Authentication with Distortion Criteria

Emin Martinian, Gregory W. Wornell, and Brian Chen

Submitted May 2002, Revised December 2003

Abstract

In a variety of applications, there is a need to authenticate a source that may have been degraded, transformed, edited, or otherwise modified, either intentionally or unintentionally. We develop a formulation of this problem, and identify and interpret the associated information-theoretic performance limits. The results are illustrated through application to binary sources with Hamming distortion measures, and to Gaussian sources with quadratic distortion measures. In each case, the associated systems are shown to perform substantially better than frequently proposed approaches based on combinations of source coding and information embedding techniques. Finally, efficient layered authentication systems are introduced as a natural extension of the basic results, and illustrated in the Gaussian-quadratic case.

Index Terms—multimedia, authentication, tamper-proofing, traitor-tracing, transaction-tracking, digital watermarking, anti-spoofing, digital signatures, information embedding, rate-distortion coding, coding with side information

1 Introduction

In traditional authentication problems, the goal is to determine whether a received message is an exact replica of what was sent. Digital signature techniques [1] are a natural tool for addressing such problems. However, in many emerging applications the message may be an audio or video waveform, and before being presented to a decoder the waveform may experience any of a variety of possible perturbations, including, for example, degradation due to noise or compression; transformation by filtering, resampling, or transcoding; or editing to annotate, enhance, or otherwise modify the waveform. Moreover, such perturbations may be intentional or unintentional, and benign or malicious. Methods for reliable authentication from such perturbed data are important as well.

This work has been supported in part by the National Science Foundation under Grant No. CCR-0073520, Microsoft Research, Hewlett-Packard through the MIT/HP Alliance, and Texas Instruments through the Leadership Universities Program. This work was presented in part at ISIT-2001, Washington, DC.

The authors are affiliated with the Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA 02139. (E-mail: {emin,gww,bchen}@mit.edu).

One motivating example involves the authentication of drivers' licenses. In such applications, one is interested in marking or otherwise modifying the pixels of the photograph to enable a decoder to determine whether — even in the presence of smudges, scratches or other artifacts of routine handling — the photograph is authentic and not a forgery. Such an encoding process can be viewed as a generalization of the current practice of imprinting holograms for such purposes. Frequently, one is constrained to keep the encoding distortion small, i.e., to choose the added markings to be effectively imperceptible so that the marked photograph can be useful even without an appropriate decoder (this is desirable e.g., for legacy systems). When a decoder is available, another goal is to use the added markings to produce a high fidelity authentic reconstruction.

An increasingly important class of motivating scenarios includes authenticatible content editing and transaction-tracking applications involving audio, video, and even text. Needs in this area arise as a result of the ease with which such content can be modified and distorted for either legitimate or fraudulent purposes. In this case, the creator of some content seeks to publish an encoding of it for editors. An editor seeks to modify the published version to enhance its value to the end-user. For the edited version to be of value to the user, it must be authenticatible, i.e., the user must be able to confirm whether it was indeed generated from what the creator published, and reliably and accurately assess what modifications were made by the editor. This requires that any signature applied during encoding survive the editing process. At the same time, for what the creator publishes to be useful to the editor, the encoding cannot differ significantly from the original content. Systems of this type can be used to facilitate authenticating scientific data, photographic images, and voice recordings in forensic and other contexts.

Researchers have proposed a variety of approaches to such problems based on digital watermarking, cryptography, and content classification [2] [3] [4] [5] [6] [7] [8] [9] [10] [11] [12] [13] [14] [15] [16] [17]. Ultimately, the methods developed to date implicitly or explicitly attempt to balance the competing goals of robustness to benign perturbations, security against tampering attacks, and encoding distortion. Some researchers [6] [13] [4] [12] propose authentication schemes which protect a signal by embedding what is referred to as a "fragile" watermark known to both encoder and decoder. The decoder extracts a watermark from the received signal and compares to the known watermark which was inserted by the encoder. The difference between the extracted watermark and the known watermark is then interpreted as a measure of authenticity. As an alternative, other

authors [3] [10] [14] propose a "robust" watermarking strategy, whereby the important features of the signal are extracted, compressed and embedded into the signal by the encoder. The decoder attempts to extract the watermark from the received signal and authenticates by comparing the features encoded in the watermark to the features in the received signal. This strategy is sometimes termed "self-embedding."

Despite the growing number of proposed systems, many basic questions remain about 1) how to best model the problem and what we mean by authentication, 2) what the associated fundamental performance limits are, and 3) what system structures can and cannot approach those limits. More generally, there are basic questions about the degree to which the authentication, digital watermarking, and data hiding problems are related or not.

While information-theoretic treatments of authentication problems is just emerging, there has been a growing literature in the information theory community on digital watermarking and data hiding problems, and more generally problems of coding with side information, much of which builds on the foundation of [18] [19] and [20]; see, e.g., [21] [22] [23] [24] [25] [26] [27] [28] [29] [30] [31] [32] [33] [34] [35] [36] [37] [38] [39] [40] [41] and the references therein. Collectively, this work provides a useful context within which to examine the topic of authentication.

Our contribution is to propose one possible model for the authentication problem, and rigorously examine its implications. In terms of performance limits, we assess the inherent trade-offs between security, robustness, and distortion, and develop the structure of systems that make these trade-offs efficiently. As we will show, these systems have important distinguishing characteristics from those proposed to date. We also see that under this model, the authentication problem is substantially different from familiar formulations of the digital watermarking and data hiding problems, and has a correspondingly different solution.

A detailed outline of the paper is as follows. We begin by briefly defining our notation in Section 2. Next in Section 3, we develop a system model and problem formulation, quantifying a notion of authentication. In Section 4, we characterize the performance limits of such systems via our main coding theorem. Section 5 contains the associated achievability proof, identifies the structure of good systems, and a converse. In Section 6 the results are applied to the case of binary sources with Hamming distortion measures, and in Section 7 to Gaussian sources with quadratic distortion measures. Section 8.1 then evaluates authentication techniques based on self-embedding

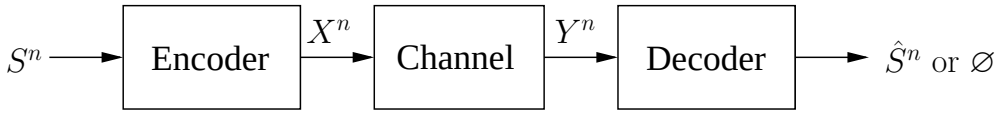

Figure 1: Authentication system model. The source S^n is encoded to create the channel input X^n, incurring some distortion. The channel models benign degradations due to routine handling and processing, transformations due to intentional editing, as well as tampering by a malicious attacker. The decoder produces from the channel output Y^n either an authentic reconstruction $\hat{S}^n$ of the source to within some fidelity, or indicates that authentication is not possible.

in the context of our problem model and system structure, and shows that self-embedding can be quite inefficient in typical regimes of interest. Next, Section 9 generalizes the results of the paper to include layered systems that support multiple levels of authentication. Finally, Section 10 contains some concluding remarks.

2 Notation

We use standard information theory notation (e.g., as found in [42]). Specifically, $E[A]$ denotes expectation of the random variable A, $H(A)$ and $I(B;C)$ denote entropy and mutual information, and $A \leftrightarrow B \leftrightarrow C$ denotes the Markov condition that random variables A and C are independent given B. We use the notation v_i^j to denote the sequence $\{v_i, v_{i+1}, \ldots, v_j\}$, and define $v^n = v_1^n$. Alphabets are denoted by uppercase calligraphic letters, e.g., $\mathcal{S}$, $\mathcal{X}$. We use $|\cdot|$ to denote the cardinality of a set or alphabet.

3 System Model and Problem Formulation

Our system model is as depicted in Fig. 1. To simplify the exposition, we model the original source as an independent and identically distributed (i.i.d.)[1] sequence $S_1, S_2, \ldots, S_n$. In practice S^n could correspond to sample values or signal representations in some suitable basis.

The encoder takes as input the block of n source samples S^n, producing an output X^n that

[1]Our results do not depend critically on the i.i.d. property, which is chosen for convenience. In fact, the i.i.d. model is sometimes pessimistic; better performance can often be obtained by taking advantage of correlation present in the source or channel. We believe that qualitatively similar results could be obtained for more general settings (e.g., using techniques from [43], [44]).

is suitably close to S^n with respect to some distortion measure. The encoded signal then passes through a channel, which captures the effects of routine handling and editing as well as any tampering, producing the channel output Y^n.

The decoder either produces, to within some fidelity as quantified by a suitable distortion measure, a reconstruction $\hat{S}^n$ of the source that is guaranteed to be free from the effects of any tampering by an attacker, or declares that it is not possible to produce such a reconstruction. We term such reconstructions "authentic."

Determining or predicting the realized channel for the authentication scenario of Fig. 1 is often difficult or even impossible. For example, an honest editor or a malicious attacker can always choose to modify the signal in an unexpected manner invalidating whatever channel model is selected. Hence we deliberately avoid choosing a channel model to try and represent the exact set of degradations likely to be encountered.

Instead, we treat permissible editing operations differently than tampering or other unanticipated modifications (e.g., more scratches, compression, or other routine degradations than expected). When the latter occur, the decoder may declare an authentication failure or even produce an authentication reconstruction, but should not be fooled into producing a reconstruction which is not authentic. For permissible editing, however, the decoder should almost always produce an authentic reconstruction.

To characterize what constitutes permissible editing, we define a "reference channel" as a probability distribution, $p(Y^n|X^n)$, which describes the relationship between the channel input and output for all benign and malicious perturbations we desire the system to overcome. For example, an authentication system for a binary source might be designed to allow a fraction p of the source samples to be flipped in order to highlight important regions and account for changes in compression format. In this case, the reference channel would be the familiar (memoryless) binary symmetric channel with cross-over probability p. When the reference channel is in effect and the received signal Y^n is generated from the encoded source X^n according to the reference channel distribution, the decoder must produce an authentic reconstruction. But if Y^n is generated according to a distribution incompatible with the reference channel (e.g., due to tampering by an attacker), then the decoder may instead declare an authentication failure.

Since the reference channel is a design parameter we naturally assume that it is known to the

encoder and decoder as well as any attackers even though the realized channel may be unknown. To properly interpret the technical results and apply them to design, model, and evaluate authentication systems we believe the reference channel view is valuable and hence we adopt this terminology throughout the paper. We will focus on memoryless, probabilistic reference channels so the analysis when the reference channel is in effect follows traditional approaches. The main conceptual difference is that while a classical channel typically defines the complete input-output relationship, the reference channel only describes the input-output relationship for permissible editing operations; potentially arbitrary tampering is not modeled directly but handled through a different mechanism.

Considering the motivating examples at the outset of the paper, given a particular reference channel, the goal of the system designer is to make the encoding distortion small, so that the what the creator publishes is a faithful replica of the original source, and to make any authentic reconstructions produced by the decoder of high fidelity, so that the user can accurately assess directly from the edited content how it differs from the original source. In general, these are conflicting objectives, and in the sequel we explore the fundamental trade-offs involved. [2]

3.1 Defining "Authentic"

Many notions of authentication are possible. The one used in digital signatures is perhaps the simplest: a received signal is authentic if and only if it is exactly the same as the encoded signal. Of course, as discussed in the introduction, this definition is too restrictive to be useful for the scenario considered in Fig. 1. When perturbations to the encoded signal are allowed, no canonical definition has yet emerged. We propose the following definition and briefly discuss other notions of authenticity and their shortcomings in Section 8.2.

Definition 1 *A reconstruction $\hat{S}^n$ produced by the decoder from the output Y^n of the channel is said to be authentic if it satisfies the Markov condition below:*

$$\hat{S}^n \leftrightarrow \{S^n, X^n\} \leftrightarrow Y^n \tag{1}$$

For example, this condition will be satisfied if $\hat{S}^n$ is a deterministic or randomized function of S^n.

[2] Similar trade-offs are present in joint source-channel coding problems with uncertain channels [45] [46] [47].

Of course, the decoder may fail to successfully decode $\hat{S}^n$ from the channel output Y^n. To avoid confusing security and decoding error, however, our security requirement is defined in the case that decoding succeeds and we deal with the probability of decoding error separately. The advantage of this approach is that if an authentication system produces a reconstruction satisfying (1), then a user can be completely confident that he will be unaffected by any actions of a malicious adversary.[3]

As our main result, in Section 4 we characterize when authentication systems are possible, and when they are not. Specifically, let D_{e} denote the encoding distortion, i.e., the distortion experienced in the absence of a channel, and let D_{r} denote the distortion in the reconstruction produced by the decoder when the signal can be authenticated, i.e., when the channel transformations are consistent with the reference distribution $p(y|x)$. Then we determine which distortion pairs $(D_{\mathrm{e}}, D_{\mathrm{r}})$ are asymptotically achievable.

3.2 An Example Distortion Region

Before developing our main result, we illustrate with an example the kinds of results that will be obtained. An illustrative achievable distortion region is depicted in Fig. 2. This example corresponds to a problem involving a symmetric Bernoulli source, Hamming distortion measures, and a binary symmetric reference channel with crossover probability p. Note that at the point $(D_{\mathrm{e}}, D_{\mathrm{r}}) = (p, p)$, the decoder completely eliminates the effects of the reference channel when it is in effect: the minimum achievable reconstruction distortion D_{r} is the same as the distortion D_{e} at the output of the encoder. Observe, too, that the case $p = 0$ corresponds to the traditional scenario for digital signatures where there is no noise. In this case, as the figure reflects, authentication is achievable without incurring any encoding distortion nor reconstruction distortion.

[3] A disadvantage is that this definition may be unnecessarily strict; a different definition may capture a satisfactory notion of authentication with fewer limits on system design. We defer further comments on other notions of authenticity to Sections 8.2 and 10.

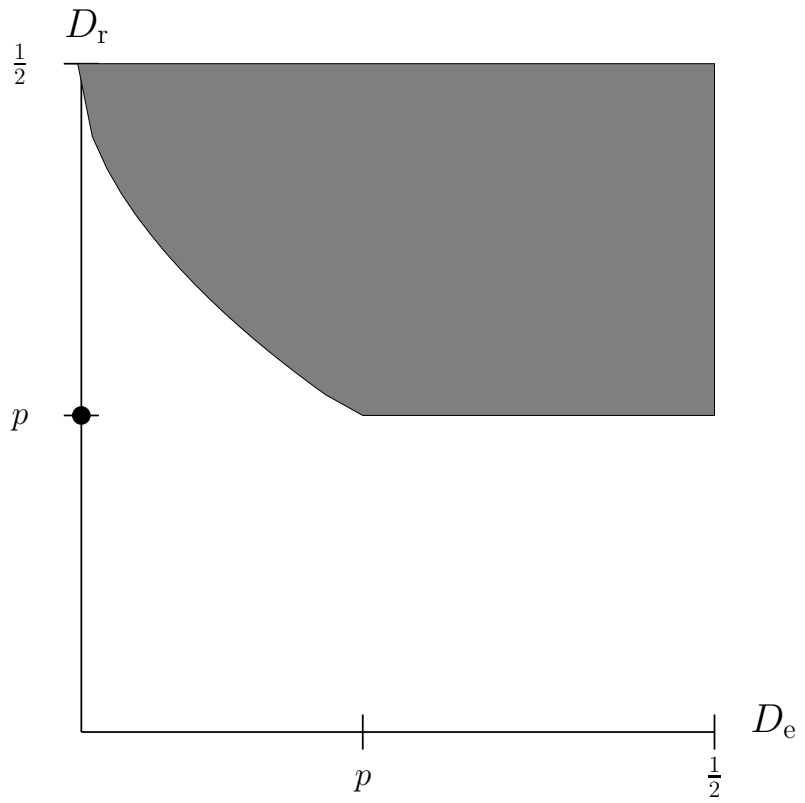

Figure 2: An achievable distortion region for a symmetric Bernoulli source transmitted over a binary symmetric reference channel with crossover probability p. Distortions are with respect to the Hamming measure. The case $p = 0$ corresponds to traditional digital signatures. If authentication was not required, the point $(D_{\mathrm{e}} = 0, D_{\mathrm{r}} = p)$ could be achieved.

