Authentication With Distortion Criteria

Emin Martinian, *Member, IEEE*, Gregory W. Wornell, *Fellow, IEEE*, and Brian Chen

***Abstract*—In a variety of applications, there is a need to authenticate content that has experienced legitimate editing in addition to potential tampering attacks. We develop one formulation of this problem based on a strict notion of security, and characterize and interpret the associated information-theoretic performance limits. The results can be viewed as a natural generalization of classical approaches to traditional authentication. Additional insights into the structure of such systems and their behavior are obtained by further specializing the results to Bernoulli and Gaussian cases. The associated systems are shown to be substantially better in terms of performance and/or security than commonly advocated approaches based on data hiding and digital watermarking. Finally, the formulation is extended to obtain efficient layered authentication system constructions.**

***Index Terms*—Coding with side information, data hiding, digital signatures, digital watermarking, information embedding, joint source–channel coding, multimedia security, robust hashing, tamper-proofing, transaction-tracking.**

I. INTRODUCTION

IN traditional authentication problems, the goal is to determine whether some content being examined is an exact replica of what was created by the author. Digital signature techniques [1] are a natural tool for addressing such problems. In such formulations, the focus on exactness avoids consideration of semantic issues. However, in many emerging applications, semantic issues are an integral aspect of the problem, and cannot be treated separably. As contemporary examples, the content of interest may be an audio or video waveform, or an image, 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, benign or malicious, and semantically significant or not. Methods for reliable authentication from such perturbed data are important as well.

The spectrum of applications where such authentication capabilities will be important is enormous, ranging from drivers' licenses, passports, and other government-issued photo identification; to news photographs and interview tapes; to state-issued currency and other monetary instruments; to legal evidence in the form of audio and video recordings in court cases. Indeed, the rapidly increasing ease with which such content can be digitally manipulated in sophisticated ways using inexpensive systems, whether for legitimate or fraudulent purposes, is of considerable concern in these applications.

Arising out of such concerns, a variety of technologies have been introduced to facilitate authentication in such settings. Examples include various physical watermarking technologies—such as hologram imprinting in images—as well as more recent digital descendents. See, e.g., [2] for some of the rich history in this area going back several hundred years. However, regardless of the implementation, all involve the process of marking or altering the content in some way, which can be viewed as a form of encoding.

A rather generic problem that encompasses essentially all the applications of interest is that of transaction-tracking in a content migration scenario. In this scenario, there are essentially three types of participants involved in the migration of a particular piece of content. There is the original author or creator of the content, who delivers an encoding of it.[1] There is the editor who makes modifications to this encoded content, and publishes the result.[2] And there is the reader or end-user for whom the published work is intended. The reader wants to be able to determine 1) whether published work being examined was derived from content originally generated by the author, and 2) how it was modified by the editor. At the same time, the editor wants the author's encoding to be (semantically) close to the original content, so that the modifications can take the semantics into account as necessary.

In the recent literature, researchers have proposed a variety of approaches to such problems based on elements of digital watermarking, cryptography, and content classification; see, e.g., [3]–[18] and the references therein. 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.

Within this literature, there are two basic types of approaches. In the first, the authentication mechanism is based on embedding what is referred to as a "fragile" watermark known to both the encoder and the decoder into the content of interest. At the decoder, a watermark is extracted and compared to the known watermark inserted by the encoder. The difference between

Manuscript received May 30, 2002; revised December 6, 2003 and March 9, 2005. This work was supported in part by the National Science Foundation under Grant CCR-0073520 and through a National Science Foundation Graduate Fellowship, Microsoft Research, Hewlett-Packard through the MIT/HP Alliance, and Texas Instruments through the Leadership Universities Program. The material in this paper was presented in part at the IEEE International Symposium on Information Theory, Washington, DC, June 2001.

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

Communicated by G. Battail, Associate Editor At Large.
Digital Object Identifier 10.1109/TIT.2005.850123

[1]There are no inherent restrictions on what can constitute "content" in this generic problem. Typical examples include video, audio, imagery, text, and various kinds of data.

[2]The motives and behavior of the editor naturally depend on the particular application and situation. At one extreme, the editor might just perform some benign resampling or other transcoding, or, at the other extreme, might attempt to create a forgery from the content. In the latter case, the editor would be considered an attacker.

the extracted watermark and the known watermark is then interpreted as a measure of authenticity. Examples of this basic approach include [5], [7], [13], [14].

