a deterministic transcoder with output $\hat{\mathbf{x}}$ such that

$$\frac{1}{n}E\left[\|\hat{\mathbf{x}}'-\hat{\mathbf{x}}\|^2\right] \leq \frac{1}{n}E\left[\|\hat{\mathbf{x}}'\|^2\right]2^{-2R}+\delta(n), \tag{45}$$

where $\delta(n) \to 0$ as $n \to \infty$. Due to the structure of the two encodings we have the Markov chain $\mathbf{x} \leftrightarrow \hat{\mathbf{x}}' \leftrightarrow \hat{\mathbf{x}}$ and hence $(\mathbf{x}-\hat{\mathbf{x}}') \leftrightarrow \hat{\mathbf{x}}' \leftrightarrow \hat{\mathbf{x}}$ as well. With this latter Markov chain we conclude from the optimality properties of MMSE estimates that (42) yields

$$E\left[\langle(\mathbf{x}-\hat{\mathbf{x}}'),\hat{\mathbf{x}}\rangle\right]=0. \tag{46}$$

To complete the proof, it suffices to observe that

$$\begin{aligned}\frac{1}{n}E\left[\|\mathbf{x}-\hat{\mathbf{x}}\|^2\right] &= \frac{1}{n}E\left[\|(\mathbf{x}-\hat{\mathbf{x}}')+(\hat{\mathbf{x}}'-\hat{\mathbf{x}})\|^2\right]\\ &= \frac{1}{n}E\left[\|\mathbf{x}-\hat{\mathbf{x}}'\|^2\right]+\frac{1}{n}E\left[\|\hat{\mathbf{x}}'-\hat{\mathbf{x}}\|^2\right] && (47)\\ &\leq d_{\mathrm{SC}}(R_\circ)+(\sigma_x^2-d_{\mathrm{SC}}(R_\circ))2^{-2R}+\epsilon+\delta(n), && (48)\end{aligned}$$

where to obtain (47) we have used the orthogonality implied by (42) and (46), and where to obtain (48) we have used the right-hand inequality in (41) with (39), and (44) with (45). ■

7 Continuous Sources in the High-Resolution Limit

In this section we show that the 1/2-bit gap bound (35) in the high-resolution limit with a quadratic distortion measure holds not just for Gaussian sources as developed in Section 6, but in fact for all sources with finite differential entropy and at least one finite moment. In fact, we believe that the high-resolution limit is the worst case and that the gap is at most 1/2-bit for all distortions.

First, from [23], we know that the successive refinement rate distortion function is within 1/2-bit of the regular rate distortion function for this scenario. Thus, it remains only to show that the successive degradation rate distortion function also lies within 1/2-bit of the regular rate distortion function as well, which we develop in the sequel.

Our proof considers two separate regimes. For $d < 2d_\circ$ we stay within the 1/2-bit bound by avoiding transcoding altogether, whereby $R = R_{\text{SC}}(d_\circ)$. To see this, note that the rate loss in this case is largest as $d_\circ \to 0$, given by

$$R_{\text{SC}}(d_\circ) - R_{\text{SC}}(d) \to \frac{1}{2}\log\left(\frac{d}{d_\circ}\right) \leq \frac{1}{2}, \tag{49}$$

where to obtain the limit in (49) we have used the asymptotic tightness of the Shannon lower bound [24]:

$$\lim_{\delta\to 0} R_{\text{SC}}(\delta) = h(\mathsf{x}) - \frac{1}{2}\log 2\pi e\delta, \tag{50}$$

with $h(\cdot)$ denoting differential entropy.

For the regime $d \geq 2d_\circ$, we use the following argument. First, we let $\hat{\mathsf{x}} = \hat{\mathsf{x}}^\circ + \mathsf{e}$ where e is a zero-mean Gaussian random variable with variance $d - d_\circ$ that is independent of $\hat{\mathsf{x}}^\circ$ and x, so the distortion is, as required,

$$\begin{aligned} E\left[(\mathsf{x}-\hat{\mathsf{x}})^2\right] &= E\left[(\mathsf{x}-\hat{\mathsf{x}}^\circ+\hat{\mathsf{x}}^\circ-\hat{\mathsf{x}})^2\right] && (51)\\ &= E\left[(\mathsf{x}-\hat{\mathsf{x}}^\circ)^2\right] + E\left[(\hat{\mathsf{x}}^\circ-\hat{\mathsf{x}})^2\right] + 2E\left[(\mathsf{x}-\hat{\mathsf{x}}^\circ)(\hat{\mathsf{x}}^\circ-\hat{\mathsf{x}})\right] && (52)\\ &= d_\circ + E\left[\mathsf{e}^2\right] + 2E\left[(\mathsf{x}-\hat{\mathsf{x}}^\circ)\,\mathsf{e}\right] && (53)\\ &= d_\circ + (d-d_\circ) + 0 && (54)\\ &= d && (55) \end{aligned}$$

where to obtain (53) and (54) we have used the definition and properties, respectively, of e. Then from (2) we see the associated rate is

$$\begin{aligned} R_{\mathrm{SD}}(d) &\leq I(\hat{\mathsf{x}};\hat{\mathsf{x}}^\circ) && (56)\\ &= h(\hat{\mathsf{x}}) - h(\hat{\mathsf{x}}|\hat{\mathsf{x}}^\circ) && (57)\\ &= h(\hat{\mathsf{x}}^\circ + \mathsf{e}) - h(\hat{\mathsf{x}}^\circ + \mathsf{e}|\hat{\mathsf{x}}^\circ) && (58)\\ &= h(\hat{\mathsf{x}}^\circ + \mathsf{e}) - h(\mathsf{e}), && (59) \end{aligned}$$

where to obtain (59) we have used that e and $\hat{\mathsf{x}}^\circ$ are independent. In turn, we can bound the rate loss as $d \to 0$ via

$$\begin{aligned} R_{\mathrm{SD}}(d) - R_{\mathrm{SC}}(d) &\leq h(\hat{\mathsf{x}}^\circ + \mathsf{e}) - h(\mathsf{e}) - R_{\mathrm{SC}}(d) && (60)\\ &= [h(\hat{\mathsf{x}}^\circ + \mathsf{e}) - h(\hat{\mathsf{x}}^\circ)] + [h(\hat{\mathsf{x}}^\circ) - h(\mathsf{x})] + [h(\mathsf{x}) - R_{\mathrm{SC}}(d)] - h(\mathsf{e}) && (61)\\ &\to 0 + 0 + \frac{1}{2}\log 2\pi e d - h(\mathsf{e}) && (62)\\ &= \frac{1}{2}\log\left(\frac{d}{d-d_\circ}\right) && (63)\\ &\leq \frac{1}{2}, && (64) \end{aligned}$$

where to obtain (60) we have used (59), and where to obtain the first and third terms in (62) we have used, respectively, the continuity of $h(\cdot)$ in [24, Theorem 1], and the high-resolution rate distortion function [24], and where to obtain (63) we have used that the differential entropy of a Gaussian random variable z of variance σ^2 is $h(\mathsf{z}) = \frac{1}{2}\log 2\pi e\sigma^2$.

To obtain that the second term in (62) is zero, it suffices to let $\hat{\mathsf{x}}^\circ = \mathsf{x} + \mathsf{e}_\circ$ where $\mathsf{e}_\circ$ is a zero-mean Gaussian random variable with variance $d_\circ$ that is independent of x, and again exploit the continuity of $h(\cdot)$ [24, Theorem 1] to conclude

$$h(\hat{\mathsf{x}}^\circ) - h(\mathsf{x}) = h(\mathsf{x} + \mathsf{e}_\circ) - h(\mathsf{x}) \to 0 \text{ as } d_\circ \to 0. \qquad (65)$$

Thus, it remains only to check that the conditional distribution associated this definition of $\mathsf{e}_\circ$ asymptotically achieves the rate distortion function $R_{\mathrm{SC}}(d_\circ)$. To see this, as $d \to 0$, whence $d_\circ \to 0$ since $d \geq d_\circ \geq 0$, we have

$$\begin{aligned} R_{\rm SC}(d) &\leq I(\hat{\mathsf{x}}^\circ, \mathsf{x}) && (66)\\ &= h(\hat{\mathsf{x}}^\circ) - h(\hat{\mathsf{x}}^\circ | \mathsf{x}) && (67)\\ &= h(\mathsf{x} + \mathsf{e}_\circ) - h(\mathsf{x} + \mathsf{e}_\circ | \mathsf{x}) && (68)\\ &= h(\mathsf{x} + \mathsf{e}_\circ) - h(\mathsf{e}_\circ) && (69)\\ &= \left[h(\mathsf{x}) - \frac{1}{2}\log 2\pi e d_\circ \right] + \left[h(\mathsf{x} + \mathsf{e}_\circ) - h(\mathsf{x}) \right] && (70)\\ &\to \left[h(\mathsf{x}) - \frac{1}{2}\log 2\pi e d_\circ \right] + 0 && (71)\\ &= \lim_{d_\circ \to 0} R_{\rm SC}(d_\circ), && (72) \end{aligned}$$

where to obtain (68), (69) and (70) we have used the definition and properties of $\mathsf{e}_\circ$, where to obtain (71) we have used the continuity of differential entropy, and where to obtain (72) we have used (50).

8 Embedding in a Quantized Source

When bit stealing is accomplished through rate splitting and successive degradation, the transcoder is given the freedom to design an entirely new codebook. This requires that the ultimate destination(s) for the quantized source be informed that transcoding has taken place so that the destination(s) can decode using the new codebook.

However, in a number of scenarios it may be either impractical or inconvenient to inform the decoder when bit stealing has taken place. Such is the case, for example, when there is an installed base of legacy source decoders in a network, or, as another example, when the bit-stealing is be covert, in which case no cooperation between transcoders and source decoders is possible. In these and other such cases, there is a need for bit stealing techniques in which the transcoder output lies in the same codebook as its input.

One natural approach to bit stealing with this constraint is based on the use of information embedding ideas. In this section, we will show that bit stealing systems of this type can, in fact, be as efficient as those implemented through rate splitting and successive degradation. From this we can conclude that the transcoder codebook constraint need not incur a loss in performance. To develop this result, it suffices to restrict our attention to the case where our original source $\mathbf{x}$ has been encoded using the classical random codebook and joint typicality encoding rule.