4 Characterization of Solution: Coding Theorems

An instance of the authentication problem consists of the tuple

$$\{\mathcal{S}, p(s), \mathcal{X}, \mathcal{Y}, p(y|x), d_{\mathrm{e}}(\cdot,\cdot), d_{\mathrm{r}}(\cdot,\cdot)\}\,. \tag{2}$$

We use $\mathcal{S}$ to denote the source alphabet—which is finite unless otherwise indicated—and $p(s)$ is its (i.i.d.) distribution. The channel input and output alphabets are $\mathcal{X}$ and $\mathcal{Y}$ and $p(y|x)$ is the (memoryless) reference channel law. Finally, $d_{\mathrm{e}}(\cdot,\cdot)$ and $d_{\mathrm{r}}(\cdot,\cdot)$ are the encoding and reconstruction distortion measures.

A solution to this problem (i.e., an authentication scheme) consists of an algorithm that returns an encoding function Υ_n, a decoding function Φ_n, and a secret key θ.[4] The secret key is shared only between the encoder and decoder; all other information is known to all parties including attackers.

The secret key θ is a k-bit sequence with k sufficiently large. The encoder is a mapping from the source sequence and the secret key to codewords, i.e.,

$$\Upsilon_n(S^n, \theta): \quad \mathcal{S}^n \times \{0,1\}^k \mapsto \mathcal{X}^n.$$

The decoder is a mapping from the channel output and the secret key to either an authentic source reconstruction $\hat{S}^n$ (i.e., one satisfying (1)) or the special symbol $\varnothing$ that indicates such a reconstruction is not possible; whence,

$$\Phi_n\left(Y^n, \theta\right): \quad \mathcal{Y}^n \times \{0,1\}^k \mapsto \mathcal{S}^n \cup \{\varnothing\}.$$

Notice that since an authentic reconstruction must satisfy (1), and since the decoder must satisfy the Markov condition $\{S^n, X^n\} \leftrightarrow Y^n \leftrightarrow \Phi_n\left(Y^n, \theta\right)$, we have that $\hat{S}^n \leftrightarrow \{S^n, X^n\} \leftrightarrow \Phi_n\left(Y^n, \theta\right)$ forms a Markov chain only *when successful decoding occurs*. Thus, the authentic reconstruction $\hat{S}^n$ should be defined as a quantity that the decoder attempts to deduce since defining $\hat{S}^n = \Phi_n\left(Y, \theta^n\right)$ will generally not satisfy (1).

Henceforth, except when there is risk of confusion, we omit both the subscript n and the secret

[4] To focus the exposition we describe only private-key schemes in this paper, but public-key implementations can be developed following, e.g., the general approach outlined in [48] and are discussed briefly in the appendix.

key argument from the encoding and decoding function notation, letting the dependence be implicit. Moreover, when the encoder and/or decoder are randomized functions, then all probabilities are taken over these randomizations as well as the source and channel law.

The relevant distortions are the encoding and decoding distortion computed as the sum of the respective (bounded) single letter distortion functions d_{e} and d_{r}, i.e.,

$$\frac{1}{n}\sum_{i=1}^{n} d_{\mathrm{e}}(S_i, X_i) \qquad \text{and} \qquad \frac{1}{n}\sum_{i=1}^{n} d_{\mathrm{r}}(S_i, \Phi_i\left(Y^n\right)).$$

Evidently,

$$d_{\mathrm{e}}: \quad \mathcal{S} \times \mathcal{X} \mapsto \mathbb{R}^+ \tag{3}$$

$$d_{\mathrm{r}}: \quad \mathcal{S} \times \mathcal{S} \mapsto \mathbb{R}^+. \tag{4}$$

The system can fail in one of three ways. The first two failure modes correspond to either the encoder introducing excessive encoding distortion, or the decoder failing to produce an authentic reconstruction with acceptable distortion when the reference channel is in effect. Accordingly, we define the overall distortion violation error event to be

$$\mathcal{E}_{\mathrm{dv}} = \mathcal{E}_{D_{\mathrm{e}}} \cup \mathcal{E}_{D_{\mathrm{r}}} \tag{5}$$

where, for any $\epsilon > 0$,

$$\mathcal{E}_{D_{\mathrm{e}}} = \left\{ \frac{1}{n}\sum_{i=1}^{n} d_{\mathrm{e}}(S_i, X_i) > D_{\mathrm{e}} + \epsilon \right\} \tag{6}$$

$$\mathcal{E}_{D_{\mathrm{r}}} = \left\{ \Phi_n\left(Y^n\right) = \varnothing \right\} \cup \left\{ \frac{1}{n}\sum_{i=1}^{n} d_{\mathrm{r}}(S_i, \Phi_i\left(Y^n\right)) > D_{\mathrm{r}} + \epsilon \right\} \cap \left\{ \Phi_n\left(Y^n\right) \neq \varnothing \right\}. \tag{7}$$

In the remaining failure mode, the system fails to produce the desired authentic reconstruction $\hat{S}^n$ from the channel output and instead of declaring that authentication is not possible produces an incorrect estimate. Specifically, we define the successful attack event according to

$$\mathcal{E}_{\mathrm{sa}} = \left\{ \Phi\left(Y^n\right) \neq \varnothing \right\} \cap \left\{ \Phi\left(Y^n\right) \neq \hat{S}^n \right\}. \tag{8}$$

Definition 2 *The achievable distortion region for the problem* (2) *is the closure of the set of pairs* $(D_{\mathrm{e}}, D_{\mathrm{r}})$ *such that there exists a sequence of authentication systems, indexed by* n*, where for every* $\epsilon > 0$ *and as* $n \to \infty$*,* $\Pr[\mathcal{E}_{\mathrm{sa}}] \to 0$ *regardless of the channel law in effect,* $\Pr[\mathcal{E}_{D_{\mathrm{e}}}] \to 0$*, and* $\Pr[\mathcal{E}_{D_{\mathrm{r}}}] \to 0$ *when the reference channel is in effect, with* $\mathcal{E}_{\mathrm{sa}}$*,* $\mathcal{E}_{D_{\mathrm{e}}}$ *and* $\mathcal{E}_{D_{\mathrm{r}}}$ *as defined in* (8)*,* (6)*, and* (7)*.*

For such systems, we have the following coding theorem:

Theorem 1 *The distortion pair* $(D_{\mathrm{e}}, D_{\mathrm{r}})$ *lies in the achievable distortion region for the problem* (2) *if and only if there exist functions* $f(\cdot,\cdot)$*,* $g(\cdot)$ *and a distribution* $p(y,x,u,s) = p(s)p(u|s)p(x|u,s)p(y|x)$ *with* X *deterministic (i.e.* $p(x|u,s) = 1_{x=f(s,u)}$*) such that*

$$I(U;Y) - I(S;U) \geq 0 \tag{9a}$$

$$E[d_{\mathrm{e}}(S, f(U,S))] \leq D_{\mathrm{e}} \tag{9b}$$

$$E[d_{\mathrm{r}}(S, g(U))] \leq D_{\mathrm{r}}. \tag{9c}$$

The alphabet $\mathcal{U}$ *of the auxiliary random variable* U *requires cardinality* $|\mathcal{U}| \leq (|\mathcal{S}|+|\mathcal{X}|+3)\cdot|\mathcal{S}|\cdot|\mathcal{X}|$*.*[5]

Essentially, the auxiliary random variable U represents an embedded description of the source which can be authenticated, X represents the channel input, and $g(U)$ in (9c) represents the authentic reconstruction. The usual condition that the channel output is determined from the channel input (i.e., the encoder does not know what the channel output will be until after the channel input is fixed) is captured by the requirement that the full joint distribution $p(y,x,u,s)$ factors as shown above. The requirement (1) that the authentic reconstruction does not depend directly on the channel output is captured by the fact that $g(\cdot)$ depends only on U and not on Y. Without the authentication requirement, the set of achievable distortion pairs can be enlarged by allowing the reconstruction to depend on the channel output, i.e. $g(U)$ in (9c) can be replaced by $g(U,Y)$.[6] Thus, as we shall see in Sections 6 and 7, security comes at a price in this problem.

As an aside, note that Theorem 1 can be contrasted with its information embedding counterpart, which as generalized from [18] in [35], states that a pair (R, D_{e}), where R is the embedding

[5]If instead $f(U,S)$ is allowed to be a non-deterministic mapping, then it is sufficient to consider distributions where the auxiliary random variable has the smaller alphabet $|\mathcal{U}| \leq |\mathcal{S}| + |\mathcal{X}| + 3$.

[6]The achievable distortion for transmitting a source across a channel can be derived without an auxiliary random variable. The advantage of using (9c) with $g(Y)$ replaced by $g(U,Y)$, however, is that it facilitates a comparison.

rate, is achievable if and only if there exists a function $f(\cdot,\cdot)$ and a distribution $p(y,x,u,s) = p(s)p(u|s)p(x|s,u)p(y|x)$ with X deterministic (i.e. $p(x|u,s) = 1_{x=f(s,u)}$) such that

$$I(U;Y) - I(S;U) \geq R \tag{10a}$$

$$E[d_{\mathrm{e}}(S, f(U,S))] \leq D_{\mathrm{e}} \tag{10b}$$

Thus we see that the authentication problem is substantially different from the information embedding problem.

Before developing the proofs of Theorem 1, to develop intuition we describe the general system structure, and its specialization to the Gaussian-quadratic case.

4.1 General System Structure

As developed in detail in Section 5, an optimal authentication system can be constructed by choosing a codebook $\mathcal{C}$ with codewords appropriately distributed over the space of possible source outcomes. A randomly chosen subset of these codewords $\mathcal{A} \subset \mathcal{C}$ are marked as admissible and the knowledge of $\mathcal{A}$ is a secret shared between the encoder and decoder, and kept from potential attackers.

The encoder maps (quantizes) the source S^n to the nearest admissible codeword U^n and then generates the channel input X^n from U^n. The decoder maps its received signal to the nearest codeword $C^n \in \mathcal{C}$. If $C^n \in \mathcal{A}$, i.e., C^n is an admissible codeword, the decoder produces the reconstruction $\hat{S}^n$ from C^n. If $C^n \notin \mathcal{A}$, i.e., C^n is not admissible, the decoder declares that an authentic reconstruction is not possible.

Observe that the $\mathcal{A}$ must have the following three characteristics. First, to avoid a successful attack the number of admissible codewords must be appropriately small. Indeed, since the attacker does not know $\mathcal{A}$, if the attacker's tampering causes the decoder to decode to any codeword other than U^n then the probability that the decoder is fooled by the tampering and does not declare a decoding failure is bounded by $|\mathcal{A}| \,/\, |\mathcal{C}|$. Second, to avoid an encoding distortion violation, the set of admissible codewords should be dense enough to allow the encoder to find an appropriate X^n near S^n. Third, to avoid a reconstruction distortion violation, the decoder should be able to distinguish the possible encoded signals at the output of the reference channel. Thus the codewords

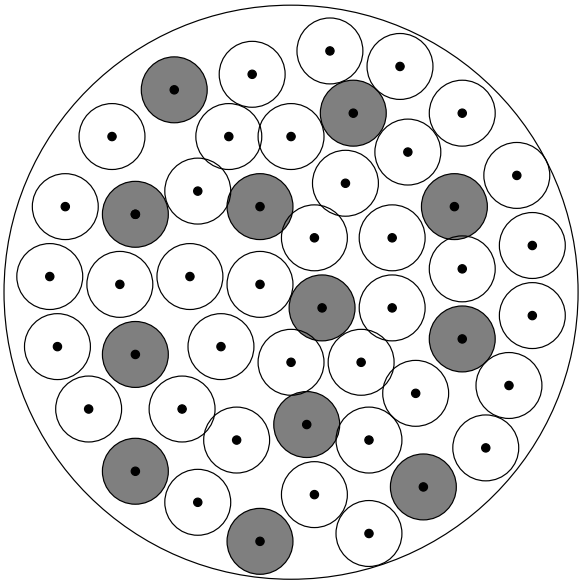

Figure 3: Codebook construction for the Gaussian-quadratic scenario. The large sphere represents the space of possible source vectors and the small spheres representing the noise are centered on codewords. When the small spheres do not overlap the codewords can be resolved at the output of the reference channel. The shaded spheres represent the admissible codewords—a secret known only to the encoder and decoder.

should be sufficiently separated that they can be resolved at the output of the reference channel.

4.1.1 Geometry for Gaussian-Quadratic Example

We illustrate the system geometry in the case of a white Gaussian source, quadratic distortion measure, and an additive white Gaussian noise reference channel, in the high signal-to-noise ratio (SNR) regime. We let σ_S^2 and σ_N^2 denote the source and channel variances, respectively. For this example, we can construct $\mathcal{C}$ by packing codewords into the space of possible source vectors such that no codeword is closer than some distance $r\sqrt{n}$ to any other, i.e., packing spheres of radius $r\sqrt{n}$ into a sphere of radius $\sigma_S\sqrt{n}$ where the center of the spheres correspond to codewords. Next, a fraction $2^{-n\gamma}$ of the codewords in $\mathcal{C}$ are chosen at random and marked as admissible to form $\mathcal{A}$. It suffices to let $\gamma = 1/\sqrt{n}$ and $r^2 = \sigma_N^2 + \epsilon$ for some $\epsilon > 0$ that is arbitrarily small. This construction is illustrated in Fig. 3.

The encoder maps the source S^n to a nearby admissible codeword U^n, which it chooses as the transmitted codeword X^n. Since the number of admissible codewords in a sphere of radius d

centered on S^n is roughly

$$\frac{|\mathcal{A}|}{|\mathcal{C}|} \cdot \left(\frac{d}{r}\right)^n,$$

on average there exists at least one codeword within distance d of the source provided $d \geq r2^{\gamma}$. Thus, the average encoding distortion is roughly $r^2 2^{2\gamma}$, which approaches $\sigma_N^2 + \epsilon$ as $n \to \infty$.

The authentic reconstruction is $\hat{S}^n = U^n$. Thus, when the decoder correctly identifies U^n, the reconstruction distortion is the same as the encoding distortion. And when the reference channel is in effect, the decoder does indeed correctly identify U^n. This follows from the fact that with high probability, the reference channel noise creates a perturbation within a noise sphere of radius $\sigma_N\sqrt{n}$ about the transmitted signal X^n, and the noise spheres do not intersect since $r > \sigma_N$.

Furthermore, when the reference channel is not in effect and the attacker tampers with the signal such that the receiver decodes to a codeword C different from the transmitted codeword U^n, then the probability that C was marked as admissible in the codebook construction phase is

$$\Pr[C \in \mathcal{A} | C \neq U^n] = \frac{|\mathcal{A}|}{|\mathcal{C}|} = 2^{-n\gamma},$$

which goes to zero as $n \to \infty$. The decoder generates $\varnothing$ if it decodes to a non-admissible codeword, so the probability of a nonauthentic reconstruction is vanishingly small.

Thus the distortions $D_{\mathrm{e}} = D_{\mathrm{r}} = \sigma_N^2$ can be approached with an arbitrarily small probability of successful attack. See [49] [48] for insights into the practical implementation of this class of systems including those designed based on a public key instead of a secret key.

5 Proofs

5.1 Forward Part: Sufficiency

Here we show that if there exist distributions and functions satisfying (9), then for every $\epsilon > 0$ there exists a sequence of authentication system with distortion at most $(D_{\mathrm{e}} + \epsilon, D_{\mathrm{r}} + \epsilon)$. Since the achievable distortion region is a closed set this implies that $(D_{\mathrm{e}}, D_{\mathrm{r}})$ lies in the achievable distortion region.

We prove this forward part of Theorem 1 by showing the existence of a random code with the desired properties.