The second type of approach is based on a "robust" watermarking strategy, whereby the important features of the content are extracted, compressed and embedded back into the content by the encoder. The decoder attempts to extract the watermark from the content it obtains and authenticates by comparing the features encoded in the watermark to the features in the content itself. This strategy is sometimes termed "self-embedding." Examples of this basic approach include [4], [11], [15].

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 are 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 [19]–[21]; see, e.g., [22]–[42] and references therein. Collectively, this work provides a useful context within which to examine the topic of authentication.

Our contribution in this paper is to propose one possible formulation for the general problem of authentication with a semantic model, and examine its implications. In particular, using distortion criteria to capture semantic aspects of the problem, we assess performance limits in terms of the inherent tradeoffs between security, robustness, and distortion, and in turn develop the structure of systems that make these tradeoffs efficiently. As we will show, these systems have important distinguishing characteristics from those proposed to date. We also see that under this model, the general 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 and terminology in Section II. Next, in Section III, we develop a system model and problem formulation, quantifying a notion of authentication. In Section IV, we characterize the performance limits of such systems via our main coding theorem. Section V contains both the associated achievability proof, which identifies the structure of good systems, and a converse. In Section VI, the results are applied to the case of binary content with Hamming distortion measures, and in Section VII, to Gaussian content with quadratic distortion measures. Section VIII then analyzes other classes of authentication techniques in the context of our framework, and shows that they are inherently either less efficient or less secure that the systems developed here. Next, Section IX generalizes the results of the paper to include layered systems that support multiple levels of authentication. Finally, Section X contains some concluding remarks.

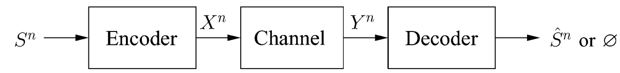

Fig. 1. Authentication system model. The source S^n is encoded by the content creator into X^n, incurring some distortion. The channel models the actions of the editor, i.e., all processing experienced by the encoded content before it is made available to the end-user. The decoder, controlled by the end-user, 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 using the special symbol $\varnothing$.

II. Notation and Terminology

We use standard information theory notation (e.g., as found in [43]). 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 $\{v_i, v_{i+1}, \dots, v_j\}$, 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.

Since the applications are quite varied, we keep our terminology rather generic. The content of interest, as well as its various encodings and recontructions, will be generically referred to as "signals," regardless of whether they refer to video, audio, imagery, text, data, or any other kind of content. The original content we will also sometimes simply refer to as the "source." Moreover, we will generally associate any manipulations of the encoded content with the "editor," regardless of whether any human is involved. However, as an exception, we will often use the term "attacker" in lieu of "editor" for cases where the manipulations are specifically of a malicious nature.

III. System Model and Problem Formulation

Our system model for the transaction-tracking scenario is as depicted in Fig. 1. To simplify the exposition, we model the original content as an independent and identically distributed (i.i.d.)[3] sequence $S_1, S_2, \dots, S_n$. 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 is suitably close to S^n with respect to some distortion measure. The encoder is under the control of the content creator. The encoded signal then passes through a channel, which models the actions of the generic "editor," and encompasses all processing experienced by the encoded signal before it is made available to the end-user as Y^n. This processing would include all effects ranging from routine handling to malicious tampering. The decoder, which is controlled by the end-user, 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 modifications by the editor, or declares that it is not possible to produce such a reconstruction. We term such reconstructions "authentic."

[3]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 would be obtained in more general settings (e.g., using techniques from [44], [45]).

Our approach to the associated channel modeling issues in the formulation of Fig. 1 has some novel features, and thus warrants special discussion. Indeed, as we now discuss, our approach to such modeling is not to *anticipate* the possible behaviors of the editor, but to effectively *constrain* them. In particular, we avoid choosing a model that tries to characterize the range of processing the editor might undertake. If we did, the security properties of the resulting system would end up being sensitive to any modeling errors, i.e., to any behavior of the editor that is inconsistent with the model.

Instead, the focus is on choosing a model that defines the range of processing the editor can undertake and have such edits accepted by the end-user. We refer to this as our "reference channel model." Specifically, we effectively design the system such that the decoder will successfully authenticate the modified content if and only if the edits are consistent with the reference channel model. Thus, the editor is free to edit the content in any way (and we make no attempt to model the range of behavior), but the subset of behaviors for which the system will authenticate is strictly controlled via the reference channel construct. Ultimately, since the end-user will not accept content that cannot be authenticated, the editor will constrain its behavior according to the reference channel.