As depicted in Fig. 6, bit stealing via embedding is implemented as follows. The transcoder embeds a message m of rate r into the index corresponding to source reconstruction codeword $\hat{\mathbf{x}}^\circ$, which is the "host." A source decoder generates the reconstruction $\hat{\mathbf{x}}$ from the received bits. Recall that informed source decoders are aware of any transcoding that has taken place — specifically, they know what rate $0 \leq r \leq R_\circ$ has been stolen. Informed decoders can exploit this information in reconstructing the source,

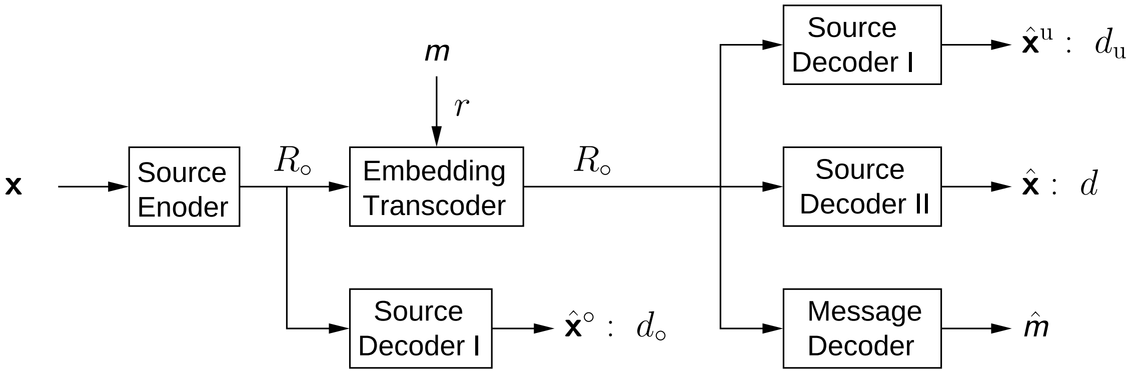


Figure 6: Bit stealing via information embedding: a message m of rate r is embedded into a source quantized to rate $R_\circ$. Source decoder I is the decoder used in the absence of transcoding, or equivalently the uninformed decoder when transcoding has taken place, in which case it produces reconstruction $\hat{\mathsf{x}}^{\mathrm{u}}$ at distortion d_{u}. Source decoder II is the decoder that is informed of the rate of any embedding that has taken place, and produces reconstruction $\hat{\mathsf{x}}$ at distortion d. The message decoder produces a reliable estimate $\hat{\mathsf{m}}$ of the embedded message with high probability.

while uninformed decoders operate as if no embedding had taken place. In the sequel we consider both informed and uninformed source decoders. A message decoder reconstructs the embedded message bits from the received bits.

A naive embedding approach would treat the source reconstruction $\hat{\mathsf{x}}^\circ$ as the host, i.e., as the "dirty paper" of [25] and use, e.g., the associated random binning code or its constructive counterparts in the form of quantization index modulation (QIM) [26] [27] [28]. Such approaches fail to exploit that $\hat{\mathsf{x}}^\circ$ is a codeword of a finite-rate codebook. However, for embedding approaches that do take such characteristics into account, the following rate distortion trade-off can be achieved.

In constructing our result, we continue to focus for simplicity on an i.i.d. source — as with successive degradation — so an instance of the problem continues to consist of the tuple (7), where as before $R_\circ$ implicitly defines $p_{\hat{\mathsf{x}}^\circ|\mathsf{x}}$, but where, now, $R = R_\circ - r$. And as before $p_{\hat{\mathsf{x}}^\circ|\mathsf{x}}(\hat{x}^\circ|x)$ denotes the conditional distribution characterizing the rate distortion function at distortion $d_\circ$ where $R_\circ = R_{\mathrm{SC}}(d_\circ)$.

Theorem 3 *For a source x quantized at rate $R_\circ$ (via typicality encoding) to a codeword $\hat{\mathsf{x}}^\circ$ in a codebook generated randomly according to $p_{\hat{\mathsf{x}}^\circ|\mathsf{x}}(\hat{x}^\circ|x)$, a distortion arbitrarily close to $d > d_\circ$ is achievable if $r < R_{\mathrm{IE}}(d)$ where [cf. (2)]*

$$R_{\mathrm{IE}}(d) = R_\circ - \inf I(\mathsf{u}; \hat{\mathsf{x}}^\circ), \tag{73}$$

where the minimization is over all conditional distributions $p_{\mathsf{u}|\hat{\mathsf{x}}^\circ}(u|\hat{x}^\circ)$ and functions $f : \mathcal{X} \mapsto \mathcal{X}$ such that the Markov constraint $\mathsf{x} \leftrightarrow \hat{\mathsf{x}}^\circ \leftrightarrow \mathsf{u}$ is satisfied, $p_{\hat{\mathsf{x}}^\circ}(x) = p_{\mathsf{u}}(x)$, and $E\left[d\left(\mathsf{x}, \hat{\mathsf{x}}\right)\right] \leq d$, where $\hat{\mathsf{x}} = f(\mathsf{u})$ and u is an auxiliary random variable with alphabet $\mathcal{X}$.

This theorem applies when the decoder is informed of the embedding, and $f(\cdot)$ can be viewed as a distortion compensation function. Achievable rates for the case in which the decoder is uninformed are obtained by constraining $f(\cdot)$ to be the identity function.

Before proving Theorem 3, we introduce some additional notation. Specifically, $d_{\mathrm{IE}}(r)$ is the distortion rate function corresponding to (73), i.e.,

$$d_{\mathrm{IE}}(r) = \inf\{d \geq 0 \ : \ R_{\mathrm{IE}}(d) \geq r\}. \tag{74}$$

Furthermore, we let $R^{\mathrm{u}}_{\mathrm{IE}}(\cdot)$ and $d^{\mathrm{u}}_{\mathrm{IE}}(\cdot)$ denote the corresponding rate distortion and distortion rate functions for *uninformed* decoders.

We establish Theorem 3 by construction. In particular, the original source is quantized using the randomly generated codebook. In particular, the source codebook $\mathcal{C}^\circ$ consists of $2^{nR_\circ}$ sequences of length-n generated according to $\prod_{i=1}^n p_{\hat{\mathsf{x}}^\circ}(\hat{x}_i^\circ)$. These sequence are labeled $\hat{\mathbf{x}}^\circ(1), \hat{\mathbf{x}}^\circ(2), \ldots, \hat{\mathbf{x}}^\circ(2^{nR_\circ})$. To encode, we find the index i such that $(\mathbf{x}, \hat{\mathbf{x}}^\circ(i)) \in T_{\mathsf{x},\hat{\mathsf{x}}^\circ}$ If there is more than one such index, choose any one of them. Transmit that i. If there is no such index, we declare an encoding failure.

The information embedding is implemented as follows. First, a message code is constructed whereby, for any $\epsilon > 0$, we randomly bin $\mathcal{C}^\circ$ into 2^{nr} subcodes $\mathcal{C}^\circ_j$ where $r = R_\circ - I(\mathsf{u}; \hat{\mathsf{x}}^\circ) - \epsilon$. Specifically, for each $\hat{\mathbf{x}}^\circ(i)$ we pick an index j uniformly distributed over $\{1, 2, \ldots, 2^{nr}\}$ and assign $\hat{\mathbf{x}}^\circ(i)$ to subcode $\mathcal{C}^\circ_j$. On average there are $2^{n(R_\circ - r)}$ sequences $\hat{\mathbf{x}}^\circ$ in each $\mathcal{C}^\circ_j$. We re-label the $\hat{\mathbf{x}}^\circ$ sequences in $\mathcal{C}^\circ$ as $\mathbf{u}(j,k)$ where $j \in \{1, 2, \ldots, 2^{nr}\}$ and $k \in \{1, 2, \ldots, |\mathcal{C}^\circ_j|\}$. This partitioning and labeling is then shared with the message decoder.

Message encoding is accomplished as follows. Given a source codeword $\hat{\mathbf{x}}^\circ(i)$, and a message $\mathsf{m} = m$, we find the index k such that $(\hat{\mathbf{x}}^\circ(i), \mathbf{u}(m,k)) \in T_{\hat{\mathsf{x}}^\circ,\mathsf{u}}$. If there is more than one such index pick any one. Transmit the index l such that $\hat{\mathbf{x}}^\circ(l) = \mathbf{u}(m,k)$. If there is no such index, we declare an embedding failure.

With this encoding and embedding, decoding is straightforward: the source reconstruction is $\hat{\mathbf{x}}$ with $\hat{\mathbf{x}}_i = f(\hat{x}_i^\circ(l))$, and the message estimate is $\hat{\mathsf{m}} = m$ such that $\hat{\mathbf{x}}(m,k) = \hat{\mathbf{x}}^\circ(l)$. It remains only to ensure the error probability vanishes and the distortion constraint is met, which we verify in the sequel.

That the probability of a source encoding failure goes to zero follows from joint strong typicality, since $R_\circ > I(\hat{x}^\circ; x)$. The probability of an embedding failure also goes to zero with large n. To see this, first note that the probability that the original source-quantization vector $\hat{\mathbf{x}}^\circ(i)$ falls into the selected bin m is 2^{-nr}, which goes to zero for n large. Moreover, conditioned on the event that $\hat{\mathbf{x}}^\circ(i)$ is not in bin m, the codewords in bin m look like i.i.d. sequences generated independently of $\hat{\mathbf{x}}^\circ(i)$ according to $\prod_{i=1}^n p_{\mathsf{u}}(x_i)$. Indeed, $\mathcal{C}^\circ_m \subset \mathcal{C}^\circ$, and the entries of $\mathcal{C}^\circ$ are generated independently according to $\prod_{i=1}^n p_{\hat{\mathsf{x}}^\circ}(x_i)$, and $p_{\hat{\mathsf{x}}^\circ}(x) = p_{\mathsf{u}}(x)$. The probability that at least one such sequence $\mathbf{u}(m,k)$ is jointly strongly typical with $\hat{\mathbf{x}}^\circ(i)$ approaches one because there are $2^{n(R_\circ - r)}$ codewords in bin m and, via (73), $R_\circ - r > I(\mathsf{u}; \hat{\mathsf{x}}^\circ)$. The probability that $\hat{\mathsf{m}} \neq m$ is zero because the decoder has direct access to the embedder output.