5.1.1 Codebook Generation

We begin by choosing some $\gamma > 0$ such that

$$I(Y;U) - I(U;S) > 3\gamma. \tag{11}$$

where γ decays to zero more slowly than $1/n$, i.e.,

$$\gamma \to 0 \text{ and } n\gamma \to \infty \text{ as } n \to \infty \tag{12}$$

Given the choice of γ, the encoder chooses a random codebook $\mathcal{C}$ of rate

$$R = I(S;U) + 2\gamma. \tag{13}$$

Each codeword in $\mathcal{C}$ is a sequence of 2^{nR} i.i.d. random variables selected according to the distribution $p(u) = \sum_{s \in \mathcal{S}} p(u|s)p(s)$. Then, for each realized codebook $\mathcal{C}$ the encoder randomly marks $2^{n(R-\gamma)}$ of the codewords in $\mathcal{C}$ as admissible and the others as forbidden. We denote this new codebook of admissible codewords as $\mathcal{A}$, which has effective rate

$$R' = R - \gamma = I(S;U) + \gamma, \tag{14}$$

where the last equality follows from substituting (13). The knowledge of which codewords are forbidden is the secret key and is revealed only to the decoder. The codebook $\mathcal{C}$ is publicly revealed.

5.1.2 Encoding and Decoding

The encoder first tries to find an admissible codeword $u^n \in \mathcal{A}$ that is δ-strongly jointly typical with its source sequence S^n according to $p(u|s)$. If the codeword $u^n \in \mathcal{A}$ is found to be typical, the encoder output is produced by mapping the pair (s^n, u^n) into x^n via $x = f(s,u)$. If no jointly typical admissible codeword exists, the encoder expects the system to fail, and thus sends an arbitrary codeword.

The decoder attempts to produce the authentic reconstruction $\hat{s}^n = g^n(u^n)$ where

$$g^n(u^n) = (g(u_1), g(u_2), \ldots, g(u_n)). \tag{15}$$

The decoder $\Phi(\cdot)$ tries to deduce $\hat{s}^n$ by searching for a unique admissible codeword $\hat{u}^n \in \mathcal{A}$ that is δ-strongly jointly typical with the received sequence Y^n. If such a codeword is found the reconstruction produced is $g^n(\hat{u}^n)$. If no such unique codeword is found, the decoder produces the output symbol $\varnothing$.

5.1.3 System Failure Probabilities

We begin by analyzing the system failure probabilities.

Probability of Successful Attack. Suppose the attacker causes the received codeword to be jointly typical with a unique codeword $c^n \in \mathcal{C}$. Since the attacker has no knowledge of which codewords are admissible, the probability that codeword c^n was chosen as admissible in the codebook construction phase is

$$\Pr[c^n \in \mathcal{A}] = \frac{|\mathcal{A}|}{|\mathcal{C}|} = \frac{2^{nR'}}{2^{nR}} = 2^{-n\gamma}.$$

where we have used (14) and (13). Therefore,

$$\Pr[\mathcal{E}_{\text{sa}}] \leq \Pr[\Phi(Y^n) \neq \varnothing \mid \Phi(Y^n) \neq \hat{S}^n] = 2^{-n\gamma}.$$

which goes to zero according to (12). Note that this argument applies regardless of the method used by the attacker since without access to the secret key its actions are statistically independent of which codewords are admissible.

Probability of Distortion Violation. The distortion violation events $\mathcal{E}_{D_{\text{e}}}$ and $\mathcal{E}_{D_{\text{r}}}$ defined in (6) and (7) can arise due to any of the following typicality failure events:

- $\mathcal{E}_{\text{st}}$: The source is not typical.
- $\mathcal{E}_{\text{et}}$: The encoder fails to find an admissible codeword that is jointly typical with its input.

- $\mathcal{E}_{\mathrm{ct}}$: The channel fails to produce an output jointly typical with its input when the reference channel law is in effect.
- $\mathcal{E}_{\mathrm{dt}}$: The decoder fails to find a codeword jointly typical with its input when the reference channel law is in effect.

A distortion violation event can also occur if there is no typicality failure but the distortion is still too high. Letting

$$\mathcal{E}_{\mathrm{tf}} = \mathcal{E}_{\mathrm{st}} \cup \mathcal{E}_{\mathrm{et}} \cup \mathcal{E}_{\mathrm{ct}} \cup \mathcal{E}_{\mathrm{dt}} \tag{16}$$

denote the typicality failure event, we have then that the probability of a distortion violation can be expressed as

$$\begin{aligned}\Pr[\mathcal{E}_{\mathrm{dv}}] &= \Pr[\mathcal{E}_{\mathrm{dv}} \mid \mathcal{E}_{\mathrm{tf}}] \cdot \Pr[\mathcal{E}_{\mathrm{tf}}] + \Pr[\mathcal{E}_{\mathrm{dv}} \mid \mathcal{E}_{\mathrm{tf}}^{\mathrm{c}}] \cdot \Pr[\mathcal{E}_{\mathrm{tf}}^{\mathrm{c}}] \leq \Pr[\mathcal{E}_{\mathrm{dv}} \mid \mathcal{E}_{\mathrm{tf}}^{\mathrm{c}}] + \Pr[\mathcal{E}_{\mathrm{tf}}] \\ &= \Pr\left[\mathcal{E}_{\mathrm{dv}} \mid \mathcal{E}_{\mathrm{tf}}^{\mathrm{c}}\right] + \Pr[\mathcal{E}_{\mathrm{st}}] + \Pr[\mathcal{E}_{\mathrm{et}} \mid \mathcal{E}_{\mathrm{st}}^{\mathrm{c}}] + \Pr[\mathcal{E}_{\mathrm{ct}} \mid \mathcal{E}_{\mathrm{st}}^{\mathrm{c}}, \mathcal{E}_{\mathrm{et}}^{\mathrm{c}}] + \Pr[\mathcal{E}_{\mathrm{dt}} \mid \mathcal{E}_{\mathrm{st}}^{\mathrm{c}}, \mathcal{E}_{\mathrm{et}}^{\mathrm{c}}, \mathcal{E}_{\mathrm{ct}}^{\mathrm{c}}] \end{aligned} \tag{17}$$

First, according to well-known properties of typical sequences [42], by choosing n large enough we can make

$$\Pr[\mathcal{E}_{\mathrm{st}}] \leq \epsilon/4 \tag{18}$$

$$\Pr[\mathcal{E}_{\mathrm{ct}} \mid \mathcal{E}_{\mathrm{st}}^{\mathrm{c}}, \mathcal{E}_{\mathrm{et}}^{\mathrm{c}}] \leq \epsilon/4. \tag{19}$$

Second, provided that the source is typical, the probability that the encoder fails to find a sequence $u^n \in \mathcal{A}$ jointly typical with the source follows from (14) as

$$\Pr[\mathcal{E}_{\mathrm{et}} \mid \mathcal{E}_{\mathrm{st}}^{\mathrm{c}}] \leq 2^{-n[R'-I(S;U)]} = 2^{-n\gamma} \tag{20}$$

from standard joint typicality arguments.

Third,

$$\Pr[\mathcal{E}_{\mathrm{dt}} \mid \mathcal{E}_{\mathrm{st}}^{\mathrm{c}}, \mathcal{E}_{\mathrm{et}}^{\mathrm{c}}, \mathcal{E}_{\mathrm{ct}}^{\mathrm{c}}] \leq 2^{-n\gamma} + \epsilon/4. \tag{21}$$

Indeed, using standard joint typicality results, the probability that the received sequence Y^n is not δ-strongly jointly typical with the correct codeword U^n selected by the encoder can be made

smaller than $\epsilon/4$ for n large enough, and the probability of it being strongly jointly typical with any other admissible codeword is, using (11) with (13), at most

$$2^{-n[I(U;Y)-R]} \leq 2^{-n\gamma}.$$

Fourth,

$$\Pr\left[\mathcal{E}_{\mathrm{dv}} \mid \mathcal{E}_{\mathrm{tf}}^{\mathrm{c}}\right] = 0. \tag{22}$$

Indeed, provided there are no typicality failures, the pair (S^n, Y^n) must be strongly jointly typical, so by the standard properties of strong joint typicality,

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} d_{\mathrm{e}}(S_i, X_i) &\leq E[d_{\mathrm{e}}(S, X)] + \delta \cdot \bar{d}_1 \\ \frac{1}{n}\sum_{i=1}^{n} d_{\mathrm{r}}(S_i, g_i(U_i)) &\leq E[d_{\mathrm{r}}(S, g(U))] + \delta \cdot \bar{d}_2,\end{aligned}$$

where $\bar{d}_1$ and $\bar{d}_2$ are bounds defined via

$$\bar{d}_1 = \sup_{(s,x)\in\mathcal{S}\times\mathcal{X}} d_{\mathrm{e}}(s, x) \tag{23}$$

$$\bar{d}_2 = \sup_{(s,\hat{s})\in\mathcal{S}\times\mathcal{S}} d_{\mathrm{r}}(s, \hat{s}). \tag{24}$$

Thus, choosing δ such that

$$\delta < \max\left(\frac{\epsilon}{\bar{d}_1}, \frac{\epsilon}{\bar{d}_2}\right)$$

and making n large enough we obtain (22).

Finally, using (18), (19), (20), (21), and (22) in (17) we obtain

$$\Pr[\mathcal{E}_{\mathrm{dv}}] \leq 3\epsilon/4 + 2 \cdot 2^{-n\gamma} \tag{25}$$

which can be made less than ϵ for n large enough. Thus $\Pr[\mathcal{E}_{D_{\mathrm{e}}}] \to 0$ and, when the reference channel is in effect, $\Pr[\mathcal{E}_{D_{\mathrm{r}}}] \to 0$.

■

5.2 Converse Part: Necessity

Here we show that if there exists an authentication system where the pair $(D_{\mathrm{e}}, D_{\mathrm{r}})$ is in the achievable distortion region, then there exists a distribution $p(u|s)$ and functions $g(\cdot)$, $f(\cdot,\cdot)$ satisfying (9). In order to apply previously developed tools, it is convenient to define the rate-function

$$R^*(D_{\mathrm{e}}, D_{\mathrm{r}}) \stackrel{\Delta}{=} \sup_{p(U|S), f:\mathcal{U}\times\mathcal{S}\mapsto\mathcal{X}, g:\mathcal{U}\mapsto\mathcal{S}: E[d_{\mathrm{e}}(S,f(U,S))]\le D_{\mathrm{e}}, E[d_{\mathrm{r}}(S,g(U))]\le D_{\mathrm{r}}} I(U;Y) - I(S;U). \tag{26}$$

Note that $R^*(D_{\mathrm{e}}, D_{\mathrm{r}}) \ge 0$ if and only if the conditions in (9) are satisfied. Thus our strategy is to assume that the sequence of encoding and decoding functions discussed in Section 4 exist with $\lim_{n\to\infty}\Pr[\mathcal{E}_{\mathrm{sa}}] = 0$, $\lim_{n\to\infty}\Pr[\mathcal{E}_{D_{\mathrm{e}}}] = 0$, and—when the reference channel is in effect—$\lim_{n\to\infty}\Pr[\mathcal{E}_{D_{\mathrm{r}}}] = 0$. We then show that these functions imply that $R^*(D_{\mathrm{e}}, D_{\mathrm{r}}) \ge 0$ and hence (9) is satisfied.

To begin we note that it suffices to choose $g(\cdot)$ to be the minimum distortion estimator of S given U. Next, by using techniques from [18] or by directly applying [35, Lemma 2] it is possible to prove that allowing X to be non-deterministic has no advantage, i.e.,

$$R^*(D_{\mathrm{e}}, D_{\mathrm{r}}) \ge \sup_{p(U|S), p(X|U,S): E[d_{\mathrm{e}}(S,X)]\le D_{\mathrm{e}}, E[d_{\mathrm{r}}(S,g(U))]\le D_{\mathrm{r}}} I(U;Y) - I(S;U). \tag{27}$$

Arguments similar to those in [18] and [35, Lemma 1] show that $R^*(D_{\mathrm{e}}, D_{\mathrm{r}})$ is monotonically non-decreasing and concave in $(D_{\mathrm{e}}, D_{\mathrm{r}})$. These properties will later allow us to make use of the following lemma, whose proof follows readily from that of Lemma 4 in [18]:

Lemma 1 *For arbitrary random variables $V, A_1, A_2, \ldots, A_n$ and a sequence of i.i.d. random variables $S_1, S_2, \ldots, S_n$,*

$$\sum_{i=1}^{n} \left[I(V, A_1^{i-1}, S_{i+1}^n; A_i) - I(V, A_1^{i-1}, S_{i+1}^n; S_i) \right] \ge I(V; A^n) - I(V; S^n). \tag{28}$$

As demonstrated by the following Lemma, a suitable U_i is

$$U_i = (\hat{S}^n, Y_1^{i-1}, S_{i+1}^n). \tag{29}$$

Lemma 2 *The choice of U_i in* (29) *satisfies the Markov relationship*

$$Y_i \leftrightarrow (S_i, X_i) \leftrightarrow U_i. \tag{30}$$

Proof:

It suffices to note that

$$p(y_i|x_i, s_i) = p(y_i|x_i) = \frac{p(y_1^i|x^n)}{p(y_1^{i-1}|x^n)} = \frac{p(y_1^i|x^n, s^n)}{p(y_1^{i-1}|x^n, s^n)} \tag{31}$$

$$= \frac{p(y_1^i|x^n, \hat{s}^n, s^n)}{p(y_1^{i-1}|x^n, \hat{s}^n, s^n)} = p(y_i|x^n, s^n, \hat{s}^n, y_1^{i-1}) \tag{32}$$

where the equalities in (31) follow from the memoryless channel model, and the first equality in (32) follows from the fact that the system generates authentic reconstructions so (1) holds. Thus, (32) implies the Markov relationship

$$Y_i \leftrightarrow (X_i, S_i) \leftrightarrow (X_1^i, X_{i+1}^n, S_1^i, S_{i+1}^n, Y_1^{i-1}, \hat{S}^n), \tag{33}$$

which by deleting selected terms from the right hand side yields (30). ■

Next, we combine these results to prove the converse part of Theorem 1 except for the cardinality bound on $\mathcal{U}$ which is derived immediately thereafter.

Lemma 3 *If a sequence of encoding and decoding functions $\Upsilon_n(\cdot)$ and $\Phi_n(\cdot)$ exist such that the decoder can generate authentic reconstructions achieving the distortion pair $(D_{\mathrm{e}}, D_{\mathrm{r}})$ when the reference channel is in effect then*

$$R^*(D_{\mathrm{e}}, D_{\mathrm{r}}) \geq 0 \tag{34}$$

Proof:

Define $D_{\mathrm{e},i}$ and $D_{\mathrm{r},i}$ as the component-wise distortions between S_i and X_i and between S_i and $\hat{S}_i$. We have

the following chain of inequalities:

$$R^*(D_{\rm e}, D_{\rm r}) = R^*\left(\frac{1}{n}\sum_{i=1}^{n} D_{{\rm e},i}, \frac{1}{n}\sum_{i=1}^{n} D_{{\rm r},i}\right) \tag{35}$$

$$\geq \frac{1}{n}\sum_{i=1}^{n} R^*(D_{{\rm e},i}, D_{{\rm r},i}) \tag{36}$$

$$\geq \frac{1}{n}\sum_{i=1}^{n}[I(U_i;Y_i) - I(U_i;S_i)] \tag{37}$$

$$\geq \frac{1}{n}\left[I(\hat{S}^n;Y^n) - I(\hat{S}^n;S^n)\right] \tag{38}$$

$$= \frac{1}{n}\left[H(\hat{S}^n|S^n) - H(\hat{S}^n|Y^n)\right] \tag{39}$$

$$\geq -\frac{1}{n}H(\hat{S}^n|Y^n) \tag{40}$$

$$\geq -\frac{1}{n} - \Pr[\Phi_n(Y^n) \neq \hat{S}^n]\log|\mathcal{S}| \tag{41}$$

The concavity of $R^*(D_{\rm e}, D_{\rm r})$ yields (36). To obtain (37), we combine Lemma 2 with (27). Next, to obtain (38), let $V = \hat{S}^n$ and $A_i = Y_i$ to apply Lemma 1 with U_i chosen according to (29). Fano's inequality yields (41).