From this perspective, the reference channel model is a system design parameter, and thus is known *a priori* to encoders, decoders, and editors. To simplify our analysis, we will restrict our attention to memoryless probabilistic reference channel models. In this case, the model is characterized by a simple conditional distribution $p(Y \mid X)$.

As our main result, in Section IV we characterize when authentication systems with the above-described behavior 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 chosen reference distribution $p(y \mid x)$. Then we determine which distortion pairs $(D_{\mathrm{e}}, D_{\mathrm{r}})$ are asymptotically achievable.

We emphasize that the distortion pair $(D_{\mathrm{e}}, D_{\mathrm{r}})$ corresponds precisely to the performance characteristics of direct interest in the system for the transaction-tracking scenario. Indeed, a small D_{e} means the editor is given a faithful version of the original content with which to work. Moreover, a small D_{r} means that the end-user is able to accurately estimate the editor's modifications by comparing the decoder input to the authentic reconstruction.

A. Defining "Authenticity"

To develop our main results, we first need to quantify the concept of an "authentic reconstruction." Recall that our intuitive notion of an authentic reconstruction is one that is free from the effects of the edits when the reference channel is in effect. Formally, this is naturally expressed as follows.

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

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

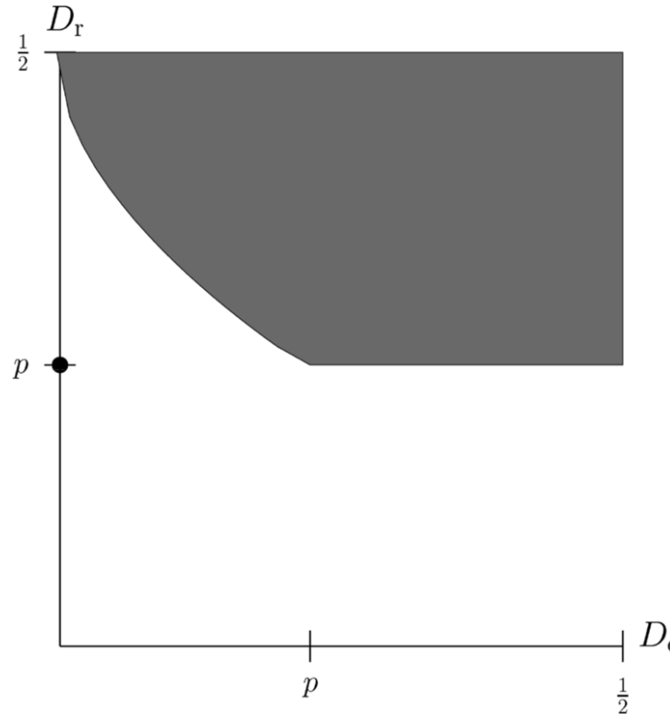

Fig. 2. The shaded area depicts the achievable distortion region for a symmetric Bernoulli source used in conjunction with a binary-symmetric reference channel of crossover probability p. Distortions are with respect to the Hamming measure. The case $p = 0$ corresponds to traditional digital signatures. If authentication was not required, the point $(D_{\mathrm{e}} = 0, D_{\mathrm{r}} = p)$ could be achieved.

Note that as special cases, this definition would include systems in which, for example, $\hat{S}^n$ is a deterministic or randomized function of S^n. More generally, this definition means that the authentic reconstructions are effectively defined by the encoder in such systems. This will have implications later in the system design.

B. An Example Distortion Region

Before developing our main result, we illustrate with an example the kinds of results that will be obtained. This example corresponds to a problem involving a symmetric Bernoulli source, Hamming distortion measures, and a (memoryless) binary-symmetric reference channel with crossover probability p.

Under this example scenario, the editor is allowed to flip a fraction p of the binary source samples, and the end-user must (almost certainly) be able to generate an authentic reconstruction from such a perturbation. If the edits are generated from a different distribution, such as a binary-symmetric channel with a crossover probability greater than p, then the decoder must (almost certainly) declare an authentication failure.