To see that the distortion constraint d is met, we first note that $(\mathbf{x}, \mathbf{u}) \in T_{\mathsf{x},\mathsf{u}}$ by the Markov lemma [9]. Indeed, $\mathsf{x} \leftrightarrow \hat{\mathsf{x}}^\circ \leftrightarrow \mathsf{u}$, and we have both $(\mathbf{x}, \hat{\mathbf{x}}^\circ) \in T_{\mathsf{x},\hat{\mathsf{x}}^\circ}$ and $(\hat{\mathbf{x}}^\circ, \hat{\mathbf{x}}^\circ(l)) \in T_{\hat{\mathsf{x}}^\circ,\mathsf{u}}$. Hence, by choosing ϵ small enough and n large enough, and by exploiting that d is bounded, we can obtain, for any $\delta > 0$,

$$E\left[d\left(\mathbf{x}, \hat{\mathbf{x}}\right)\right] = \frac{1}{n}\sum_{i=1}^{n} E\left[d\left(x_i, f(\hat{x}_i^\circ(l))\right)\right] = \sum_{x,u} E\left[T_{\mathbf{x},\hat{\mathbf{x}}^\circ(l)}(x,u)\right] d\left(x, f(u)\right) \leq d + \delta, \tag{75}$$

which establishes the theorem. ■

Because of the transcoder output codebook constraint, we have, in general,

$$R_{\mathrm{SD}}(d) + R_{\mathrm{IE}}(d) \leq R_\circ. \tag{76}$$

We now develop cases in which (76) holds with a strict inequality, and when it holds with equality.

8.1 Binary-Hamming Case

As we now show by example, constraining the transcoder output codebook to coincide with the input codebook according to Theorem 3 generally incurs a loss in performance.

To see this, consider again the case of a Bernoulli-p source. In this case, because of the codebook constraint $p_{\hat{\mathsf{x}}^\circ}(x) = p_{\mathsf{u}}(x)$, the information embedding test channel is no longer a binary symmetric channel, but rather

$$p_{\hat{\mathsf{x}}^\circ|\mathsf{u}}(0|1) = \alpha \tag{77}$$

$$p_{\hat{\mathsf{x}}^\circ|\mathsf{u}}(1|0) = \beta. \tag{78}$$

Moreover, for this binary case we can without loss of generality skip the distortion compensation in Theorem 3 (i.e., let $f(\cdot)$ be the identity function so that $\hat{\mathsf{x}} = \mathsf{u}$). Thus, in a manner analogous to the way we obtained the successive degradation rate distortion function for this case, we obtain, via (73), that $r = R_\circ - H_{\mathrm{B}}(q) + \Pr[\mathsf{u} = 1] H_{\mathrm{B}}(\alpha) + \Pr[\mathsf{u} = 0] H_{\mathrm{B}}(\beta)$ and $d = d_\circ + [\Pr[\mathsf{u} = 1]\alpha + \Pr[\mathsf{u} = 0]\beta](1 - 2d_\circ)$. Defining

$$q = p_{\hat{\mathsf{x}}^\circ}(1) = \frac{p - d_\circ}{1 - 2d_\circ}, \tag{79}$$

we incorporate the output codebook constraint $p_{\mathsf{u}}(1) = p_{\hat{\mathsf{x}}^\circ}(1) = q$ to get

$$\beta = \alpha \cdot \frac{q}{1-q}, \tag{80}$$

from which we obtain

$$R_{\mathrm{IE}}(d) = \max\left\{R_\circ - H_{\mathrm{B}}(q) + qH_{\mathrm{B}}\left(\frac{d - d_\circ}{2q(1-2d_\circ)}\right) + (1-q)H_{\mathrm{B}}\left(\frac{d - d_\circ}{2(1-q)(1-2d_\circ)}\right), 0\right\} \tag{81}$$

Comparing $R_{\mathrm{SD}}(d)$ in (27) with $R_\circ - R_{\mathrm{IE}}(d)$ in (81) we see that this embedding strategy is in general less efficient, it takes a higher residual rate to describe the source to a target distortion level: the transcoder output codebook constraint exacts a price in performance. The gap for the case $p = 2/5$ and $R_\circ = 1/4$ (for which $d_\circ \simeq 0.2$) is depicted in Fig. 7 over the relevant distortion range: $d_\circ \leq d \leq p$. Note the step discontinuity at $d = p$: when stealing all the rate, so that $R_\circ - R_{\mathrm{IE}}(p) = 0$, it suffices for the decoder to ignore all the received data and reconstruct the all zero sequence.

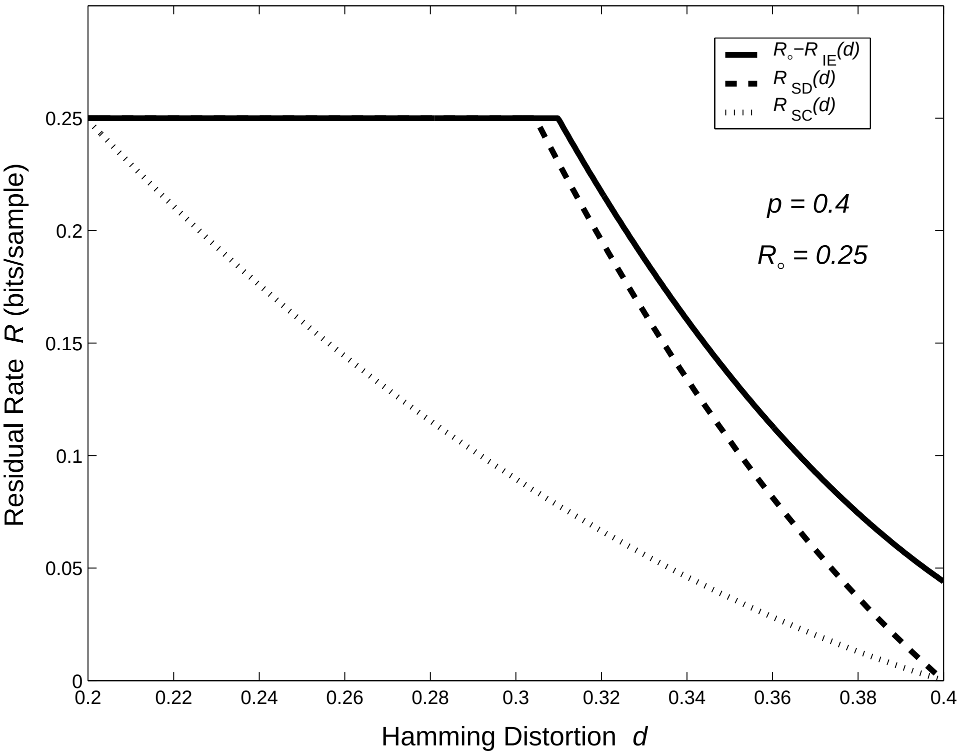


Figure 7: Rate distortion loss for bit stealing via information embedding in the case of a Bernoulli-2/5 source and Hamming distortion measure, with $R_\circ = 1/4$. The progressively lower solid, dashed, and dotted curves correspond to $R_\circ - R_{\mathrm{IE}}(d)$, $R_{\mathrm{SD}}(d)$, and $R_{\mathrm{SC}}(d)$, respectively, i.e, using bit-stealing by embedding with an informed decoder, using bit-stealing by successive degradation, which doesn't have the transcoder output codebook constraint, and using successive refinement (informed encoding).

Note that in the special case $p = 1/2$ (whence $q = 1/2$), no loss of performance is realized, i.e., the transcoder codebook constraint is not limiting: $R_{\mathrm{SD}}(d) + R_{\mathrm{IE}}(d) = R_\circ$. We now consider another important situation in which this property holds.

8.2 Gaussian-Quadratic Case

In this section, we show not only that constraining the transcoder codebook to be the same as its input need not incur a loss in performance vis-*à*-vis rate-splitting in the Gaussian-quadratic case, but also that the only additional decoder processing required to ensure there is no loss is simple distortion compensation in the form of (embedding-rate-dependent) attenuation of the source reconstruction. We further show that even an uninformed decoder suffers no more than a 3 dB distortion penalty — or equivalently a 0.21-bit rate penalty — relative to the informed decoder with distortion compensation.

As in Section 6, we let x be a length-n i.i.d. sequence of Gaussian random variables with variance σ_{x}^2, and consider the quadratic distortion measure $d(a, b) = (a - b)^2$. In this case, we obtain that the distortion rate function (74) for informed decoders takes the form

$$\frac{d_{\mathrm{IE}}(r)}{\sigma_{\mathsf{x}}^2} = 2^{-2R_\circ} + \left(1 - 2^{-2R_\circ}\right) 2^{-2(R_\circ - r)}, \tag{82}$$

and is achieved using distortion compensation of the form

$$\hat{\mathsf{x}} = f(\mathsf{u}) = \beta \mathsf{u} \tag{83}$$

where

$$\beta = \sqrt{1 - 2^{-2(R_\circ - r)}}. \tag{84}$$

In contrast, the corresponding distortion rate function for uninformed decoders takes the form

$$\frac{d_{\mathrm{IE}}^{\mathrm{u}}(r)}{\sigma_{\mathsf{x}}^2} = 1 + \left(1 - 2^{-2R_\circ}\right)\left(1 - 2\sqrt{1 - 2^{-2(R_\circ - r)}}\right). \tag{85}$$

Eqs. (82) and (85) are obtained from (73) and (74) as follows. First, as in Section 6.1 we generate the usual conditional distribution for achieving the rate distortion bound for Gaussian sources according to $\hat{\mathsf{x}}^\circ = \alpha \mathsf{x} + \mathsf{e}_\circ$, where α is given by (31), i.e., $\alpha = (1 - d_\circ/\sigma_{\mathsf{x}}^2)$, and where $\mathsf{e}_\circ$ is a zero-mean Gaussian random variable with variance $\alpha d_\circ$ that is independent of x. Using [19, Lemma A.3], we know that the optimizing distribution in (73) is Gaussian. When we further constrain the distribution so that $p_{\mathsf{u}}(x) = p_{\hat{\mathsf{x}}^\circ}(x)$, we obtain that it must be of the form (cf. [25]) $\mathsf{u} = \gamma \hat{\mathsf{x}}^\circ + \mathsf{e}$, where $\gamma \geq 0$ is a parameter and e is a zero-mean Gaussian random variable with variance $(1 - \gamma^2)(\sigma_{\mathsf{x}}^2 - d_\circ)$ that is independent of $\hat{\mathsf{x}}^\circ$.