Finally, using (in order) Bayes' law, (8), and (7), we obtain

$$\Pr[\Phi_n(Y^n) \neq \hat{S}^n] = \Pr[\mathcal{E}_{\rm sa}] + \Pr[\{\Phi_n(Y^n) \neq \hat{S}^n\} \cap \{\Phi_n(Y^n) = \varnothing\}] \tag{42}$$

$$\leq \Pr[\mathcal{E}_{\rm sa}] + \Pr[\{\Phi_n(Y^n) = \varnothing\}] \tag{43}$$

$$\leq \Pr[\mathcal{E}_{\rm sa}] + \Pr[\mathcal{E}_{D_{\rm r}}]. \tag{44}$$

Therefore exploiting that the system generates an authentic reconstruction ($\lim_{n\to\infty}\Pr[\mathcal{E}_{\rm sa}] = 0$) of the right distortion ($\lim_{n\to\infty}\Pr[\mathcal{E}_{D_{\rm r}}] = 0$) and that the alphabet of S is finite, we have that (41) and (44) imply (34). ■

The following proposition bounds the cardinality of $\mathcal{U}$.

Proposition 1 *Any point in the achievable distortion region defined by* (9) *can be attained with U distributed over an alphabet $\mathcal{U}$ of cardinality at most $(|\mathcal{S}| + |\mathcal{X}| + 3)\cdot|\mathcal{S}|\cdot|\mathcal{X}|$ with $p(x|u,s)$ singular or over an alphabet $\mathcal{U}$ of cardinality at most $|\mathcal{S}| + |\mathcal{X}| + 3$ if $p(x|u,s)$ is not required to be singular.*

Proof:

This can be proved using standard tools from convex set theory. Essentially, we define a convex set of continuous functions $f_j(\mathbf{p})$ where $\mathbf{p}$ represents a distribution of the form $\Pr(S = s, X = x|U = u)$ and the $f_j(\cdot)$ functions capture the features of the distributions relevant to (9). According to Carathéodory's Theorem [42, Theorem 14.3.4], [50], there exist j-max +1 distributions $\mathbf{p}_1$ through $\mathbf{p}_{j\text{-max}+1}$ such that any vector of function values, $(f_1(\mathbf{p}'), f_2(\mathbf{p}'), \ldots, f_{j\text{-max}}(\mathbf{p}'))$, achieved by some distribution $\mathbf{p}'$ can be achieved with a convex combination of the $\mathbf{p}_i$ distributions. Since each distribution corresponds to a particular choice for U, at most j-max + 1 possible values are required for U. Specifically, the desired cardinality bound for our problem can be proved by making the following syntactical modifications to the argument in [51, bottom left of p. 634]:

1. Replace $\Pr(X = x \mid U = u)$ with $\Pr(S = s, X = x \mid U = u)$ which is represented by the notation $\mathbf{p}$.
2. Choose
$$f_j(\mathbf{p}) = \sum_x \Pr(S = j, X = x \mid U = u) \tag{45}$$
for $j \in \{1, 2, \ldots, n\}$ where $n = |\mathcal{S}|$.
3. Choose
$$f_{n+1}(\mathbf{p}) = \sum_s \sum_x d_{\mathrm{e}}(x, s) \Pr(S = s, X = x \mid U = u). \tag{46}$$
4. Choose
$$f_{n+2}(\mathbf{p}) = \sum_s \sum_x d_{\mathrm{r}}(g(u), s) \Pr(S = s, X = x \mid U = u). \tag{47}$$
5. Choose
$$f_{n+3}(\mathbf{p}) = \sum_s \left(\sum_x \Pr(S = s, X = x \mid U = u) \right) \log \left(\sum_x \Pr(S = s, X = x \mid U = u) \right). \tag{48}$$
6. Choose
$$f_{n+4}(\mathbf{p}) = \sum_y \left(\sum_x \sum_s \Pr(Y = y \mid X = x) \Pr(S = s, X = x \mid U = u) \right) \cdot \log \left(\sum_x \sum_s \Pr(Y = y \mid X = x) \Pr(S = s, X = x \mid U = u) \right). \tag{49}$$
7. Choose
$$f_{n+5+j}(\mathbf{p}) = \sum_s \Pr(S = s, X = j \mid U = u) \tag{50}$$
for $j \in \{1, 2, \ldots, |\mathcal{X}|\}$.

Since the $f_j(\mathbf{p})$ determine $\Pr[S = s]$ (and therefore $H(S)$ as well), D_{e}, D_{r}, $H(S|U)$, $H(Y|U)$, and $\Pr[X = x]$ (and therefore $\Pr[Y = y]$ and $H(Y)$ also), they can be used to identify all points in the distortion region. According to [51, Lemma 3], for every point in this region obtained over the alphabet $\mathcal{U}$ there exists a U^* from alphabet $\mathcal{U}^*$ with cardinality $|\mathcal{U}^*|$ at most one greater than the dimension of the space spanned by the vectors f_i. The f_i corresponding to $\Pr[S = s]$ and $\Pr[X = x]$ contribute $|\mathcal{S}| - 1$ and $|\mathcal{X}| - 1$ dimensions while the other f_i contribute four more dimensions. Thus it suffices to choose $|\mathcal{U}^*| \leq |\mathcal{X}| + |\mathcal{S}| + 3$. Note that this cardinality bound applies to the general case where X is not necessarily a deterministic function of S and U^*.

By directly applying [35, Lemma 2] to each pair (u^*, s) in $\mathcal{U}^* \times \mathcal{S}$, we can split each u^* into $|\mathcal{X}|$ new symbols, u^{**} such that the mapping from (u^{**}, s) to x is deterministic. The new auxiliary random variable U^{**} takes values over the alphabet $\mathcal{U}^{**}$ where

$$|\mathcal{U}^{**}| = |\mathcal{U}^*| \cdot |\mathcal{S}| \cdot |\mathcal{X}| = (|\mathcal{X}| + |\mathcal{S}| + 3) \cdot |\mathcal{S}| \cdot |\mathcal{X}|. \tag{51}$$

Furthermore, this process does not change the distortion or violate the mutual information constraint. Thus a deterministic mapping from the source and auxiliary random variable to the channel input can be found with no loss of optimality provided a potentially larger alphabet is allowed for the auxiliary random variable. ■

We next apply Theorem 1 to two example scenarios of interest—one discrete and one continuous.

6 Example: the Binary-Hamming Scenario

Some applications of authentication are inherently discrete. For example, we might be interested in authenticating a passage of text, some of whose characters may have been altered in a benign manner through errors in optical character recognition process or error-prone human transcription during scanning, as well as by tampering.

As perhaps the simplest model representative of such discrete problems, we consider a symmetric binary source with a binary symmetric reference channel. Specifically, we model the source as an i.i.d. sequence where each S_i is a Bernoulli(1/2) random variable[7] and the reference channel output is $Y_i = X_i \oplus N_i$, where $\oplus$ denotes modulo-2 addition and where N^n is an i.i.d. sequence of Bernoulli(p) random variables. Finally, we adopt the Hamming distortion measure:

$$d(a,b) = \begin{cases} 0, & \text{if } a = b \\ 1, & \text{otherwise .} \end{cases}$$

For this problem, a suitable auxiliary random variable is

$$U = \{S \oplus (A \cdot T) \oplus [(1-A) \cdot V]\} + 2 \cdot (1-A), \tag{52}$$

where A, T, and V are Bernoulli α, τ, and ν random variables, respectively, and are independent of each other and S and N. Without loss of generality, the parameters τ and ν are restricted to the range $(0, 1/2)$. Note that $\mathcal{U} = \{0, 1, 2, 3\}$.

The encoder function $X = f(S, U)$ is, in turn, given by

$$X = \begin{cases} U, & \text{if } U \in \{0,1\} \\ S, & \text{if } U \in \{2,3\}, \end{cases} \tag{53}$$

from which it is straightforward to verify via (52) that the encoding distortion is

$$D_{\mathrm{e}} = \alpha\tau. \tag{54}$$

[7] We adopt the convention that all Bernoulli random variables take values in the set $\{0, 1\}$.

The corresponding decoder function $\hat{S} = g(U)$ takes the form

$$\hat{S} = U \bmod 2, \tag{55}$$

from which it is straightforward to verify via (52) that the reconstruction distortion is

$$D_{\mathrm{r}} = \alpha\tau + (1-\alpha)\nu. \tag{56}$$

In addition, $I(U;S)$ takes the form

$$\begin{aligned} I(U;S) &= H(S) - H(S|U) \\ &= H(S) - H(S,A|U) + H(A|U,S) \\ &= H(S) - H(S|U,A) - H(A|U) + H(A|U,S) \\ &= 1 - \alpha \cdot h(\tau) - (1-\alpha) \cdot h(\nu), \end{aligned} \tag{57}$$

where the second and third equalities follow from the entropy chain rule, where the last two terms on the third line are zero because knowing U determines A, and where the last equality follows from (52), with $h(\cdot)$ denoting the binary entropy function, i.e., $h(q) = -q\log q - (1-q)\log(1-q)$ for $0 \le q \le 1$. Similarly, $I(U;Y)$ takes the form

$$\begin{aligned} I(U;Y) &= H(Y) - H(Y|U) \\ &= H(Y) - H(Y,A|U) + H(A|U,Y) \\ &= H(Y) - H(Y|U,A) - H(A|U) + H(A|U,Y) \qquad (58) \\ &= 1 - \alpha\, h(p) - (1-\alpha) h\left(p(1-\nu) + (1-p)\nu\right). \qquad (59) \end{aligned}$$

For a fixed p, varying the parameters α, ν, and τ such that (59) is at least as big as (57) as required by (9a) generates the achievable distortion region shown in Fig. 4. Note from (59), (57), (54) and (56) that the boundary point $D_{\mathrm{e}} = D_{\mathrm{r}} = p$, in particular, is obtained by the parameter values $\alpha = 1$ and $\tau = p$ (with any choice of ν). Numerical optimization over all $p(u|s)$ and all (not necessarily singular) $p(x|s,u)$ with the alphabet size $|\mathcal{U}| = 7$ chosen in accordance with Proposition 1 confirms that Fig. 4 captures all achievable distortion pairs.

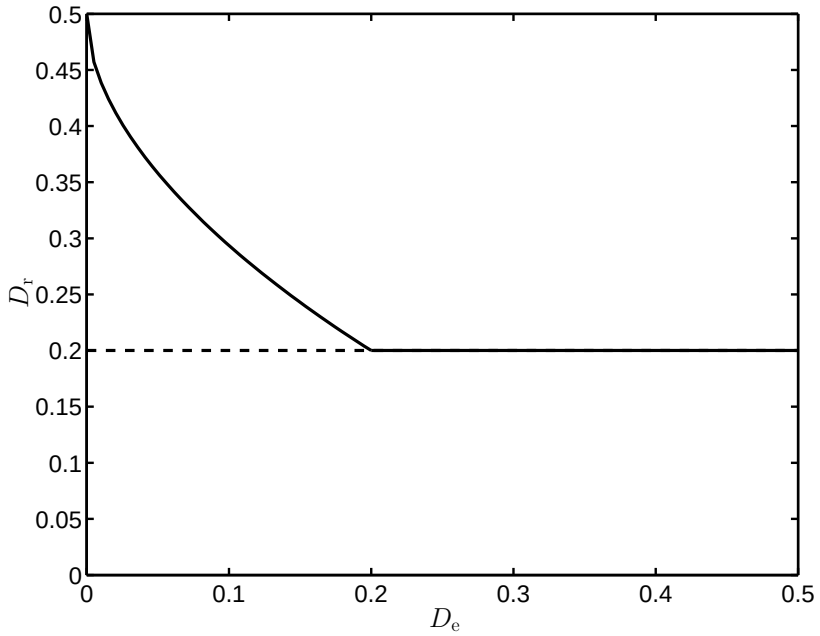

Figure 4: The solid line represents the frontier of the achievable distortion region for a binary symmetric source and a binary symmetric reference channel with cross-over probability $p = 0.2$. The dashed line represents the distortion region achievable when authentication is not required.

For comparison, we can also develop the achievable distortion region when authentication is not required. In this setting the goal is to provide a representation of the source which allows a decoder to obtain a good reconstruction from the reference channel output while keeping the encoding distortion small. Although in general hybrid analog-digital coding schemes can be used [35], uncoded transmission is optimal for the binary-Hamming case and thus all points in the region $D_\mathrm{e} \geq 0$ and $D_\mathrm{r} \geq p$ are achievable, as also shown in Fig. 4. Thus we see that the requirement that reconstructions be authentic strictly decreases the achievable distortion region as shown in Fig 4.

7 Example: the Gaussian-Quadratic Scenario

Some applications of authentication are inherently continuous. Examples involve sources such as imagery, video, or audio, that may encounter degradations and transformations as a result of routine handling including compression, transcoding, resampling, printing, and scanning, intentional editing and enhancements, as well as tampering attacks.

As perhaps the simplest model representative of such continuous problems, we consider a white

Gaussian source with a white Gaussian reference channel. Specifically, we model the source as an i.i.d. Gaussian sequence where each S_i has mean zero and variance σ_S^2, and the independent reference channel noise as an i.i.d. sequence whose ith element N_i has mean zero and variance σ_N^2. Furthermore, we adopt the quadratic distortion measure $d(a,b) = (a-b)^2$.

While our proofs in Section 5 exploited that our signals were drawn from finite alphabets and that all distortion measures were bounded to simplify our development, the results can be generalized to continuous-alphabet sources with unbounded distortion measures using standard methods. In the sequel, we assume without proof that the coding theorems hold for Gaussian sources with quadratic distortion. Since we have been unable to determine the optimal distribution for U in closed form,[8] we develop inner and outer bounds on the boundary of the achievable distortion region.

7.1 Unachievable Distortions: Inner Bounds

To derive an inner bound, we ignore the requirement that reconstructions be authentic, i.e., satisfy (1), and study the distortions possible in this case.

For a given transmit power P, it is well-known that the minimum possible reconstruction distortion D_{r} in the transmission of the source over the channel can be achieved without either source or channel coding in this Gaussian scenario, and the resulting distortion is

$$D_{\mathrm{r}} = \frac{\sigma_N^2 \sigma_S^2}{\sigma_N^2 + P}. \tag{60}$$

Moreover, for a scheme with encoding distortion D_{e}, the Cauchy-Schwarz inequality implies that the P is bounded according to

$$P = E[X^2] = E[(X-S+S)^2] = E[(X-S)^2] + E[S^2] + 2E[(X-S)S] \leq D_{\mathrm{e}} + \sigma_S^2 + 2\sqrt{D_{\mathrm{e}}\sigma_S^2}, \tag{61}$$

where equality holds if and only if $X = \left(1 + \sqrt{D_{\mathrm{e}}/\sigma_S^2}\right) S$. Thus, substituting (61) into (60) yields the inner bound

$$D_{\mathrm{r}} = \frac{\sigma_N^2 \sigma_S^2}{\sigma_N^2 + \left(\sqrt{D_{\mathrm{e}}} + \sigma_S\right)^2}. \tag{62}$$

[8] An analysis using the calculus of variations suggests that the optimal distribution is not even Gaussian.

7.2 Achievable Distortions: Outer Bounds

To derive outer bounds we will consider codebooks where (S, U, X) are jointly Gaussian. Since it is sufficient to consider X to be a deterministic function of U and S, the innovations form

$$T \sim N(0, \sigma_T^2), \quad E[TS] = 0 \tag{63a}$$

$$U = aS + cT \tag{63b}$$

$$X = bU + dT \tag{63c}$$

conveniently captures the desired relationships.[9] We examine two regimes: a low D_{e} regime in which we restrict our attention to the parameterization $(a, b, c, d) = (1, 1, 1/\alpha, 1)$, and a high D_{e} regime in which we restrict our attention to the parameterization $(a, b, c, d) = (1, \beta, 1, 0)$. As we will see, time-sharing between these parameterizations yields almost the entire achievable distortion region for Gaussian codebooks.

Low D_{e} Regime

We obtain an encoding that is asymptotically good at low D_{e} by using a distribution with structure similar to that used to achieve capacity in the related problem of information embedding [19]. In the language of [25], the encoding process involves distortion-compensation. In particular, the source is amplified by a factor $1/\alpha$, quantized to the nearest codeword, attenuated by α, and then a fraction of the resulting quantization error is added back to produce the final encoding, i.e.,

$$X^n = \alpha Q[S^n/\alpha] + (1-\alpha)(S^n - \alpha Q[S^n/\alpha]) \tag{64}$$

where $Q[\cdot]$ denotes the quantizer function.

With this encoding structure, it is convenient to make the assignment $U^n = \alpha Q[S^n/\alpha]$, so that we may write

$$U = S + T/\alpha \tag{65}$$

$$X = U + (1-\alpha)(S - U) = S + T \tag{66}$$

[9] It can be shown that choosing either $a = 1$ or $c = 1$ incurs no loss of generality.

where T is a Gaussian random variable with mean zero and variance σ_T^2 independent of both the source S and the channel noise N.

We choose $g(\cdot)$ to be the minimum mean-square estimate of S given U. Thus the resulting distortions are, via (65) and (66),

$$D_{\rm e} = E[(X-S)^2] = E[(S+T-S)^2] = \sigma_T^2 \tag{67}$$

and, in turn,

$$D_{\rm r} = E[S^2]\left(1 - \frac{E[SU]^2}{E[S^2]E[U^2]}\right) = \frac{\sigma_S^2(\sigma_T^2+\alpha^2\sigma_S^2) - \alpha^2\sigma_S^4}{\sigma_T^2+\alpha^2\sigma_S^2} = \frac{\sigma_S^2 D_{\rm e}}{D_{\rm e}+\alpha^2\sigma_S^2}. \tag{68}$$

To show that distortions (67) and (68) are achievable requires proving that (9a) holds. In [19], the associated difference of mutual informations is computed (using slightly different notation) as

$$I(U;Y) - I(S;U) = \frac{1}{2}\log\frac{\sigma_T^2(\sigma_T^2+\sigma_S^2+\sigma_N^2)}{\sigma_T^2\sigma_S^2(1-\alpha)^2 + \sigma_N^2(\sigma_T^2+\alpha^2\sigma_S^2)} \tag{69}$$

which implies that to keep the difference of mutual informations nonnegative we need

$$\sigma_T^2(\sigma_T^2+\sigma_S^2+\sigma_N^2) \geq \sigma_T^2\sigma_S^2(1-\alpha)^2 + \sigma_N^2(\sigma_T^2+\alpha^2\sigma_S^2). \tag{70}$$

Collecting terms in powers of α yields

$$\alpha^2(\sigma_T^2\sigma_S^2+\sigma_N^2\sigma_S^2) - 2\alpha\sigma_T^2\sigma_S^2 - \sigma_T^4 = (\alpha-r_+)(\alpha-r_-) \leq 0 \tag{71}$$

where

$$r_+ = \frac{1+\sqrt{1+\sigma_T^2/\sigma_S^2+\sigma_N^2/\sigma_S^2}}{1+\sigma_N^2/\sigma_T^2} \geq 0 \tag{72}$$

$$r_- = \frac{1-\sqrt{1+\sigma_T^2/\sigma_S^2+\sigma_N^2/\sigma_S^2}}{1+\sigma_N^2/\sigma_T^2} \leq 0. \tag{73}$$

Therefore to satisfy the mutual information constraint we need $r_- \leq \alpha \leq r_+$.

To minimize the distortions, (68) and (67) imply we want $|\alpha|$ as large as possible subject to the

constraint (71). Thus we choose $\alpha = r_+$, from which we see that

$$\frac{\alpha_{\text{auth}}}{\alpha_{\text{ie}}} = \left(1 + \sqrt{1 + \frac{\sigma_T^2 + \sigma_N^2}{\sigma_S^2}}\right), \tag{74}$$

where $\alpha_{\text{ie}} = \sigma_T^2/(\sigma_T^2 + \sigma_N^2)$ is the corresponding information embedding scaling parameter determined by Costa [19]. Evidently, the scaling parameter for the authentication problem is at least twice the scaling for information embedding and significantly larger when either the SNR σ_S^2/σ_N^2 or signal-to-(encoding)-distortion ratio (SDR) σ_S^2/σ_T^2 is small.

High D_{e} Regime

An encoder that essentially amplifies the quantization of the source to overcome the reference channel noise is asymptotically good at high D_{e}. A system with this structure corresponds to choosing the encoder random variables according to

$$U = S + T \tag{75}$$

$$X = \beta U. \tag{76}$$

In turn, choosing as $g(\cdot)$ the minimum mean-square error estimator of S given U yields the distortions

$$D_{\text{e}} = (1-\beta)^2\sigma_S^2 + \beta^2\sigma_T^2 \tag{77}$$

$$D_{\text{r}} = \frac{\sigma_S^2\sigma_T^2}{\sigma_S^2 + \sigma_T^2}. \tag{78}$$

It remains only to determine β. Since

$$I(U;S) = \frac{1}{2}\log\frac{\sigma_S^2 + \sigma_T^2}{\sigma_T^2} \tag{79}$$

and

$$I(U;Y) = \frac{1}{2}\log\frac{\beta^2(\sigma_S^2 + \sigma_T^2) + \sigma_N^2}{\sigma_N^2}, \tag{80}$$

the mutual information constraint (9a) implies that

$$\beta \geq \sqrt{\frac{\sigma_S^2 \sigma_N^2}{\sigma_T^2(\sigma_S^2 + \sigma_T^2)}} \tag{81}$$

7.3 Comparing and Interpreting The Bounds

Using (68) with α given by (72) and varying σ_T^2 yields one outer bound. Using (77) and (78) with (81) and again varying σ_T^2 yields the other outer bound. The lower convex envelope of this pair of outer bounds is depicted in Fig. 5 at different SNR's. To see that the first and second outer bounds are asymptotically the best achievable for low and high $D_{\rm e}$, respectively, we superimpose on these figures the best Gaussian codebook performance, as obtained by numerically optimizing the parameters in (63).

By using (62), (68), and (78), it is possible to show that for any fixed $D_{\rm e} \geq \sigma_N^2$ the inner and outer bounds converge asymptotically in SNR in the sense that

$$\lim_{\rm SNR \to \infty} \frac{D_{r,\rm outer}}{D_{r,\rm inner}} = 1$$

where $D_{r,\rm inner}$ and $D_{r,\rm outer}$ represent the inner and outer bounds corresponding to the fixed value of $D_{\rm e}$. Thus, in this high SNR regime, Gaussian codebooks are optimal, and (62) accurately characterizes their performance as reflected in Fig. 5.

The figure also indicates (and it is possible to prove) that for any fixed SNR, the inner and outer bounds converge asymptotically in $D_{\rm e}$ in the sense that

$$\lim_{D_{\rm e} \to \infty} \frac{D_{r,\rm outer}(D_{\rm e})}{D_{r,\rm inner}(D_{\rm e})} = 1$$

where $D_{r,\rm inner}(D_{\rm e})$ and $D_{r,\rm outer}(D_{\rm e})$ represent the inner and outer bounds as a function of the encoding distortion $D_{\rm e}$. Evidently in this high encoding distortion regime, $D_{\rm r}/\sigma_N^2$ can be made arbitrarily small by using Gaussian codebooks and making $D_{\rm e}/\sigma_N^2$ sufficiently large. While this implies that, in principle, there is no fundamental limit to how small we can make $D_{\rm r}$ by increasing $D_{\rm e}$ through amplification of the source, in practice secondary effects not included in the model such as saturation or clipping will provide an effective limit.

Finally, note that the cost of providing authentication is readily apparent since the inner bound

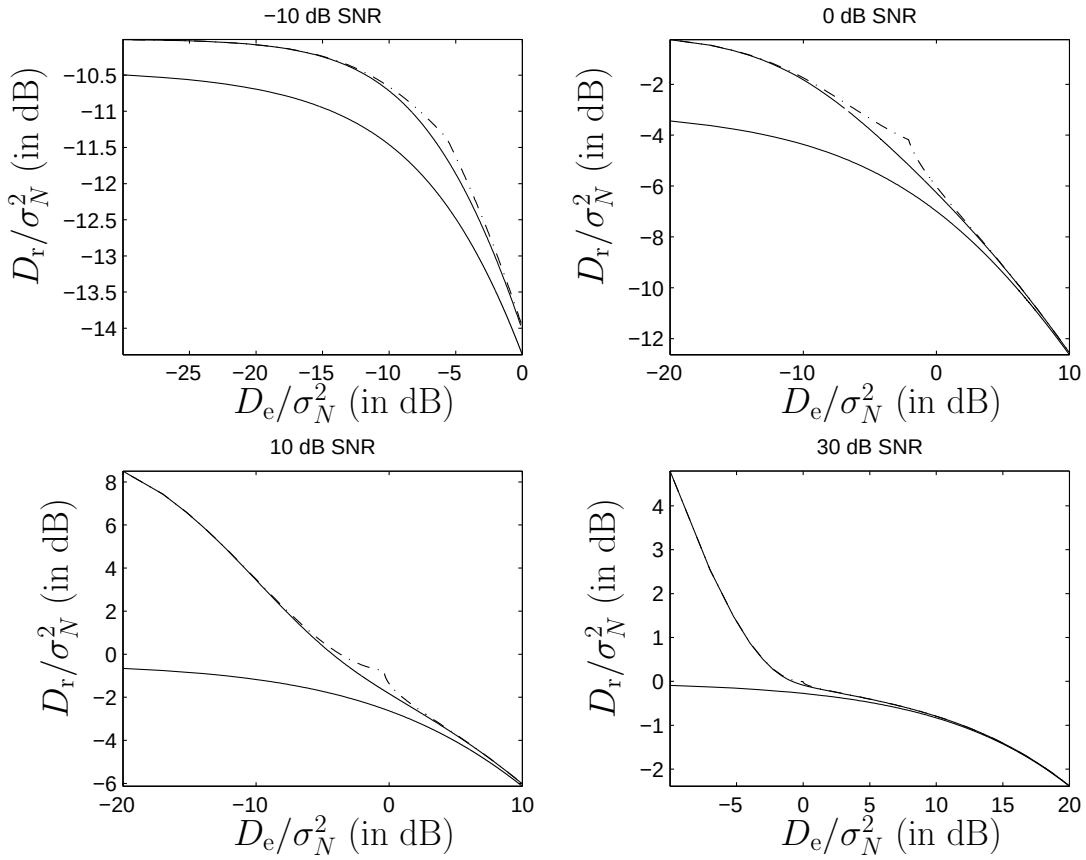

Figure 5: Bounds on the achievable distortion region for the Gaussian-quadratic problem. The lowest solid line is the inner bound, which corresponds to the boundary of the achievable region when reconstructions need not be authentic. The numerically obtained upper solid line is the outer bound resulting from the use of Gaussian codebooks. The dashed curve corresponds to the lower convex envelope of the low and high $D_{\rm e}$ outer bounds derived in the text.

from (62) represents the distortions achievable when the reconstruction need not be authentic. Since for a fixed SNR, the bounds converge asymptotically for large D_e, and for a fixed $D_e \geq \sigma_N^2$ the bounds converge asymptotically for large SNR, we conclude that the price of authentication is negligible in these regimes. However, for low D_{e} regimes of operation, requiring authenticity strictly reduces the achievable distortion region. This behavior is analogous to that observed in the binary-Hamming case.

8 Comparing Authentication Architectures

The most commonly studied architectures for authentication are robust watermarking (sometimes referred to as self-embedding) and fragile watermarking. In the sequel we compare these architectures to that developed in this paper.

8.1 Quantize-and-Embed Authentication Systems

A variety of researchers have considered a quantize-and-embed strategy, sometimes termed "robust watermarking" or "self-embedding," for authentication problems with distortion criteria [14] [9] [10] [3] [15]. The idea is to encode as follows. First, the source S^n is quantized to a representation in terms of bits using a source coding (compression) algorithm. Second the bits are protected using a cryptographic technique such as a digital signature or hash function. Finally, the protected bits are embedded into the original source using an information embedding (digital watermarking) algorithm. At the decoder, the embedded bits are extracted. If their authenticity is verified via the appropriate cryptographic technique, a reconstruction of the source is produced from the bits. Otherwise, the decoder declares that an authentic reconstruction is not possible.

It is straightforward to develop the information-theoretic limits of such approaches, and to compare the results to the optimum systems developed in the preceding sections. In particular, if we use optimum source coding and information embedding in the quantize-and-embed approach, it follows that, in contrast to Theorem 1, the distortion pair $(D_{\mathrm{e}}, D_{\mathrm{r}})$ lies in the achievable distortion region for a quantize-and-embed structured solution to the problem (2) if and only if there exists

distributions $p(\hat{s}|s)$ and $p(u|s)$, and a function $f(\cdot,\cdot)$, such that

$$I(U;Y) - I(S;U) \geq I(S;\hat{S}) \tag{82a}$$

$$E[d_{\mathrm{e}}(S, f(U,S))] \leq D_{\mathrm{e}} \tag{82b}$$

$$E[d_{\mathrm{r}}(S,\hat{S})] \leq D_{\mathrm{r}}. \tag{82c}$$

These results follow from the characterization of the rate-distortion function of a source [42] and the capacity of information embedding systems with distortion constraints as developed in [35] as an extension of [18].

To appreciate the reduction in the achievable region, we consider our two example scenarios.

8.1.1 Example: Binary-Hamming Case

In this scenario, the rate-distortion function is [42]

$$R(D_{\mathrm{r}}) = 1 - h(D_{\mathrm{r}}), \tag{83}$$

while the information embedding capacity is (see [35]) the upper concave envelope of the function

$$g_p(D_{\mathrm{e}}) = \begin{cases} 0, & \text{if } 0 \leq d < p, \\ h(D_{\mathrm{e}}) - h(p), & \text{if } p \leq D_{\mathrm{e}} \leq 1/2, \end{cases} \tag{84}$$

i.e.,

$$C(D_{\mathrm{e}}) = \begin{cases} \dfrac{g_p(D_p)}{D_p} D_{\mathrm{e}}, & \text{if } 0 \leq D_{\mathrm{e}} \leq D_p, \\ g_p(D_{\mathrm{e}}), & \text{if } D_p < D_{\mathrm{e}} \leq 1/2, \end{cases} \tag{85}$$

where $D_p = 1 - 2^{-h(p)}$. Equating R in (83) to C in (85), we obtain a relation between D_{r} and D_{e}. This curve is depicted in Fig. 6 for different reference channel parameters. As this figure reflects, the optimum quantize-and-embed system performance lies strictly inside the achievable region for the binary-Hamming scenario developed in Section 6, with the performance gap largest for the cleanest reference channels.