Next, it is straightforward to confirm that the optimum distortion compensation $f(\cdot)$ must be the MMSE estimator for x from u. In turn, since we have concluded these are jointly Gaussian random variables, this estimator is linear, whence (83).

It remains only to optimize over the remaining parameters γ and β. In terms of our parameterized distribution, we have

$$R_\circ - r = I(\hat{x}^\circ; u) = \frac{1}{2}\log\left(\frac{1}{1-\gamma^2}\right) \tag{86}$$

whence

$$\gamma = \sqrt{1-2^{-2(R_\circ - r)}}. \tag{87}$$

Thus, the distortion takes the form

$$\begin{aligned} d = E\left[(\hat{x}-x)^2\right] &= (\alpha\beta\gamma-1)^2\sigma_x^2 + \alpha\beta^2\gamma^2 d_\circ + \beta^2(1-\gamma^2)(\sigma_x^2 - d_\circ) \\ &= \sigma_x^2\left[\beta^2\left(1-\frac{d_\circ}{\sigma_x^2}\right) - 2\beta\left(1-\frac{d_\circ}{\sigma_x^2}\right)\sqrt{1-\frac{d_\circ}{\sigma_x^2}2^{2r}} + 1\right]. \end{aligned} \tag{88}$$

where via (30) we have $d_\circ = \sigma_x^2 2^{-2R_\circ}$, and where to obtain (88) we have substituted for α and γ according to (31) and (87), respectively.

For uninformed decoders, it suffices to substitute $\beta = 1$ and into (88) to obtain (85). For informed decoders, simple optimization of the quadratic (88) with respect to β yields (82) with β given by (84).

The corresponding rate distortion functions are readily obtained from (82) and (85), and take the forms, respectively,

$$R_{\mathrm{IE}}(d) = \max\left\{R_\circ - \frac{1}{2}\log\left(\frac{\sigma_x^2 - d_\circ}{d - d_\circ}\right), 0\right\} \tag{89}$$

and

$$R_{\mathrm{IE}}^{\mathrm{u}}(d) = \max\left\{R_\circ + \frac{1}{2}\log\left[1 - \frac{1}{4}\left(1 + \frac{\sigma_x^2 - d}{\sigma_x^2 - d_\circ}\right)^2\right], 0\right\}. \tag{90}$$

Comparing (29) and (89) we see that (76) holds with equality in the Gaussian case: the transcoder codebook constraint does not exact a price in performance provided the decoder is informed.

The embedding rate distortion functions (89) and (90) take the form depicted in Fig. 8. As with bit-stealing by successive degradation, there is a distortion threshold below which no embedding can be performed and still meet the distortion constraint. One can view this threshold as the minimum amount of distortion that must be incurred if any embedding is used.[7] For informed decoders, this is given by (34) since successive degradation and embedding have identical performance characteristics in this case. For uninformed decoders, the threshold is

$$\frac{d_*^{\mathrm{u}}}{\sigma_x^2} = \frac{d_{\mathrm{IE}}^{\mathrm{u}}(0)}{\sigma_x^2} = 1 + \left(1 - 2^{-2R_\circ}\right)\left(1 - 2\sqrt{1-2^{-2R_\circ}}\right). \tag{91}$$

We quantify the loss in performance suffered by an uninformed decoder relative to an informed one in terms of both distortion and rate. We look first at the large r regime, where comparing (82) and (85), and

[7] This threshold is strictly positive because to embed the transcoder must replace codewords it receives to other codewords in the codebook, and these codewords have some average minimum distance — in fact $2d_\circ$ in the high rate limit — from one another.

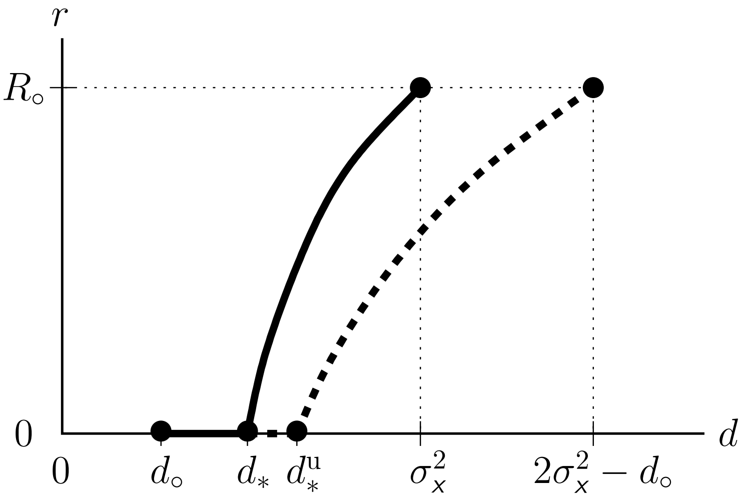


Figure 8: Comparing the rate distortion function for bit stealing via information embedding. The informed and uninformed decoder performances are depicted by the solid and dashed curves, respectively. Below the distortion thresholds d_* and d_*^{u} for the informed and uninformed decoders, respectively, no embedding should be used.

consistent with Fig. 8 we see that the loss is largest. Accordingly, we obtain

$$\frac{d_{\mathrm{IE}}^{\mathrm{u}}(r)}{d_{\mathrm{IE}}(r)} \leq \frac{d_{\mathrm{IE}}^{\mathrm{u}}(R_\circ)}{d_{\mathrm{IE}}(R_\circ)} = 2 - 2^{-2R_\circ} < 2 \tag{92}$$

which corresponds to a 3 dB gap in the large $R_\circ$ limit.[8] The corresponding rate loss comes from comparing (89) and (90), where we see the loss is again largest when r is largest. Thus,

$$R_{\mathrm{IE}}(d) - R_{\mathrm{IE}}^{\mathrm{u}}(d) \leq R_{\mathrm{IE}}(\sigma_x^2) - R_{\mathrm{IE}}^{\mathrm{u}}(\sigma_x^2) = \frac{1}{2}\log\frac{4}{3} \approx 0.21 \text{ bit}, \tag{93}$$

which we note is independent of $R_\circ$.

Turning next to the performance losses in the small r regime, it is straightforward to verify the the distortion gap $d_{\mathrm{IE}}^{\mathrm{u}}(r)/d_{\mathrm{IE}}(r)$ is small as $r \to 0$ — indeed, it is at most 0.2834 dB, which occurs for $R_\circ \approx 0.2528$. At the two extremes in this small stolen rate regime — $R_\circ = 0$ and $R_\circ \to \infty$ — the distortion gap is zero.

To compute the associated rate gap, $R_{\mathrm{IE}}(d) - R_{\mathrm{IE}}^{\mathrm{u}}(d)$, we begin by noting that there exists a rate threshold for $R_\circ$ below which $d_*^{\mathrm{u}} > \sigma_x^2$. In this region $R_{\mathrm{IE}}^{\mathrm{u}}(d) = 0$. In particular, it is straightforward to verify from (91) that $d_*^{\mathrm{u}} > \sigma_x^2$ whenever $0 < R_\circ < R_\circ^*$, where [cf. (93)]

$$R_\circ^* = \frac{1}{2}\log\frac{4}{3} \approx 0.21. \tag{94}$$

Thus the gap in the small stolen rate regime below this threshold is $R_{\mathrm{IE}}(d) - R_{\mathrm{IE}}^{\mathrm{u}}(d) = R_{\mathrm{IE}}(d) \leq R_{\mathrm{IE}}(\sigma_x^2) =$

[8] This gap arises because with $r = R_\circ$, the source is completely overwritten by the embedded message. An informed decoder will ignore the received source codeword, reproducing $\hat{\mathbf{x}} = E[\mathbf{x}]$ and experiencing distortion σ_x^2. However, an uninformed decoder does not know to ignore what it receives, so it experiences an additional distortion of $\sigma_x^2 - d_\circ$, the variance of the received codeword.

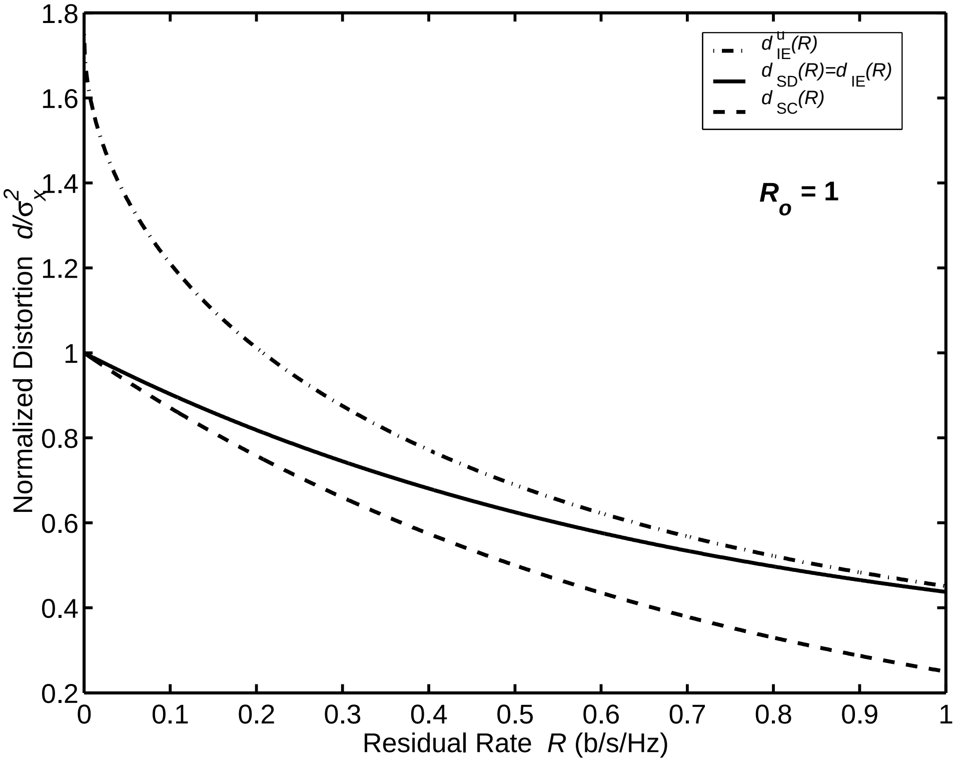


Figure 9: Distortion rate trade-offs for bit stealing in the Gaussian-quadratic case, with $R_\circ = 1$. The progressively lower dash-dotted, solid, and dashed curves correspond to $d^{\mathrm{u}}_{\mathrm{IE}}(R_\circ - R)$, $d_{\mathrm{SD}}(R) = d_{\mathrm{IE}}(R_\circ - R)$, and $d_{\mathrm{SC}}(R)$, respectively, i.e, using information without an informed decoder, using information embedding with an informed decoder or using successive degradation, and using successive refinement (informed encoding).