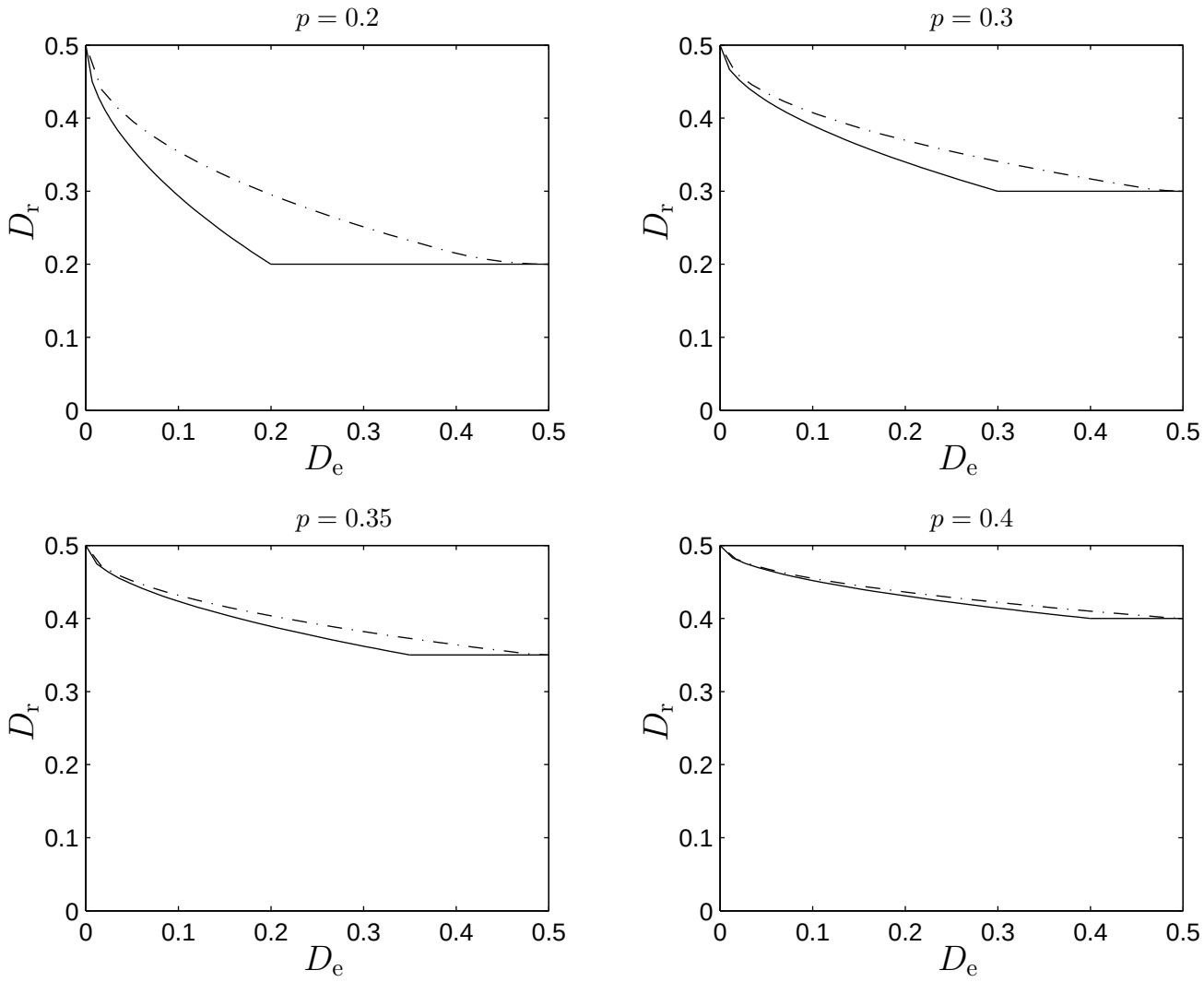

Figure 6: Performance loss of quantize-and-embed systems for the Binary-Hamming scenario with reference channel crossover probability p. The solid curve depicts the boundary of the achievable regions for the optimum system; the dashed curve depicts that of the best quantize-and-embed system.

8.1.2 Example: Gaussian-Quadratic Case

In this scenario, the rate-distortion function is [42]

$$R(D_{\mathrm{r}}) = \begin{cases} \frac{1}{2}\log\frac{\sigma_S^2}{D_{\mathrm{r}}}, & 0 \le D_{\mathrm{r}} \le \sigma_S^2 \\ 0, & D_{\mathrm{r}} > \sigma_S^2, \end{cases} \tag{86}$$

while the information embedding capacity is [19]

$$C(D_{\mathrm{e}}) = \frac{1}{2}\log\left(1 + \frac{D_{\mathrm{e}}}{\sigma_N^2}\right). \tag{87}$$

Again, equating R in (86) to C in (87), we obtain the following relation between D_{r} and D_{e} for all $D_{\mathrm{e}} > 0$:

$$D_{\mathrm{r}} = \frac{\sigma_S^2}{(1 + D_{\mathrm{e}}/\sigma_N^2)}. \tag{88}$$

This curve is depicted in Fig. 7 for different reference channel SNRs. This figure reflects that the optimum quantize-and-embed system performance lies strictly inside the achievable region for the Gaussian-quadratic scenario developed in Section 7. Likewise, the performance gap is largest for the highest SNR reference channels. Indeed, comparing the inner bound (62) on the performance of the optimum system with that of quantize-and-embed, i.e., (88), we see that while quantize-and-embed incurs no loss at low SNR:

$$\frac{D_{\mathrm{r}}^{\mathrm{qe}}}{D_{\mathrm{r}}} \to 1 \quad \text{as} \quad \frac{\sigma_S^2}{\sigma_N^2} \to 0, \tag{89}$$

at high SNR the loss is as much as SNR/2 for $D_{\mathrm{e}} \ge \sigma_N^2$:

$$\frac{\sigma_N^2}{\sigma_S^2}\frac{D_{\mathrm{r}}^{\mathrm{qe}}}{D_{\mathrm{r}}} \to \frac{1}{1 + D_{\mathrm{e}}/\sigma_N^2} \le \frac{1}{2} \quad \text{as} \quad \frac{\sigma_S^2}{\sigma_N^2} \to \infty, \tag{90}$$

where we have used $D_{\mathrm{r}}^{\mathrm{qe}}$ to denote the quantize-and-embed reconstruction distortion (88).

Disadvantages of Quantize-and-Embed: The main disadvantage of quantize-and-embed systems as characterized by (82) seem to be that at high SNR and low encoding distortion, only a few bits representing the original signal can be embedded in the source. Thus the resulting reconstruction distortion is much higher than a system employing a joint source–channel–authentication

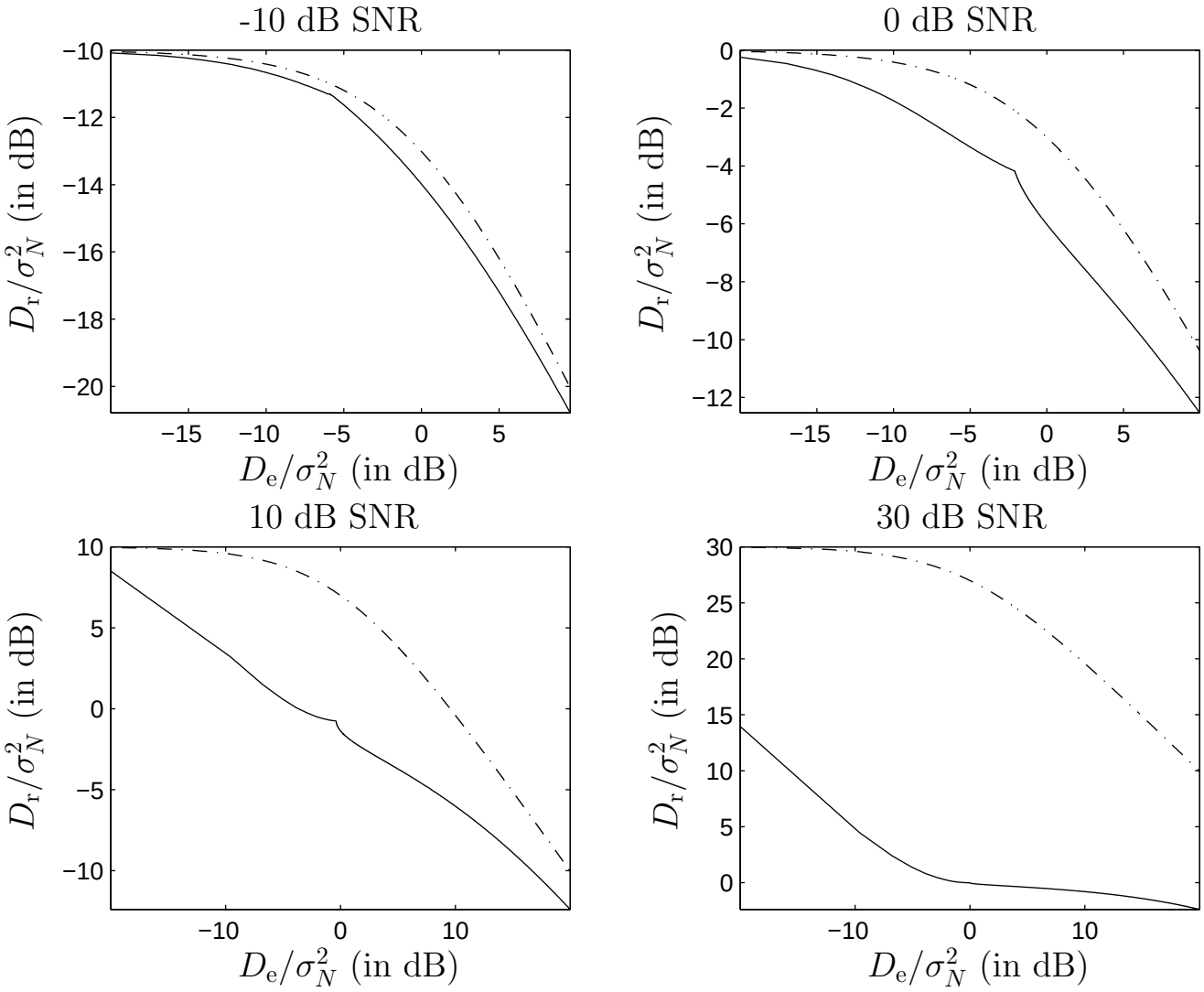

Figure 7: Performance loss of quantize-and-embed systems for the Gaussian-quadratic scenario at various reference channel SNRs. The solid curve depicts the Gaussian codebook inner bound of the achievable regions for the optimum system; the dashed curve depicts that of the best quantize-and-embed system.

coding strategy based on the principles in Section 5.

8.2 Fragile Watermarking Authentication Systems

Another popular authentication architecture is based on the idea of fragile watermarking [6] [13] [4] [12]. A watermark message, M, known only to the encoder and decoder (and kept secret from the attacker) is embedded into the source signal by the encoder. The decoder attempts to extract a watermark, $\hat{M}$, from the received signal. If $\hat{M}$ and M are close enough (*e.g.*, the distortion between $\hat{M}$ and M using a scalar distortion measures is below some threshold), then the received signal is declared authentic. Developing a complete information theoretic characterization of such schemes is beyond the scope of this paper. Instead we point out that while fragile watermarking systems may achieve significantly lower encoding and reconstruction distortions when the reference channel is in effect, they also achieve significantly lower security.

Specifically, imagine that the received signal in a fragile watermarking system using encoding distortion D_{e} is declared authentic. Various signal restoration algorithms (e.g., linear filtering or techniques from the theory of coding remote sources [52]) can be use to enhance the received signal and we denote the reconstruction as $g(Y^n, M)$. Let $\mathcal{R}$ denote the set of all possible values for $g(Y^n, M)$. If $|\mathcal{R}| = 1$ then the security requirement (1) will be satisfied and the attacker can not influence the reconstruction. In this case, the reconstruction distortion will be at least as large as in Theorem 1. As $|\mathcal{R}|$ increases, the reconstruction can more closely follow the channel output and so (when the reference channel is in effect) the reconstruction distortion decreases, but the amount of tampering possible by a malicious attacker increases. Hence while fragile watermarking systems may allow a lower reconstruction distortion they provide a weaker type of security.

Essentially, arguments in favor of this type of security assume that if the changes in a signal (as measured by a scalar distortion measure) are numerically insignificant they are also semantically insignificant. Thus, even though an attacker may have tampered with a signal which is declared authentic, the tampering is immaterial. A major drawback to this view is that current methods of measuring distortion are extremely crude. Hence, even if a signal differs only slightly according to a crude distortion measure, the meaning may be drastically altered. Conversely, routine degradations such as re-sampling or compression which are semantically insignificant may introduce a large numerical distortion. While more accurate distortion measures may mitigate these problems, there

is always the concern that a malicious adversary will fool the system by finding a flaw in the distortion measure.[10]

9 Layered Authentication: Broadcast Reference Channels

More elaborate authentication systems provide multiple layers of security, and arise naturally out of the use of broadcast reference channel models. For such layered authentication systems, we might reasonably seek the following characteristics, which generalize those developed earlier. Starting from a channel input that is some fixed encoding of the source that incurs some distortion, for channel outputs that are consistent with any of a fixed set of reference channels, a decoder produces an authentic reconstruction of some corresponding fidelity. Otherwise, the decoder declares that an authentic reconstruction is not possible.

For the purpose of illustration, we focus on the two-receiver memoryless degraded broadcast channel [42] as our reference channel. For convenience, we refer to the strong channel as the "fine" one, and the weak channel, which is a degraded version of the strong one, as the "coarse" one. In this case, for any prescribed level of encoding distortion D_{e}, there is a trade-off between the qualities $D_{\mathrm{r}}^{\mathrm{c}}$ and $D_{\mathrm{r}}^{\mathrm{f}}$ of authentic reconstructions that can be achieved by decoders whose inputs are consistent with the coarse and fine channels, respectively. In general, achieving smaller values of $D_{\mathrm{r}}^{\mathrm{c}}$ requires accepting larger values of $D_{\mathrm{r}}^{\mathrm{f}}$ and vice-versa. Using the ideas of this paper, one can explore the fundamental nature of such trade-offs.

9.1 Achievable Distortion Regions

The scenario of interest is depicted in Fig. 8. As a natural generalization of its definition in the single-layer context (2), an instance of the layered authentication problem consists of the tuple

$$\{\mathcal{S}, p(s), \mathcal{X}, \mathcal{Y}, p(y_{\mathrm{c}}|y_{\mathrm{f}}), p(y_{\mathrm{f}}|x), d_{\mathrm{e}}(\cdot,\cdot), d_{\mathrm{r}}(\cdot,\cdot)\}\,, \tag{91}$$

[10] Inaccurate distortion measures are also a problem in measuring fidelity in other settings such as compression, denoising, and signal enhancement. For these applications, it seems unlikely that Nature will craft distortions specially tailored to exploit the shortcomings of a particular distortion measure. But this is exactly what a malicious adversary will do. Thus inaccurate distortion measures are particularly dangerous for authentication even though they may be acceptable for other applications.

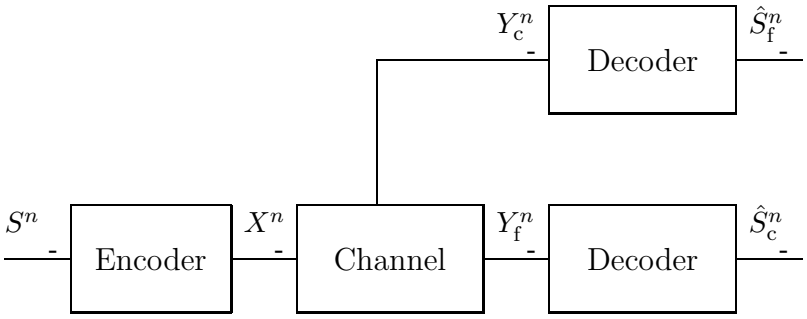

Figure 8: Two-layered authentication system operation when the reference channel is in effect. From the coarse and fine outputs of the degraded broadcast reference channel, the authentic reconstructions $\hat{S}_{\rm c}^n$ and $\hat{S}_{\rm f}^n$ are produced. The common encoding obtained from the source S^n is X^n.

where, since our reference channel is a degraded broadcast channel, the reference channel law takes the form

$$p(y_{\rm c}^n, y_{\rm f}^n|x^n) = p(y_{\rm c}^n|y_{\rm f}^n)\, p(y_{\rm f}^n|x^n). \tag{92}$$

Let $\hat{S}_{\rm c}^n$ denote the authentic reconstruction when decoder input is consistent with the coarse output of the reference channel, and let $\hat{S}_{\rm f}^n$ denote the authentic reconstruction when decoder input is consistent with the fine output of the reference channel. In turn, the corresponding two reconstruction distortions are defined according to

$$D_{\rm r}^{\rm c} = \frac{1}{n}\sum_{i=1}^{n} d_{\rm r}(S^n, \hat{S}_{\rm c}^n) \qquad \text{and} \qquad D_{\rm r}^{\rm f} = \frac{1}{n}\sum_{i=1}^{n} d_{\rm r}(S^n, \hat{S}_{\rm f}^n). \tag{93}$$

The following theorem develops trade-offs between the encoding distortion $D_{\rm e}$, and the two reconstruction distortions (93) that are achievable.