$R_\circ$, which decreases monotonically to zero as $R_\circ$ decreases from $R^*_\circ$ to 0. Above $R_\circ = R^*_\circ$, the rate gap is largest and equals

$$R_{\mathrm{IE}}(d^{\mathrm{u}}_*) = \frac{1}{2}\log\left[2^{1+2R_\circ}\left(1-\sqrt{1-2^{-2R_\circ}}\right)\right] \tag{95}$$

which decreases monotonically to zero with increasing $R_\circ$. Hence, regardless of $R_\circ$ the rate gap is, again, at most $R^*_\circ$ or 0.21 bits.

The different rate distortion trade-offs for bit stealing in the Gaussian-quadratic case are summarized in Fig. 9, where normalized distortion is plotted versus residual source coding rate $R = R_\circ - r$, for $R_\circ = 1$. The common distortion for successive degradation and embedding with an informed decoder, i.e., $d_{\mathrm{IE}}(R_\circ - R) = d_{\mathrm{SD}}(R)$, appears as the solid middle curve. The distortion for embedding with an uninformed decoder, i.e., $d^{\mathrm{u}}_{\mathrm{IE}}(R_\circ - R)$, appears as the dash-dotted upper curve. Finally, the distortion for successive refinement, which corresponds to an informed encoder, is the regular Gaussian-quadratic distortion rate function $d_{\mathrm{SC}}(R)$ and appears as the dashed lower curve.

9 Conclusions

In this paper, we show a variety of results on transcoding. In particular, we show that transcoding with uninformed encoders and/or decoders need not incur significant losses relative to their informed counterpart for some meaningful models, at least if not applied repeatedly. In some sense, this means that almost all good source codes are effectively almost successively refinable even when not designed as such. Thus it may make sense for system designers to avoid imposing the successive refinement constraint in code design, and use the associated degrees of freedom in other ways. Of course, while we have argued that the transcoder performance need not significantly suffer, the complexity of transcoding may be increase. More generally, this raises interesting questions for future work regarding which transcoding complexity/performance trade-offs are most attractive in different scenarios.

We further show that information embedding approaches to bit stealing that allow for both informed and uninformed source decoders can also be quite efficient. In particular, for reasonable models, they do as well as bit stealing approaches that are not so constrained. When a large fraction of the bits are being stolen, uninformed decoders can incur a substantial performance loss relative to informed decoders. However, informed decoders differ only by incorporating distortion compensation in the form of simple post-reconstruction scaling. Thus, for many audiovisual-oriented sources of practical interest, and the associated gain-invariant distortion measures arising out of human perceptual characteristics, the uninformed and informed source decoder outputs are equivalently good. In this case, the price for enabling uninformed source decoders is increased complexity for extracting the stolen bits, at least relative to successive degradation.

A Proof of Lemma 1

Before proceeding, we need the following Lemma.

Lemma 2 *For any event* $\mathcal{B}$ *with* $\Pr[\mathcal{B}] > 1/2$,

$$|\Pr[\mathcal{A} \mid \mathcal{B}] - \Pr[\mathcal{A}]| < 2(1 - \Pr[\mathcal{B}]) \tag{96}$$

Proof:
We begin by upper bounding $\Pr[\mathcal{A} \mid \mathcal{B}]$ via

$$\Pr[\mathcal{A} \mid \mathcal{B}] = \frac{\Pr[\mathcal{A} \cap \mathcal{B}]}{\Pr[\mathcal{B}]} \le \frac{\Pr[\mathcal{A}]}{\Pr[\mathcal{B}]} < \Pr[\mathcal{A}] \cdot (3 - 2\Pr[\mathcal{B}]) \le \Pr[\mathcal{A}] + 2\,(1 - \Pr[\mathcal{B}])\,, \tag{97}$$

where (97) follows since $1/\Pr[\mathcal{B}] < 3 - 2\Pr[\mathcal{B}]$ if $\Pr[\mathcal{B}] > 1/2$.

Next, we lower bound $\Pr[\mathcal{A} \mid \mathcal{B}]$ via

$$\Pr[\mathcal{A} \mid \mathcal{B}] = \frac{\Pr[\mathcal{A}] + \Pr[\mathcal{B}] - \Pr[\mathcal{A} \cup \mathcal{B}]}{\Pr[\mathcal{B}]} \ge \Pr[\mathcal{A}] + \Pr[\mathcal{B}] - \Pr[\mathcal{A} \cup \mathcal{B}] > \Pr[\mathcal{A}] - (1 - \Pr[\mathcal{B}]) \ge \Pr[\mathcal{A}] - 2\,(1 - \Pr[\mathcal{B}])\,. \tag{98}$$

Combining these upper and lower bounds establishes (96) ■

Returning to our proof of Lemma 1, let

$$\mathcal{G}_\delta \triangleq \{(\mathbf{a}, \mathbf{b}) : D\left(T_{\mathbf{a},\mathbf{b}} \| p_{\mathbf{v},u}\right) \le \delta \text{ and } T_{\mathbf{a},\mathbf{b}} = T_{\mathbf{v}(j),\mathbf{u}} \text{ for some } j\}, \tag{99}$$

where $\mathbf{v}(j)$ is the jth codeword in the quantization codebook, for $j = 1, 2, \ldots, 2^{nR^\delta}$. Then

$$\begin{aligned}
\big|E\{\theta[\mathbf{v}, \mathbf{u}, g(\mathbf{v})]\} - &E\{\theta[\mathbf{w}, \mathbf{u}, g(\mathbf{w})]\}\big| \\
&= |\Pr\{\theta[\mathbf{v}, \mathbf{u}, g(\mathbf{v})] = 1\} - \Pr\{\theta[\mathbf{w}, \mathbf{u}, g(\mathbf{w})] = 1\}| \qquad (100)\\
&\le |\Pr\{\theta[\mathbf{v}, \mathbf{u}, g(\mathbf{v})] = 1\} - \Pr\{\theta[\mathbf{v}, \mathbf{u}, g(\mathbf{v})] = 1 \mid \mathcal{E}^c\}| \\
&\quad + |\Pr\{\theta[\mathbf{w}, \mathbf{u}, g(\mathbf{w})] = 1 \mid (\mathbf{w}, \mathbf{u}) \in \mathcal{G}_\delta\} - \Pr\{\theta[\mathbf{w}, \mathbf{u}, g(\mathbf{w})] = 1\}| \\
&\quad + |\Pr\{\theta[\mathbf{v}, \mathbf{u}, g(\mathbf{v})] = 1 \mid \mathcal{E}^c\} - \Pr\{\theta[\mathbf{w}, \mathbf{u}, g(\mathbf{w})] = 1 \mid (\mathbf{w}, \mathbf{u}) \in \mathcal{G}_\delta\}| \qquad (101)\\
&\le 2\Pr[\mathcal{E}] + 2\Pr[\mathcal{E}] + 0, \qquad (102)
\end{aligned}$$

where (100) follows from the fact that $\theta \in \{0, 1\}$, where (101) follows from two applications of the triangle inequality, and where the first and second terms in (102) come from applications of Lemma 2. Note that to obtain the second term we have used that, in accordance with the dithered quantization rule, $\Pr\left[(\mathbf{w}, \mathbf{u}) \in \mathcal{G}_\delta\right] = 1 - \Pr[\mathcal{E}]$.

To see that the third term in (102) is zero, note that

$$\begin{aligned}
\Pr\{\theta[\mathbf{v}, \mathbf{u}, g(\mathbf{v})] = 1 \mid \mathcal{E}^c\} - &\Pr\{\theta[\mathbf{w}, \mathbf{u}, g(\mathbf{w})] = 1 \mid (\mathbf{w}, \mathbf{u}) \in \mathcal{G}_\delta\} \\
&= \sum_{\mathbf{a},\mathbf{b},\mathbf{c}} \theta[\mathbf{a}, \mathbf{b}, \mathbf{c}] \; \left[p_{\mathbf{v},\mathbf{u},g(\mathbf{v})|\mathcal{E}^c}(\mathbf{a}, \mathbf{b}, \mathbf{c}) - p_{\mathbf{w},\mathbf{u},g(\mathbf{w})|(\mathbf{w},\mathbf{u})\in\mathcal{G}_\delta}(\mathbf{a}, \mathbf{b}, \mathbf{c})\right] \\
&= \sum_{\mathbf{a},\mathbf{b},\mathbf{c}} \theta[\mathbf{a}, \mathbf{b}, \mathbf{c}]\, p_{g(\mathbf{v})|\mathbf{v}}(\mathbf{c}|\mathbf{a}) \; \left[p_{\mathbf{v},\mathbf{u}|\mathcal{E}^c}(\mathbf{a}, \mathbf{b}) - p_{\mathbf{w},\mathbf{u}|(\mathbf{w},\mathbf{u})\in\mathcal{G}_\delta}(\mathbf{a}, \mathbf{b})\right], \qquad (103)
\end{aligned}$$

where to obtain (103) we have used the map independence property, which yields the Markov relationships $\mathbf{u} \leftrightarrow \mathbf{v} \leftrightarrow g(\mathbf{v})$ and $\mathbf{u} \leftrightarrow \mathbf{w} \leftrightarrow g(\mathbf{w})$. Finally, that the term in brackets in (103) is zero follows immediately from the way that $\mathbf{v}$ is generated from a noisy observation $\mathbf{w}$ in the dithered quantization rule when the quantization is successful.