Theorem 2 *The distortion triple* $(D_{\rm e}, D_{\rm r}^{\rm c}, D_{\rm r}^{\rm f})$ *lies in the achievable distortion region for the layered authentication problem* (91) *if there exist distributions* $p(u,t|s)$ *and* $p(x|u,t,s)$*, and functions* $g_{\rm c}(\cdot)$

and $g_{\mathrm{f}}(\cdot,\cdot)$ *such that*

$$I(U;Y_{\mathrm{c}}) - I(S;U) \geq 0 \tag{94a}$$

$$I(T;Y_{\mathrm{f}}|U) - I(S;T|U) \geq 0 \tag{94b}$$

$$E[d_{\mathrm{e}}(S,X)] \leq D_{\mathrm{e}} \tag{94c}$$

$$E[d_{\mathrm{r}}(S,g_{\mathrm{c}}(U))] \leq D_{\mathrm{r}}^{\mathrm{c}}. \tag{94d}$$

$$E[d_{\mathrm{r}}(S,g_{\mathrm{f}}(U,T))] \leq D_{\mathrm{r}}^{\mathrm{f}}. \tag{94e}$$

In this theorem, the achievable distortion region is defined in a manner that is the natural generalization of that for single-layer systems, i.e., Definition 2.

In the interests of brevity and since it closely parallels that for the single-layer case, we avoid a formal proof of this result. Instead, we sketch the key ideas of the construction.

Sketch of Proof:

First a codebook $\mathcal{C}_{\mathrm{c}}$ is created for the coarse layer at rate $R_{\mathrm{c}} = I(U;S) + 2\gamma$ where only $2^{n(R_{\mathrm{c}}+\gamma)}$ codewords are marked as admissible as in Theorem 1. Then for each codeword, $c_{\mathrm{c}} \in \mathcal{C}_{\mathrm{c}}$, an additional random codebook, $\mathcal{C}_{\mathrm{f}}(c_{\mathrm{c}})$ of rate $R_{\mathrm{f}} = I(T;S|U) + 2\gamma$ is created according to the marginal distribution $p(t|u)$ where only $2^{n(R_{\mathrm{f}}+\gamma)}$ codewords are marked as admissible.

The encoder first searches $\mathcal{C}_{\mathrm{c}}$ for an admissible codeword c_{c} jointly typical with the source and then searches $\mathcal{C}_{\mathrm{f}}(c_{\mathrm{c}})$ for a refinement c_{f} that is jointly typical with the source. The pair $(c_{\mathrm{c}}, c_{\mathrm{f}})$ is then mapped into the channel according to $p(x|u,t,s)$. By standard arguments the encoding will succeed with high probability provided that $R_{\mathrm{c}} > I(U;S)$ and $R_{\mathrm{f}} > I(T;S|U)$.

When the received signal is consistent with either output of the reference channel, the decoder locates an admissible codeword $\hat{c}_{\mathrm{c}} \in \mathcal{C}_{\mathrm{c}}$ jointly typical with the received signal. If the received signal is consistent with the coarse output of the reference channel, in particular, the decoder then produces the coarse authentic reconstruction $\hat{S}_{\mathrm{c}}^n = g_{\mathrm{c}}(\hat{c}_{\mathrm{c}})$. However, if the received signal is consistent with the fine output of the reference channel, the decoder then proceeds to locate an admissible $\hat{c}_{\mathrm{f}} \in \mathcal{C}_{\mathrm{f}}(\hat{c}_{\mathrm{c}})$ and produces the fine authentic reconstruction $\hat{S}_{\mathrm{f}}^n = g_{\mathrm{f}}(\hat{c}_{\mathrm{c}}, \hat{c}_{\mathrm{f}})$.

By arguments similar to those used in the single-layer case (i.e., proof of Theorem 1), this strategy achieves vanishingly small probabilities of successful attack, and when the reference channel is in effect meets the distortion targets provided that $R_{\mathrm{c}} < I(U;Y_{\mathrm{c}})$ and $R_{\mathrm{f}} < I(T;Y_{\mathrm{f}}|U)$.

■

9.2 Example: Gaussian-Quadratic Case

The Gaussian-quadratic case corresponds to the fine and coarse outputs of the reference channel taking the forms $Y_{\mathrm{f}} = X + N$ and $Y_{\mathrm{c}} = Y_{\mathrm{f}} + V$, respectively, where N and V are Gaussian random variables independent of each other, as well as S and X.

For this case, a natural approach to the layered authentication system design has the structure depicted in Figure 9, which generalizes that of the single-layer systems developed in Section 7. The encoder determines the codeword T^n nearest the source S^n, then perturbs T^n so as to reduce the encoding distortion, producing the transmitted vector X^n. If the channel output stays within the darkly shaded sphere centered about T^n, e.g., producing Y_{f}^n as shown, the decoder produces a fine-grain authentic reconstruction from T^n. If the channel output is outside the darkly shaded sphere, but inside the encompassing lightly shaded sphere centered about U^n, e.g., producing Y_{c}^n as shown, the decoder produces a coarse-grain authentic reconstruction from U^n. If the channel output is outside any shaded region, e.g., producing Z^n, the decoder indicates that an authentic reconstruction is not possible.

To illustrate a possible achievable distortion region for the layered authentication scenario, we extend the Gaussian-Quadratic example from Section 7.

To develop the performance of such a system, we apply Theorem 2. In particular, we choose the auxiliary random variables according to

$$U = S + A/\alpha \tag{95}$$

$$T = S + B/\beta \tag{96}$$

$$X = S + A + B. \tag{97}$$

where A and B are Gaussian random variables independent of S. Choosing $g_{\mathrm{c}}(\cdot)$ and $g_{\mathrm{f}}(\cdot,\cdot)$ to be the minimum mean-square error estimates of S from U and (U,T), respectively, yields

$$D_{\mathrm{e}} = \sigma_A^2 + \sigma_B^2 \tag{98}$$

$$D_{\mathrm{r}}^{\mathrm{c}} = \sigma_S^2\left(1 - \frac{E[SU]^2}{E[S^2]E[U^2]}\right) = \frac{\sigma_S^2\sigma_A^2}{\sigma_A^2 + \alpha^2\sigma_S^2} \tag{99}$$

$$D_{\mathrm{r}}^{\mathrm{f}} = \sigma_S^2 - \Lambda_{S,[UT]}\Lambda_{[UT]}^{-1}\Lambda_{[UT],S} = \frac{\sigma_S^2\sigma_A^2\sigma_B^2}{\beta^2\sigma_S^2\sigma_A^2 + \sigma_A^2\sigma_B^2 + \alpha^2\sigma_S^2\sigma_B^2}, \tag{100}$$

where Λ with a single subscript denotes the covariance of its argument, and Λ with a subscript pair denotes the cross-covariance between its arguments.

To produce $\hat{S}_{\mathrm{c}}^n$, a decoder essentially views B as additive channel noise. Therefore, we can

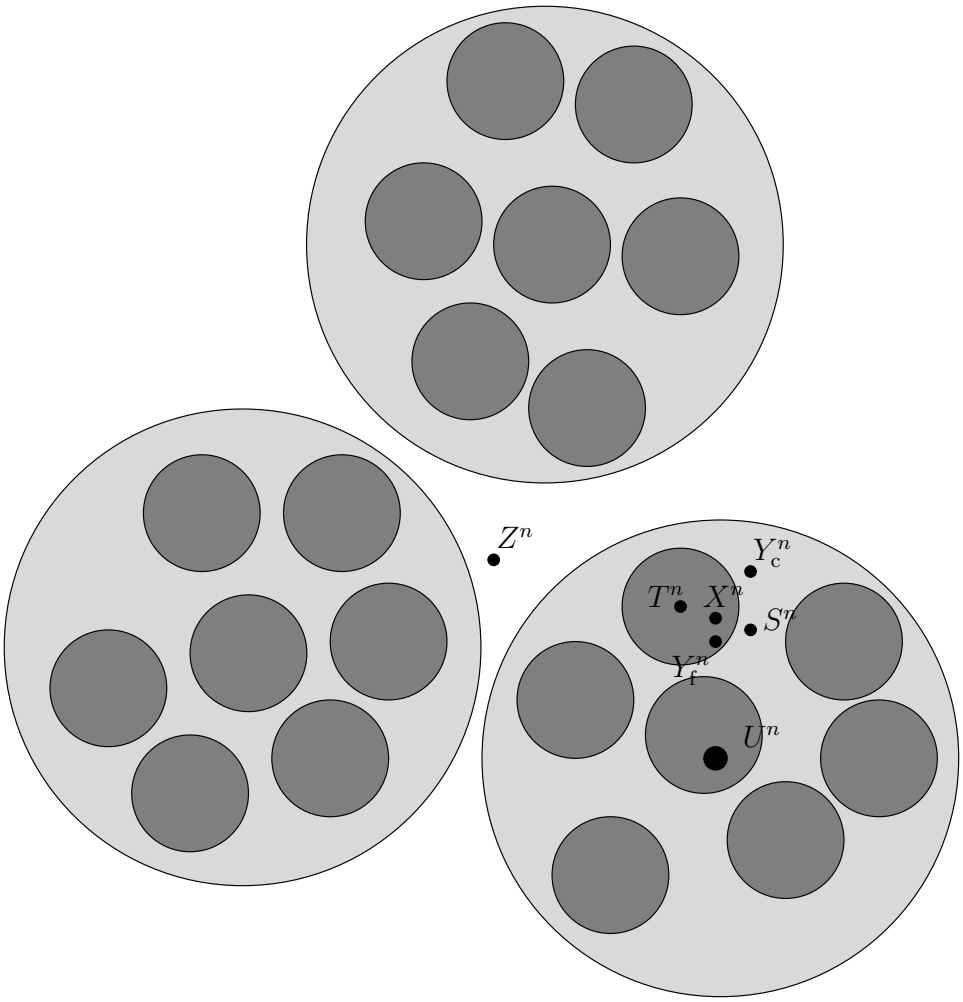

Figure 9: Illustration of geometry for a two-layer authentication system for the Gaussian-quadratic scenario.

immediately apply the arguments from Section 7.2 to obtain

$$I(U;Y_{\rm c}) - I(S;U) = \frac{1}{2}\log\frac{\sigma_A^2(\sigma_A^2+\sigma_S^2+\sigma_N^2+\sigma_V^2+\sigma_B^2)}{\sigma_A^2\sigma_S^2(1-\alpha)^2+(\sigma_N^2+\sigma_V^2+\sigma_B^2)(\sigma_A^2+\alpha^2\sigma_S^2)}. \tag{101}$$

From this we can solve for α as in the single-layer case of Section 7.2 by simply replacing σ_T^2 and σ_N^2 with σ_A^2 and $\sigma_N^2+\sigma_V^2+\sigma_B^2$, respectively, in (72).

Finally, since

$$I(S;T|U) - I(T;Y_{\rm f}|U) = H(T|U,Y_{\rm f}) - H(T|U,S) = H(T,U,Y_{\rm f}) - H(U,Y_{\rm f}) - H(T,U,S) + H(U,S).$$

we see that (94b) implies

$$\frac{\det(\Lambda_{[TUY_{\rm f}]})}{\det(\Lambda_{[UY_{\rm f}]})} \le \frac{\det(\Lambda_{[TUS]})}{\det(\Lambda_{[US]})}. \tag{102}$$

By varying σ_A^2, σ_B^2 and β such that (102) is satisfied we can trace out the volume of an achievable distortion region. Fig. 10 shows slices of this three dimensional region by plotting the fine and coarse reconstruction distortions $D_{\rm r}^{\rm f}$ and $D_{\rm r}^{\rm c}$ for various values of the encoding distortion $D_{\rm e}$. Note that it follows from our single-layer inner bounds that for a particular choice of encoding distortion $D_{\rm e}$, the achievable trade-offs between $D_{\rm r}^{\rm c}$ and $D_{\rm r}^{\rm f}$ are contained within the region

$$D_{\rm r}^{\rm c} \ge \frac{\sigma_S^2(\sigma_N^2+\sigma_V^2)}{\sigma_N^2+\sigma_V^2+\left(\sqrt{D_{\rm e}}+\sigma_S\right)^2} \tag{103}$$

$$D_{\rm r}^{\rm f} \ge \frac{\sigma_S^2\sigma_N^2}{\sigma_N^2+\left(\sqrt{D_{\rm e}}+\sigma_S\right)^2}, \tag{104}$$

where obviously the lower bound of (104) is smaller than that of (103).

A simple alternative to the layering system for such authentication problems is time-sharing, whereby some fraction of time the encoder uses a codebook appropriate for the coarse reference channel, and for the remaining time uses a codebook appropriate for the fine reference channel. When the coarse reference channel is in effect, the decoder produces the coarse authentic reconstruction for the fraction of time the corresponding codebook is in effect and produces zero the rest of the time. When the fine reference channel is in effect, the decoder produces the fine authentic reconstruction during the fraction of time the corresponding codebook is in effect, and produces the coarse reconstruction for the remaining time (since the broadcast channel is a degraded one).

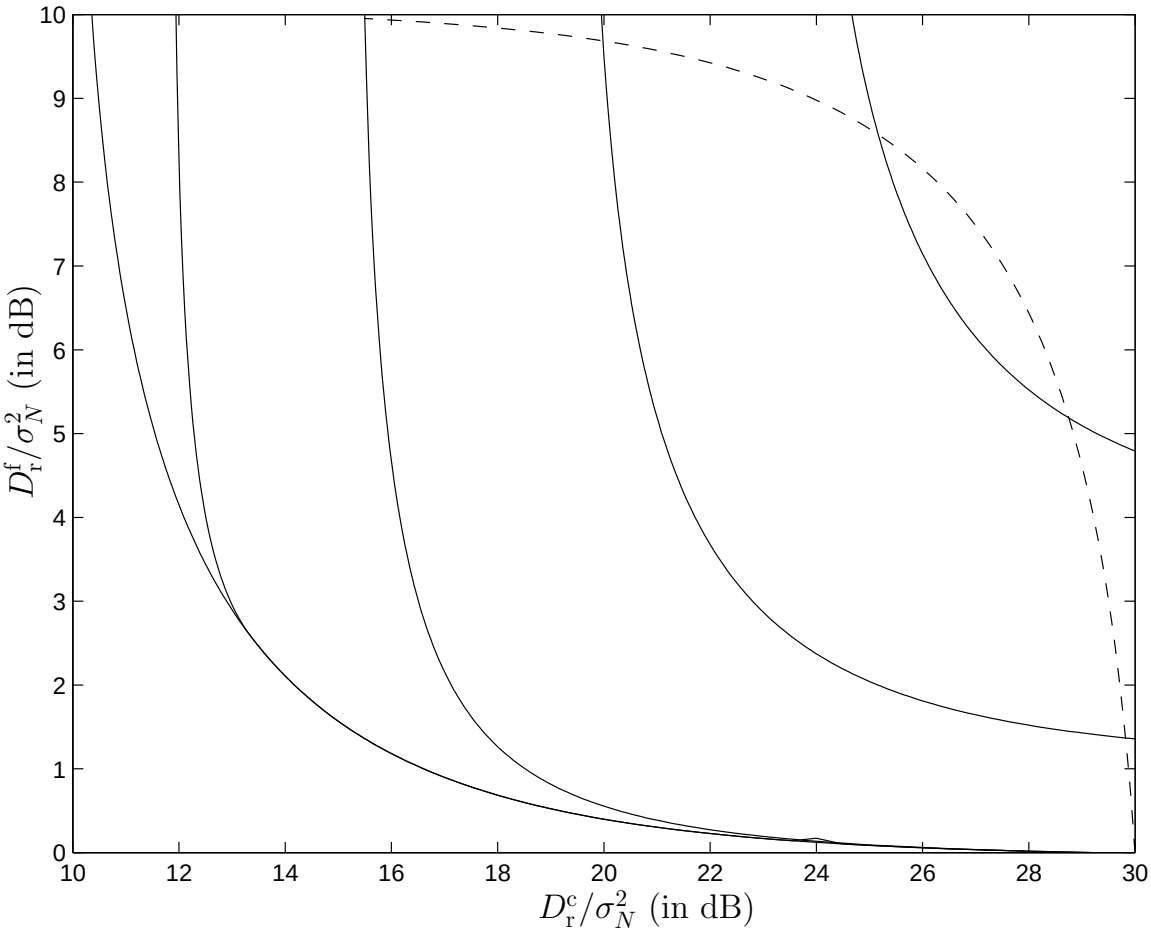

Figure 10: Achievable reconstruction distortion pairs in a layered authentication system for the Gaussian-quadratic case with $\sigma_S^2/\sigma_N^2 = 30$ dB, $\sigma_N^2 = 1$ and $\sigma_V^2/\sigma_N^2 = 10$ dB. From left to right the curves correspond to encoding distortions of $D_{\rm e}/\sigma_N^2 = 10, 5, 0, -5, -10$ dB. Using a layered framework, authentic reconstructions with distortions $D_{\rm r}^{\rm c}$, and $D_{\rm r}^{\rm f}$ can be obtained depending on the severity of the perturbations encountered. Increasing $D_{\rm e}$ (the distortion caused by the encoder) increases the range of trade-offs possible between $D_{\rm r}^{\rm c}$ and $D_{\rm r}^{\rm f}$. The dashed curve corresponds time-sharing between two points on the $D_{\rm e}/\sigma_N^2 = 0$ dB curve.

However, as Fig. 10 also illustrates, this approach is in general particularly inefficient: the use of such time-sharing results in a substantially smaller achievable region.

10 Concluding Remarks

While this paper explores suitable architectures for authentication systems, many aspects of the detailed design and implementation of such systems remain to be addressed. As the wide array of proposed systems demonstrates, building authentication systems requires tools from a variety of fields including information theory, communication theory, signal processing and cryptography.

On the information theory side, analyzing systems without the i.i.d. assumption in the source (e.g. image models with correlated pixel values) or the channel (e.g. blurs, smudges, or other localized or geometric degradations) is an important area for investigation. Systems that take advantage of correlations in the source sequence can likely be designed to be significantly more robust to additive white Gaussian noise in the Gaussian-quadratic scenario of Section 7, for example. Also, the current lack of accurate, tractable source and channel models for authentication systems suggests that universal encoders or decoders which perform well regardless of the particular source and/or channel (or those which perform well under mismatch conditions [53] [54]) could be especially valuable. Finally, while the distortion region characterized by Theorem 1 in Section 4 represents fundamental asymptotic limits, bounds for finite block length systems—possibly using techniques in [37]—would also provide useful insights.

Examples of interesting future work in communication theory relating to authentication systems includes the design and analysis of practical authentication codes approaching the fundamental limits as well as investigating synchronization and registration techniques. Specifically, turbo-codes or similar coding structures suitable for iterative decoding can be expected to closely approximate the performance promised by the random coding argument used in Section 5.

Some signal processing issues of interest include appropriate source and channel models as well as the choice of a good signal basis. For example, multi-scale models and associated wavelet expansions have proven valuable in other areas of image, audio, and video processing and the multi-resolution capabilities of such techniques could prove especially valuable for designing layered authentication systems.

Also, while we have described only private-key authentication systems, in practice, public-key systems are often desirable. One approach to adapting our framework to public-key systems is outlined in the appendix for the purpose of illustration. However, more generally the issue of appropriate cryptographic tools and how they can best be combined with the other facets of authentication remains largely unexplored. For example, focusing on computationally bounded attackers may enlarge the achievable distortion region.

Finally, an interesting area of future research is the analysis of alternative notions of authentication using a different definition of security than in (1). More relaxed security requirements may yield a larger distortion region or other compelling advantages. We believe that a similar analysis for other notions of authentic (e.g., the notion of authenticity for fragile watermarking systems discussed in Section 8.2) or a form of Definition 1 requiring a finite but non-zero limit on the influence of the channel on $\hat{S}^n$) would be a major step forward in authentication research.

Appendix: Public-Key Adaptation of Private-Key Systems

To simplify the analysis we have studied private key systems where the encoder and decoder share a secret key, θ, which is hidden from the attacker. In many practical applications, however, it is more convenient to use public key systems where a public key θ_p is known to all parties (including the attacker) while a signing key, θ_s, is known only to the encoder. The advantage of public key systems is that while only the encoder possessing θ_s can encode, anyone possessing θ_p can decode and verify a properly encoded signal. In this section, we briefly describe how a secret key authentication system can be combined with a generic digital signature scheme to yield a public key system as discussed in further detail in [48] [49].

A digital signature scheme consists of a signing function $\tau = \mathcal{S}(m, \theta_s)$ and verifying function $\mathcal{V}(m, \tau, \theta_p)$. Specifically, the signing function maps an arbitrary length message m to a γ bit tag τ using the signing key θ_s. The verifying function returns true (with high probability) when given a message, public key, and tag generated using the signing function with the corresponding signing key. Furthermore, it is computationally infeasible to produce a tag accepted by the verifier without using the signing key. Many such digital signature schemes have been described in the cryptography literature where τ requires a number of bits that is sub-linear in n or even finite.

Encoding:

1. The key for the secret key authentication system is published along with the public key of the digital signature scheme.

2. The encoder uses the original authentication system to map the source S^n to $\Upsilon_n(S^n)$.

3. For a system like the one described in Section 5.1, there are a finite number of possible values for the authentic reconstruction $\hat{S}^n$ and the authentic reconstruction is a deterministic function of S^n. Thus each reconstruction can be assigned a bit representation $c(\hat{S}^n)$ and the encoder computes $\tau = \mathcal{S}(c(\hat{S}^n), \theta_s)$ using the digital signature algorithm.

4. Finally the digital signature tag τ is embedded into $\Upsilon_n(S^n)$ using an information embedding or digital watermarking algorithm. Since τ only requires a sub-linear number of bits, this process incurs an asymptotically negligible encoding distortion.

Decoding:

1. The decoder extracts an estimate $\hat{\tau}$ of the embedded tag τ. Since the size of τ is sub-linear, the probability that $\hat{\tau} \neq \tau$ when the reference channel is in effect can be made negligible.

2. Next, the decoder uses the secret key authentication decoder to produce $\hat{S}^n$ and its bit representation $c(\hat{S}^n)$.

3. The decoder checks whether the digital signature verifying algorithm $\mathcal{V}(c(\hat{S}^n), \hat{\tau}, \theta_p)$ accepts the received data as valid.

4. If so, then the decoder produces the authentic reconstruction $\hat{S}^n$. Otherwise, the decoder declares a decoding failure.

Security, Robustness, and Distortion: The security of the public key schemes is the same as the security of the public key digital signature scheme used. Specifically, the only way an attacker can fool the system is to find a matching $\hat{S}^n$ and τ accepted by the digital signature verifying algorithm. The robustness and distortion of the public key system are asymptotically the same as the underlying secret key system since embedded digital signature tag requires only a sub-linear fraction of bits.

Acknowledgments

The authors are grateful to Prof. Ram Zamir for many helpful suggestions including an improved proof of the converse part of Theorem 1. The authors would also like to thank the reviewers for their careful reading of the manuscript and suggestions for improvement.

References

[1] W. Diffie and M. E. Hellman, "New directions in cryptography," *IEEE Trans. Inform. Theory*, vol. 67, pp. 644–654, Nov. 1976.

[2] J. Fridrich, "Methods for tamper detection in digital images," *Proc. Multimedia and Security Workshop at ACM Multimedia*, 1999.

[3] C. Rey and J.-L. Dugelay, "Blind detection of malicious alterations on still images using robust watermarks," in *IEE Seminar Secure Images and Image Authentication*, 2000, pp. 7/1–7/6.

[4] R. B. Wolfgang and E. J. Delp, "A watermark for digital images," in *Proc. Int. Conf. Image Processing (ICIP)*, 1996, vol. 3, pp. 219–222.

[5] G. L. Friedman, "The trustworthy digital camera: Restoring credibility to the photographic image," *IEEE Trans. Consumer Electronics*, vol. 39, pp. 905–910, Nov. 1993.

[6] D. Kundur and D. Hatzinakos, "Digital watermarking for telltale tamper proofing and authentication," in *Proc. IEEE*, July 1999, vol. 87, pp. 1167–1180.

[7] P. W. Wong, "A public key watermark for image verification and authentication," in *Proc. Int. Conf. Image Processing (ICIP)*, 1998, vol. 1, pp. 445–459.

[8] M. Wu and B. Liu, "Watermarking for image authentication," in *Proc. Int. Conf. Image Processing (ICIP)*, 1998, vol. 2, pp. 437–441.

[9] M. P. Queluz, "Towards robust, content based techniques for image authentication," in *Proc. Workshop Multimedia Signal Processing (MMSP)*, 1998, pp. 297–302.

[10] S. Bhattacharjee and M. Kutter, "Compression tolerant image authentication," in *Proc. Int. Conf. Image Processing (ICIP)*, 1998, vol. 1, pp. 435–439.

[11] B. Macq and J.-L. Dugelay, "Watermarking technologies for authentication and protection of images," *Ann. Telecomm.*, vol. 55, no. 3–4, pp. 92–100, Mar.-Apr. 2000.

[12] J. J. Eggers and B. Girod, "Blind watermarking applied to image authentication," in *Proc. Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, Salt Lake City, Utah, May 2001.

[13] M. M. Yeung and F. Mintzer, "An invisible watermarking technique for image verification," in *Proc. Int. Conf. Image Processing (ICIP)*, 1997, vol. 2, pp. 680–683.

[14] M. Schneider and S. Chang, "A robust content based digital signature for image authentication," in *Proc. Int. Conf. Image Processing (ICIP)*, 1996, vol. 3, pp. 227–230.

[15] C.-Y Lin and S.-F. Chang, "A robust image authentication method distinguishing jpeg compression from malicious manipulation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 11, no. 2, pp. 153–168, Feb. 2001.

[16] L. Me and G. R. Arce, "A class of authentication digital watermarks for secure multimedia communication," *IEEE Trans. Image Processing*, vol. 10, no. 11, pp. 1754–1764, Nov. 2001.

[17] C.-S Lu and H. Liao, "Multipurpose watermarking for image authentication and protection," in *IEEE Trans. Image Processing*, 2001, vol. 10, pp. 1579–1592.

[18] S. I. Gel'Fand and M. S. Pinsker, "Coding for channel with random paramters," *Prob. Contr. Inform. Theory*, vol. 9, no. 1, pp. 19–31, 1980.

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

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

[21] J. A. O'Sullivan, P. Moulin, and J. M. Ettinger, "Information-theoretic analysis of steganography," in *Proc. Int. Symp. Inform. Theory*, Cambridge, MA, Aug. 1998, p. 297.

[22] B. Chen and G. W. Wornell, "Quantization index modulation: A class of provably good methods for digital watermarking and information embedding," in *Proc. Int. Symp. Inform. Theory*, Sorrento, Italy, June 2000.

[23] A. Cohen and A. Lapidoth, "On the Gaussian watermarking game," in *IEEE Int. Symp. Inform. Theory*, June 2000, p. 48.

[24] P. Moulin and J.A. O'Sullivan, "Information-theoretic analysis of information hiding," in *IEEE Int. Symp. Inform. Theory*, June 2000, p. 19.

[25] 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, no. 4, pp. 1423–1443, May 2001.

[26] P. Moulin and J. O'Sullivan, "Information-theoretic analysis of information hiding," Preprint; Dec. 2001 revision., Sep. 1999.

[27] Y. Steinberg and N. Merhav, "Identification in the presence of side information with application to watermarking," *IEEE Trans. Inform. Theory*, vol. 47, no. 4, pp. 1410–1422, May 2001.

[28] A. S. Cohen and A. Lapidoth, "The Gaussian watermarking game," *IEEE Trans. Inform. Theory*, June 2002, To appear.

[29] M. D. Swanson, M. Kobayashi, and A. H. Tewfik, "Multimedia data-embedding and watermarking technologies," in *Proc. IEEE*, June 1998, vol. 86, pp. 1064–1087.

[30] N. Memon and P. W. Wong, "Protecting digital media content," *Commun. ACM*, vol. 41, no. 7, pp. 35–42, July 1998.

[31] I. J. Cox and J.-P. M. G. Linnartz, "Some general methods for tampering with watermarks," *IEEE J. Select. Areas Commun.*, vol. 16, no. 4, pp. 587–593, May 1998.

[32] J. Chou, S. S. Pradhan, and K. Ramchandran, "On the duality between distributed source coding and data hiding," in *Proc. Asilomar Conf. Signals, Systems, Computers*, Pacific Grove, CA, 1999.

[33] S. S. Pradhan, J. Chou, and K. Ramchandran, "Duality between source and channel coding with side information," Tech. Rep. M01/34, Univ. Calif., Berkeley, 2001.

[34] J. K. Su, J. J. Eggers, and B. Girod, "Illustration of the duality between channel coding and rate distoriton with side information," in *Proc. Asilomar Conf. Signals, Systems, Computers*, Pacific Grove, CA, Nov. 2000.

[35] R. J. Barron, Brian 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, no. 5, pp. 1159–1180, May 2003.

[36] R. J. Barron, B. C. Chen, and G. W. Wornell, "The duality between information embedding and source coding with side information and some applications," in *Proc. Int. Symp. Inform. Theory*, Washington, DC, June 2001.

[37] N. Merhav, "On random coding error exponents of watermarking systems," *IEEE Trans. Inform. Theory*, vol. 46, no. 2, pp. 420–430, Mar. 2000.

[38] M. Chiang and T. M. Cover, "Unified duality of channel capacity and rate distortion with state information," in *Proc. Int. Symp. Inform. Theory*, Washington, DC, June 2001.

[39] U. Erez, S. Shamai, and R. Zamir, "Capacity and lattice-strategies for cancelling known interference," in *Proc. Int. Symp. Inform. Theory & Appl.*, Honolulu, HI, Nov. 2000, pp. 681–684.

[40] R. Zamir, S. Shamai, and U. Erez, "Nested linear/lattice codes for structured multiterminal binning," *IEEE Trans. Inform. Theory*, June 2002, To appear.

[41] A. Sutivong, T.M. Cover, Mung Chiang, and Young-Han Kim, "Rate vs. distortion trade-off for channels with state information," in *Proc. International Symposium on Information Theory*, July 2002, p. 226.

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

[43] S. Verdu and Te Sun Han, "A general formula for channel capacity," *IEEE Trans. Inform. Theory*, vol. 40, no. 4, pp. 1147–1157, Jul. 1994.

[44] Y. Steinberg and S. Verdu, "Simulation of random processes and rate-distortion theory," *IEEE Trans. Inform. Theory*, vol. 42, no. 1, pp. 63–86, Jan. 1996.

[45] U. Mittal and N. Phamdo, "Hybrid digital-analog (HDA) joint source-channel codes for broadcasting and robust communications," *IEEE Trans. Inform. Theory*, vol. 48, no. 5, pp. 1082–1102, May 2002.

[46] Z. Reznic, R. Zamir, and M. Feder, "Joint source-channel coding of a Gaussian mixture source over the gaussian broadcast channel," *IEEE Trans. Inform. Theory*, vol. 48, no. 3, pp. 776–781, Mar. 2002.

[47] S. Shamai, S. Verdu, and R. Zamir, "Systematic lossy source/channel coding," *IEEE Trans. Inform. Theory*, vol. 44, no. 2, pp. 564–579, Mar. 1998.

[48] E. Martinian, B. Chen, and G. W. Wornell, "Information theoretic approach to the authentication of multimedia," in *Proc. SPIE: Security and Watermarking of Multimedia Contents III (part of Electronic Imaging 2001)*, 2001.

[49] E. Martinian, "Authenticating multimedia in the presence of noise," M.S. thesis, Massachusetts Institute of Technology, Cambridge, MA, 2000.

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

[51] R. Ahlswede and J. Körner, "Source coding with side information and a converse for degraded broadcast channels," *IEEE Trans. Inform. Theory*, vol. IT-21, no. 6, pp. 629–637, Nov. 1976.

[52] T. Berger, *Rate Distortion Theory: A Mathematical Basis For Data Compression*, Prentice-Hall, Englewood Cliffs, NJ, 1971.

[53] N. Merhav, G. Kaplan, A. Lapidoth, and S. Shitz, "On information rates for mismatched decoders," *IEEE Trans. Inform. Theory*, vol. 40, no. 6, pp. 1953–1967, Nov. 1994.

[54] A. Lapidoth, "Nearest neighbor decoding for additive non-Gaussian noise channels," *IEEE Trans. Inform. Theory*, vol. 42, no. 5, pp. 1520–1529, Sep. 1996.