B Proof of Proposition 1

If $\mathbf{y}$ is the output of the dithering (13) when $\mathbf{x}$ is the input, and $\hat{\mathbf{x}}^\circ$ is the codeword to which $\mathbf{y}$ is mapped, then when encoding succeeds we have

$$\begin{aligned}
d\left(\mathbf{x}, \hat{\mathbf{x}}^\circ\right) &= d\left(\mathbf{x}, \mathbf{y}\right) && (104)\\
&= \sum_{\hat{x}^\circ, x} d\left(x, \hat{x}^\circ\right) T_{\mathbf{y},\mathbf{x}}(\hat{x}^\circ, x) && (105)\\
&= E\left[d\left(x, \hat{x}^\circ\right)\right] + \sum_{\hat{x}^\circ, x} d\left(x, \hat{x}^\circ\right)\left[T_{\mathbf{y},\mathbf{x}}(\hat{x}^\circ, x) - p_{\hat{\mathsf{x}}^\circ,\mathsf{x}}(\hat{x}^\circ, x)\right] && (106)\\
&\leq E\left[d\left(x, \hat{x}^\circ\right)\right] + d_{\max} \sum_{\hat{x}^\circ, x} \left|T_{\mathbf{y},\mathbf{x}}(\hat{x}^\circ, x) - p_{\hat{\mathsf{x}}^\circ,\mathsf{x}}(\hat{x}^\circ, x)\right| && (107)\\
&\leq E\left[d\left(x, \hat{x}^\circ\right)\right] + d_{\max} \sqrt{2 \ln 2 \cdot D\left(T_{\mathbf{y},\mathbf{x}} \| p_{\hat{\mathsf{x}}^\circ,\mathsf{x}}\right)} && (108)\\
&\leq E\left[d\left(x, \hat{x}^\circ\right)\right] + d_{\max} \sqrt{2\delta \ln 2} && (109)
\end{aligned}$$

where (107) follows from the triangle inequality and since $d\left(x, \hat{x}^\circ\right) \leq d_{\max}$, (108) follows, like (10), from [9, Lemma 12.6.1, p. 300], and (109) is a consequence of dithered encoding step 2. The remaining steps follow from simple algebraic manipulations and the definition of types.

Finally, we obtain

$$\begin{aligned}
\Pr\left[d\left(\mathbf{x}, \hat{\mathbf{x}}^\circ\right) > E\left[d\left(x, \hat{x}^\circ\right)\right] + d_{\max}\sqrt{2\delta \ln 2}\right] &= \Pr\left[d\left(\mathbf{x}, \hat{\mathbf{x}}^\circ\right) > E\left[d\left(x, \hat{x}^\circ\right)\right] + d_{\max}\sqrt{2\delta \ln 2} \mid \mathcal{E}\right] \Pr[\mathcal{E}] && (110)\\
&\quad + \Pr\left[d\left(\mathbf{x}, \hat{\mathbf{x}}^\circ\right) > E\left[d\left(x, \hat{x}^\circ\right)\right] + d_{\max}\sqrt{2\delta \ln 2} \mid \mathcal{E}^c\right] \Pr[\mathcal{E}^c] && (111)\\
&\leq 1 \cdot \Pr[\mathcal{E}] + 0 \cdot \Pr[\mathcal{E}^c] && (112)
\end{aligned}$$

where to obtain the zero in the last line we have used (109), yielding (15) as desired.

C Proof of Proposition 2

First, we need the following lemma, which establishes that for any pair of sequences $(\mathbf{y}, \mathbf{x})$ satisfying encoding step 2, the empirical type $T_{\mathbf{y},\mathbf{x}}$ is about as close to $p_{\hat{\mathsf{x}}^\circ} p_{\mathsf{x}}$ as $p_{\hat{\mathsf{x}}^\circ,\mathsf{x}}$ is to $p_{\hat{\mathsf{x}}^\circ} p_{\mathsf{x}}$.

Lemma 3 *For any empirical joint type $T_{\mathbf{y},\mathbf{x}}$ where $(\mathbf{y}, \mathbf{x})$ satisfies the condition in dithered encoding step 2,*

$$D\left(T_{\mathbf{y},\mathbf{x}} \| p_{\hat{\mathsf{x}}^\circ} p_{\mathsf{x}}\right) \leq I(\hat{x}^\circ; x) + \delta - \log\left[\left(p^{\min}_{\hat{\mathsf{x}}^\circ,\mathsf{x}}\right)^3\right] \sqrt{2\delta \ln 2}, \tag{113}$$

where the superscript $^{\min}$ *denotes the smallest nonzero value of the distribution that is its argument, i.e., for an arbitrary distribution p_{w},*

$$p^{\min}_{\mathsf{w}} = \min_{\{w | p_{\mathsf{w}}(w) > 0\}} p_{\mathsf{w}}(w) \tag{114}$$

Proof:

Eq. (113) is obtained via

$$\begin{aligned}
D\left(T_{\mathbf{y},\mathbf{x}} \| p_{\hat{\mathsf{x}}^\circ} p_{\mathsf{x}}\right) &= I(\hat{x}^\circ; x) + D\left(T_{\mathbf{y},\mathbf{x}} \| p_{\hat{\mathsf{x}}^\circ} p_{\mathsf{x}}\right) - D\left(p_{\hat{\mathsf{x}}^\circ,\mathsf{x}} \| p_{\hat{\mathsf{x}}^\circ} p_{\mathsf{x}}\right) && (115)\\
&= I(\hat{x}^\circ; x) + H(p_{\hat{\mathsf{x}}^\circ,\mathsf{x}}) - H(T_{\mathbf{y},\mathbf{x}}) - \sum_{\hat{x}^\circ, x}\left[T_{\mathbf{y},\mathbf{x}}(\hat{x}^\circ, x) - p_{\hat{\mathsf{x}}^\circ,\mathsf{x}}(\hat{x}^\circ, x)\right] \log\left[p_{\hat{\mathsf{x}}^\circ}(\hat{x}^\circ) p_{\mathsf{x}}(x)\right] && (116)
\end{aligned}$$

$$
\begin{aligned}
&\le I(\hat{x}^\circ;\mathsf{x}) + |H(T_{\mathbf{y},\mathbf{x}}) - H(p_{\hat{\mathsf{x}}^\circ,\mathsf{x}})| - \sum_{\hat{x}^\circ,x} |T_{\mathbf{y},\mathbf{x}}(\hat{x}^\circ,x) - p_{\hat{\mathsf{x}}^\circ,\mathsf{x}}(\hat{x}^\circ,x)| \log\left[p_{\hat{\mathsf{x}}^\circ}(\hat{x}^\circ)p_{\mathsf{x}}(x)\right] && (117)\\
&\le I(\hat{x}^\circ;\mathsf{x}) + D\left(T_{\mathbf{y},\mathbf{x}}\|p_{\hat{\mathsf{x}}^\circ,\mathsf{x}}\right) - \sum_{\hat{x}^\circ,x} |T_{\mathbf{y},\mathbf{x}}(\hat{x}^\circ,x) - p_{\hat{\mathsf{x}}^\circ,\mathsf{x}}(\hat{x}^\circ,x)| \log\left[p_{\hat{\mathsf{x}}^\circ}(\hat{x}^\circ)p_{\mathsf{x}}(x)p_{\hat{\mathsf{x}},\hat{\mathsf{x}}^\circ}(\hat{x},\hat{x}^\circ)\right] && (118)\\
&\le I(\hat{x}^\circ;\mathsf{x}) + \delta - \log\left[(p_{\hat{\mathsf{x}}^\circ}p_{\mathsf{x}})^{\min} p^{\min}_{\hat{\mathsf{x}}^\circ,\mathsf{x}}\right] \sum_{\hat{x}^\circ,x} |T_{\mathbf{y},\mathbf{x}}(\hat{x}^\circ,x) - p_{\hat{\mathsf{x}}^\circ,\mathsf{x}}(\hat{x}^\circ,x)| && (119)\\
&\le I(\hat{x}^\circ;\mathsf{x}) + \delta - \log\left[\left(p^{\min}_{\hat{\mathsf{x}}^\circ,\mathsf{x}}\right)^3\right] \sqrt{D\left(T_{\mathbf{y},\mathbf{x}}\|p_{\hat{\mathsf{x}}^\circ,\mathsf{x}}\right) 2\ln 2} && (120)\\
&\le I(\hat{x}^\circ;\mathsf{x}) + \delta - \log\left[(p^{\min}_{\hat{\mathsf{x}}^\circ,\mathsf{x}})^3\right] \sqrt{2\delta \ln 2}, && (121)
\end{aligned}
$$

where (117) follows from fact that $\log p_{\hat{\mathsf{x}}^\circ}p_{\mathsf{x}} \le 0$, (118) follows from the inequality

$$
\begin{aligned}
|H(T_{\mathbf{y},\mathbf{x}}) - H(p_{\hat{\mathsf{x}}^\circ,\mathsf{x}})| &= \left| D\left(T_{\mathbf{y},\mathbf{x}}\|p_{\hat{\mathsf{x}}^\circ,\mathsf{x}}\right) + \sum_{\hat{x}^\circ,x} \left[T_{\mathbf{y},\mathbf{x}}(\hat{x}^\circ,x) - p_{\hat{\mathsf{x}}^\circ,\mathsf{x}}(\hat{x}^\circ,x)\right] \log\left[p_{\hat{\mathsf{x}},\hat{\mathsf{x}}^\circ}(\hat{x},\hat{x}^\circ)\right] \right| && (122)\\
&\le D\left(T_{\mathbf{y},\mathbf{x}}\|p_{\hat{\mathsf{x}}^\circ,\mathsf{x}}\right) - \sum_{\hat{x}^\circ,x} |T_{\mathbf{y},\mathbf{x}}(\hat{x}^\circ,x) - p_{\hat{\mathsf{x}}^\circ,\mathsf{x}}(\hat{x}^\circ,x)| \log\left[p_{\hat{\mathsf{x}},\hat{\mathsf{x}}^\circ}(\hat{x},\hat{x}^\circ)\right], && (123)
\end{aligned}
$$

and (119) follows from successful quantization in encoding step 2 and the following argument. If for some $(\hat{x}^\circ, x)$ the marginal product satisfies $p_{\hat{\mathsf{x}}^\circ}(\hat{x}^\circ)p_{\mathsf{x}}(x) = 0$ then $p_{\hat{\mathsf{x}}^\circ,\mathsf{x}}(\hat{x}^\circ,x) = T_{\mathbf{y},\mathbf{x}}(\hat{x}^\circ,x) = 0$. This is because $p_{\hat{\mathsf{x}}^\circ,\mathsf{x}}$ must have at least the same zeros as $p_{\hat{\mathsf{x}}^\circ}p_{\mathsf{x}}$. In this case $T_{\mathbf{y},\mathbf{x}}(\hat{x}^\circ,x) = 0$ since $D\left(T_{\mathbf{y},\mathbf{x}}\|p_{\hat{\mathsf{x}}^\circ,\mathsf{x}}\right)$ cannot be infinite because we know the dithered encoder succeeded in Step 2. Hence, the largest term in the sum in (118) is $\log\left[(p_{\hat{\mathsf{x}}^\circ}p_{\mathsf{x}})^{\min}p^{\min}_{\hat{\mathsf{x}}^\circ,\mathsf{x}}\right]$, and (119) follows. The remaining two inequalities (120) and (121) follow from analogous arguments in the proof of Proposition 1, together with the fact that $(p_{\hat{\mathsf{x}}^\circ}p_{\mathsf{x}})^{\min} > \left(p^{\min}_{\hat{\mathsf{x}}^\circ,\mathsf{x}}\right)^2$.
■

Next we need a Lemma lower bounding the probability that the encoder could encode to the ith codeword.

Lemma 4 *For any empirical joint type $T_{\mathbf{y},\mathbf{x}}$ (and in particular the ones where $(\mathbf{y},\mathbf{x})$ satisfies the condition in the dithered encoding step 2),*

$$
\Pr\left[T_{\hat{\mathbf{x}}^\circ(i),\mathbf{x}} = T_{\mathbf{y},\mathbf{x}}\right] \ge 2^{-n\left[D(T_{\mathbf{y},\mathbf{x}}\|p_{\hat{\mathsf{x}}^\circ}p_{\mathsf{x}}) + |\mathcal{X}|^2\cdot\frac{\log(n+1)}{n}\right]}. \qquad (124)
$$

Proof:
The desired result follows from the chain of inequalities

$$
\begin{aligned}
\Pr\left[T_{\hat{\mathbf{x}}^\circ(i),\mathbf{x}} = T_{\mathbf{y},\mathbf{x}}\right] &= \Pr[T_{\hat{\mathbf{x}}^\circ(i),\mathbf{x}} = T_{\mathbf{y},\mathbf{x}} \mid \boldsymbol{\mathsf{x}} = \mathbf{x}] && (125)\\
&= \frac{\Pr\left[T_{\hat{\mathbf{x}}^\circ(i),\mathbf{x}} = T_{\mathbf{y},\mathbf{x}}\right]\cdot \Pr[\boldsymbol{\mathsf{x}} = \mathbf{x} \mid T_{\hat{\mathbf{x}}^\circ(i),\mathbf{x}} = T_{\mathbf{y},\mathbf{x}}]}{\Pr[\boldsymbol{\mathsf{x}} = \mathbf{x}]} && (126)\\
&= \frac{\Pr\left[T_{\hat{\mathbf{x}}^\circ(i),\mathbf{x}} = T_{\mathbf{y},\mathbf{x}}\right]\cdot 2^{-n\cdot H(T_{\mathbf{x}})}}{2^{-n(H(T_{\mathbf{x}}) + D(T_{\mathbf{x}}\|p_{\mathsf{x}}))}} && (127)\\
&\ge \Pr\left[T_{\hat{\mathbf{x}}^\circ(i),\mathbf{x}} = T_{\mathbf{y},\mathbf{x}}\right] && (128)\\
&\ge (n+1)^{-|\mathcal{X}|^2}\cdot 2^{-nD\left(T_{\mathbf{y},\mathbf{x}}\|p_{\hat{\mathsf{x}}^\circ}p_{\mathsf{x}}\right)}. && (129)
\end{aligned}
$$

Eq. (125) follows from the fact that $\hat{\boldsymbol{\mathsf{x}}}^\circ$ and $\boldsymbol{\mathsf{x}}$ are independent since the codewords are generated independently of the source, and (126) follows from the definition of conditional probability. Eq. (127) follows from [9, Theorem 12.1.2, p. 281] and the observation that $T_{\hat{\mathbf{x}}^\circ(i),\mathbf{x}} = T_{\mathbf{y},\mathbf{x}}$ implies $T_{\boldsymbol{\mathsf{x}}} = T_{\mathbf{x}}$. Eq. (128) follows from the fact that relative entropy is non-negative. Finally, (129) is a consequence of [9, Theorem 12.1.4, p. 285]. ■

Now we are ready to prove Proposition 2. First, we express the dithered quantization failure event $\mathcal{E}$ as

$$
\mathcal{E} = \mathcal{E}_D \cup \mathcal{E}_T, \qquad (130)
$$

where $\mathcal{E}_D$ denotes the event that the relative entropy $D\left(T_{\mathbf{y},\mathbf{x}}\|p_{\hat{\mathsf{x}}^\circ,\mathsf{x}}\right)$ is too large in step 2 and $\mathcal{E}_T$ denotes the event that no sequences exist in the codebook such that $T_{\hat{\mathbf{x}}^\circ,\mathbf{x}} = T_{\mathbf{y},\mathbf{x}}$ in step 3.

From (130), we obtain, via the union bound,

$$\Pr[\mathcal{E}] \leq \Pr[\mathcal{E}_D] + \Pr[\mathcal{E}_T \mid \mathcal{E}_D^{\mathrm{c}}], \tag{131}$$

whose components we now bound in turn.

Via [9, Theorem 12.2.1, p. 287], we obtain

$$\Pr[\mathcal{E}_D] \leq 2^{-n\left(\delta - |\mathcal{X}|^2 \cdot \frac{\log(n+1)}{n}\right)}. \tag{132}$$

Hence there exists an $\alpha_1 > 0$ and n_1 such that for all $n > n_1$,

$$\Pr[\mathcal{E}_D] < \exp(-n\alpha_1). \tag{133}$$

For example, it suffices to choose $\alpha_1 = \delta/2$ and $n_1 = 4\,|\mathcal{X}|^4/\delta^2$.

The second term in (131) is bounded by the following chain of inequalities:

$$\begin{aligned}
\Pr[\mathcal{E}_T \mid \mathcal{E}_D^{\mathrm{c}}] &= \sum_{\mathbf{y},\mathbf{x}} \Pr[\mathcal{E}_T \mid (\mathbf{x},\mathbf{y}) = (\mathbf{x},\mathbf{y}), \mathcal{E}_D^{\mathrm{c}}] \Pr[(\mathbf{x},\mathbf{y}) = (\mathbf{x},\mathbf{y}) \mid \mathcal{E}_D^{\mathrm{c}}] && (134)\\
&\leq \sum_{\mathbf{x},\mathbf{y}} \max_{\mathbf{x},\mathbf{y}} \{\Pr[\mathcal{E}_T \mid (\mathbf{x},\mathbf{y}) = (\mathbf{x},\mathbf{y}), \mathcal{E}_D^{\mathrm{c}}]\} \Pr[(\mathbf{x},\mathbf{y}) = (\mathbf{x},\mathbf{y}) \mid \mathcal{E}_D^{\mathrm{c}}] && (135)\\
&= \max_{\mathbf{x},\mathbf{y}} \Pr[\mathcal{E}_T \mid (\mathbf{x},\mathbf{y}) = (\mathbf{x},\mathbf{y}), \mathcal{E}_D^{\mathrm{c}}] && (136)\\
&= \max_{\mathbf{x},\mathbf{y}} \prod_{i=1}^{2^{nR_\circ}} \left(1 - \Pr[T_{\hat{\mathbf{x}}^\circ(i),\mathbf{x}} = T_{\mathbf{y},\mathbf{x}} \mid \mathcal{E}_D^{\mathrm{c}}]\right) && (137)\\
&= \max_{\mathbf{x},\mathbf{y}} \left(1 - \Pr[T_{\hat{\mathbf{x}}^\circ(1),\mathbf{x}} = T_{\mathbf{y},\mathbf{x}} \mid \mathcal{E}_D^{\mathrm{c}}]\right)^{2^{nR_\circ}} && (138)\\
&\leq \max_{\mathbf{x},\mathbf{y}} \left(1 - 2^{-n\left[D(T_{\mathbf{y},\mathbf{x}} \| p_{\hat{x}^\circ} p_x) + |\mathcal{X}|^2 \cdot \frac{\log(n+1)}{n}\right]}\right)^{2^{nR_\circ}} && (139)\\
&\leq \max_{\mathbf{x},\mathbf{y}} \exp\left\{-2^{-n\left[D(T_{\mathbf{y},\mathbf{x}} \| p_{\hat{x}^\circ} p_x) + |\mathcal{X}|^2 \cdot \frac{\log(n+1)}{n}\right]} \cdot 2^{nR_\circ}\right\} && (140)\\
&= \max_{\mathbf{x},\mathbf{y}} \exp\left\{-2^{n\left[R_\circ - D(T_{\mathbf{y},\mathbf{x}} \| p_{\hat{x}^\circ} p_x) - |\mathcal{X}|^2 \cdot \frac{\log(n+1)}{n}\right]}\right\} && (141)\\
&\leq \exp\left\{-2^{n\left[R_\circ - I(\hat{x}^\circ; x) - \delta + \log\left[\left(p_{\hat{x}^\circ,x}^{\min}\right)^3\right]\sqrt{2\delta \ln 2} - |\mathcal{X}|^2 \cdot \frac{\log(n+1)}{n}\right]}\right\}. && (142)
\end{aligned}$$

where (138) follows from symmetry, (139) follows from Lemma 4, (140) follows from the inequality [9, Lemma 13.5.3, p. 353] $(1-y)^n \leq \exp(-yn)$, and (142) follows from Lemma 3.

Let $\delta_{\max}$ be the largest value of δ such that

$$R_\circ > I(\hat{x}^\circ; x) + \delta - \log\left[\left(p_{\hat{x}^\circ,x}^{\min}\right)^3\right]\sqrt{2\delta \ln 2}. \tag{143}$$

Note that since $R_\circ > I(x; \hat{x}^\circ)$ we have $\delta_{\max} > 0$. Hence, for every $\delta \in (0, \delta_{\max})$, there exists an n_2 such that for all $n > n_2$

$$\Pr[\mathcal{E}_T \mid \mathcal{E}_D^{\mathrm{c}}] \leq \exp(-\alpha_2 n). \tag{144}$$

for some $\alpha_2 > 0$.

Finally, combining the exponential bounds (133) and (144) with (131) we conclude there exists an n_0 such that for all $n > n_0$

$$\Pr[\mathcal{E}] \leq 2\exp[-n\min(\alpha_1, \alpha_2)]. \tag{145}$$

D Proof of Proposition 3

For our proof, we require the following lemma.

Lemma 5 *If $\hat{\mathbf{x}}^\circ$ is the output of an admissible encoder and $p_{\hat{x}^\circ}$ is its distribution, then for all n*

$$\Pr[D\left(T_{\hat{\mathbf{x}}^\circ} \| p_{\hat{x}^\circ}\right) \leq \epsilon_0(n)] \geq 1 - \epsilon_0(n), \tag{146}$$

where $\epsilon_0(n)$ is a function satisfying $\lim_{n\to\infty} \epsilon_0(n) = 0$.

Proof:
For admissible encoders we know $D\left(T_{\hat{\mathbf{x}}^\circ,\mathbf{x}} \| p_{\hat{x}^\circ,x}\right)$ converges to zero in probability and therefore so does $D\left(T_{\hat{\mathbf{x}}^\circ} \| p_{\hat{x}^\circ}\right)$, i.e., for every $\epsilon > 0$ there exists an $n_0(\epsilon)$ such that for all $n > n_0(\epsilon)$

$$\Pr[D\left(T_{\hat{\mathbf{x}}^\circ} \| p_{\hat{x}^\circ}\right) \leq \epsilon] \geq 1 - \epsilon. \tag{147}$$

Without loss of generality, we can assume that the function is a mapping of the form $n_0(\epsilon) : (0,1) \mapsto \mathbb{R}$ and monotonically increases as ϵ decreases. As such, it possesses an inverse that is the desired $\epsilon_0(n)$ in our lemma. ■

For our main result, we use proof by contradiction. Suppose that the probability that the dithered transcoder fails does not converge to 0 with n, but stays above some fixed $\epsilon_1 > 0$. This implies, when combined with Lemma 5, that there is set of sequences $\mathcal{S}_n$ that cannot be transcoded successfully and such that for all $\hat{\mathbf{x}}^\circ \in \mathcal{S}_n$,

$$D\left(T_{\hat{\mathbf{x}}^\circ} \| p_{\hat{x}^\circ}\right) \leq \epsilon_0(n). \tag{148}$$

Since by their construction dithered transcoders treat all inputs in a given type identically, the set $\mathcal{S}_n$ must contain some number of whole type classes. Denote these type classes $\mathcal{T}_1, \mathcal{T}_2, \ldots, \mathcal{T}_K$. Let $\mathcal{T}_*$ denote the worst of these type classes, i.e.,

$$\mathcal{T}_* = \underset{\mathcal{T}\in\{\mathcal{T}_1,\mathcal{T}_2,\ldots,\mathcal{T}_K\}}{\arg\max} \Pr[\mathcal{E} \mid \hat{\mathbf{x}}^\circ \in \mathcal{T}], \tag{149}$$

where $\mathcal{E}$ denotes the event that the dithered transcoding fails.

Then $\mathcal{T}_*$ satisfies the following:

$$\begin{aligned} \epsilon_1 = \Pr[\mathcal{E}] &= \sum_{i=1}^{K} \Pr[\mathcal{E} \mid \hat{\mathbf{x}}^\circ \in \mathcal{T}_i] \Pr[\hat{\mathbf{x}}^\circ \in \mathcal{T}_i] && (150)\\ &\leq \sum_{i=1}^{K} \Pr[\mathcal{E} \mid \hat{\mathbf{x}}^\circ \in \mathcal{T}_i] && (151)\\ &\leq \Pr[\mathcal{E} \mid \hat{\mathbf{x}}^\circ \in \mathcal{T}_*] \cdot (n+1)^{|\mathcal{X}|}. && (152) \end{aligned}$$

where to obtain the last inequality we have used (149) and that there are at most $(n+1)^{|\mathcal{X}|}$ type classes [9, Theorem 12.1.1, p. 280]. From (152), we obtain

$$\Pr[\mathcal{E} \mid \hat{\mathbf{x}}^\circ \in \mathcal{T}_*] \geq \frac{\epsilon_1}{(n+1)^{|\mathcal{X}|}}. \tag{153}$$

Now let us further suppose that the transcoder input is generated in an i.i.d. manner according to the distribution $p_{\hat{x}^\circ}$, which is a valid output of an admissible source encoder. Then [9, Theorem 12.1.4, p. 285]

$$\Pr[\hat{\mathbf{x}}^\circ \in \mathcal{T}_*] \geq \frac{2^{-nD(\mathcal{T}_* \| p_{\hat{x}^\circ})}}{(n+1)^{|\mathcal{X}|}} \geq \frac{2^{-n\epsilon_0(n)}}{(n+1)^{|\mathcal{X}|}}. \tag{154}$$

where to obtain the second inequality we have used (148).

From (153) and (154), we see that the probability of dithered transcoder failure does not decay exponentially in n, since

$$\Pr[\mathcal{E}] \geq \Pr[\mathcal{E} \mid \hat{\mathbf{x}}^\circ \in \mathcal{T}_*] \Pr[\hat{\mathbf{x}}^\circ \in \mathcal{T}_*] \geq \frac{\epsilon_1}{(n+1)^2 \, |\mathcal{X}|}. \tag{155}$$

where $\epsilon_0(n) \to 0$ as $n \to \infty$.

But this contradicts Proposition 2 which states that the probability of a dithered quantization failure must decrease exponentially in n when the quantizer input is i.i.d.. Thus we conclude that the probability of dithered transcoder failure must approach zero as $n \to \infty$.

Acknowledgment

The authors wish to than Prof. Ioannis Kontoyiannis and Prof. Ram Zamir for helpful interactions.

References

[1] W. H. Equitz and T. M. Cover, "Successive refinement of information," *IEEE Trans. Inform. Theory*, vol. 37, pp. 269–275, Mar. 1991.

[2] D. Karakos and A. Papamarcou, "A relationship between quantization and watermarking rates in the presence of additive Gaussian attacks," *IEEE Trans. Inform. Theory*, vol. 49, pp. 1970–1982, Aug. 2003.

[3] D. Kundur, "Implications for high capacity data hiding in the presence of lossy compression," in *Proc. IEEE Int. Conf. On Info. Tech.: Coding & Comp.*, pp. 16–21, 2000.

[4] M. Ramkumar and A. N. Akansu, "Information theoretic bounds for data hiding in compressed images," in *Proc. IEEE Workshop on Multimedia Sig. Proc.*, 1998.

[5] H. Yamamoto, "Source coding theory for cascade and branching communications systems," *IEEE Trans. Info. Theory*, vol. 27, pp. 299–308, May 1981.

[6] B. Rimoldi, "Successive refinement of information: Characterization of the achievable rates," *IEEE Trans. Inform. Theory*, vol. 40, pp. 253–259, Jan. 1994.

[7] E.-H. Yang and Z. Zhang, "The redundancy of source coding with a fidelity criterion—part II: Coding at a fixed rate level with unknown statistics," *IEEE Trans. Inform. Theory*, vol. 47, pp. 126–145, Jan. 2001.

[8] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems.* Budapest, Hungary: Akadémiai Kiadó, 1986.

[9] T. M. Cover and J. A. Thomas, *Elements of Information Theory.* John Wiley and Sons, 1991.

[10] A. Kanlis, S. Khudanpur, and P. Narayan, "Typicality of a good rate-distortion code," *Problems of Information Transmission*, vol. 32, pp. 112–121, Jan. 1996.

[11] A. Kanlis, *Compression and Transmission of Information at Multiple Resolutions.* PhD thesis, University of Maryland at College Park, 1998.

[12] S. Shamai and S. Verdu, "The empirical distribution of good codes," *IEEE Trans. Inform. Theory*, vol. IT-43, pp. 836–846, May 1997.

[13] J. K. Wolf and J. Ziv, "Transmission of noisy information to a noisy receiver with minimum distortion," *IEEE Trans. Info. Theory*, vol. 16, pp. 406–411, July 1970.

[14] T. Berger, “Multiterminal source coding,” in *The Information Theory Approach to Communications* (G. Longo, ed.), ch. 4, Springer-Verlag, 1977.

[15] T. Berger, *Rate Distortion Theory.* Prentice-Hall, 1971.

[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] R. L. Dobrushin and B. S. Tsybakov, “Information transmission with additional noise,” *IEEE Trans. Inform. Theory*, vol. 8, pp. 293–304, Sept. 1962.

[18] A. D. Wyner, “The rate-distortion function for source coding with side information at the decoder–II: General sources,” *Information and Control*, vol. 38, pp. 60–80, 1978.

[19] A. Cohen and A. Lapidoth, “The Gaussian watermarking game,” *IEEE Trans. Inform. Theory*, vol. 48, June 2002.

[20] R. Zamir and M. Feder, “On universal quantization by randomized uniform/lattice quantizers,” *IEEE Trans. Inform. Theory*, vol. 38, pp. 428–436, Mar. 1992.

[21] R. Zamir and M. Feder, “On lattice quantization quantization noise,” *IEEE Trans. Inform. Theory*, vol. 42, pp. 1152–1159, July 1996.

[22] A. Lapidoth, “On the role of mismatch in rate distortion theory,” *IEEE Trans. Inform. Theory*, vol. 43, pp. 38–47, Jan. 1997.

[23] L. Lastras and T. Berger, “All sources are nearly successive refinable,” *IEEE Trans. Inform. Theory*, vol. 47, pp. 918–926, Mar. 2001.

[24] T. Linder and R. Zamir, “On the asymptotic tightness of the Shannon lower bound,” *IEEE Trans. Inform. Theory*, vol. 40, pp. 2026–2031, Nov. 1994.

[25] M. H. Costa, “Writing on dirty paper,” *IEEE Trans. Inform. Theory*, vol. 29, pp. 439–441, May 1983.

[26] 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.

[27] R. Zamir, S. Shamai, and U. Erez, “Nested linear/lattice codes for structured multiterminal binning,” *IEEE Trans. Inform. Theory*, vol. 48, pp. 1250–1276, June 2002.

[28] R. J. Barron, B. Chen, and G. W. Wornell, “The duality between information embedding and source coding with side information and some applications,” *IEEE Trans. Inform. Theory*, vol. 49, p. 1159, 2003.