The corresponding achievable distortion region is depicted in Fig. 2. Several points on the frontier are worth discussing. First, note that the upper left point on the frontier, i.e., $(D_{\mathrm{e}}, D_{\mathrm{r}}) = (0, 1/2)$, reflects that if no encoding distortion is allowed, then authentic reconstructions are not possible, since the maximum possible distortion is incurred. At the other extreme, the lower right point of the frontier, i.e., $(D_{\mathrm{e}}, D_{\mathrm{r}}) = (1/2, p)$, corresponds to a system in which the source is first source coded to distortion p, after which the resulting bits are digitally signed and channel coded for the binary-symmetric channel (BSC).

While no amount of encoding distortion can reduce the reconstruction distortion below p, the point $(D_\mathrm{e}, D_\mathrm{r}) = (p, p)$ on the frontier establishes that a reconstruction distortion of p is actually achievable with much less encoding distortion than the lower right point suggests. In fact, because the required encoding distortion is only p, the decoder can be viewed as completely eliminating 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.

The more general structure of the frontier is also worth observing. In particular, D_r is a decreasing function of D_e along the frontier. This reflects that the objectives of small D_e (which the editor wants) and a small D_r (which the end-user wants) are conflicting and a fundamental tradeoff is involved for any given reference channel. In fact, as we will see in the sequel, this behavior is not specific to this example, but a more general feature of our authentication problem formulation.[4]

Finally, observe that the achievable region decreases monotonically with p, the severity of edits allowed. Thus, if one has particular target encoding and reconstruction distortions, then this effectively limits how much editing can be tolerated. As the extreme point, the case $p = 0$ in which no editing is allowed corresponds to the traditional scenario for digital signatures. In this case, as the figure reflects, authentication is achievable without incurring any encoding distortion nor reconstruction distortion. It is worth noting that the nature of the interplay between the severity of the reference channel and the achievable distortion region is not specific to this example, but arises more generally with this formulation of the authentication problem.

IV. Characterization of Solution: Coding Theorems

An instance of the authentication problem consists of the seven-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. The secret key is shared only between the encoder and decoder; all other information is known to all parties including editors. (For the interested reader, straightforward adaptations of our solutions to public-key implementations are summarized in the Appendix. However, we otherwise restrict our attention to private-key schemes in the paper to focus the exposition.)

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.,

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

[4]This should not be surprising, since such tradeoffs frequently arise in joint source–channel coding problems with uncertain channels; see, e.g., [46]–[48].

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

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

Notice that since an authentic reconstruction must satisfy (1), and since the decoder must satisfy the Markov condition

$$\{S^n, X^n\} \leftrightarrow Y^n \leftrightarrow \Phi_n(Y^n, \theta)$$

we have that

$$\hat{S}^n \leftrightarrow \{S^n, X^n\} \leftrightarrow \Phi_n(Y^n, \theta)$$

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 $\hat{S}^n = \Phi_n(Y, \theta^n)$ will generally not satisfy (1).

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

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

Evidently,

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

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

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 $\epsilon > 0$

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

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

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

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

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 $n \to \infty$, $\Pr[\mathcal{E}_{\rm sa}] \to 0$ regardless of the channel law in effect, $\Pr[\mathcal{E}_{D_{\rm e}}] \to 0$, and $\Pr[\mathcal{E}_{D_{\rm r}}] \to 0$ when the reference channel is in effect, with $\mathcal{E}_{\rm sa}, \mathcal{E}_{D_{\rm e}}$, and $\mathcal{E}_{D_{\rm r}}$ as defined in (8), (6), and (7).

For such systems, we have the following coding theorem.

Theorem 1: The distortion pair $(D_{\rm e}, D_{\rm r})$ 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

$$I(U;Y) - I(S;U) \geq 0 \tag{9a}$$
$$E[d_{\rm e}(S, f(U,S))] \leq D_{\rm e} \tag{9b}$$
$$E[d_{\rm r}(S, g(U))] \leq D_{\rm r}. \tag{9c}$$

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 that can be authenticated, X represents the encoding of the source S, 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 editor's manipulations—i.e., the realization of the reference channel—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 $g(U,Y)$. Thus, as we shall see in Sections VI and VII, security comes at a price in this problem.

Theorem 1 has some interesting features. First, it is worth noting that since the problem formulation is inherently "analog," dealing only with waveforms, we might expect the best solutions to the problem to be analog in nature. However, what the theorem suggests, and what its proof confirms, is that digital solutions are in fact sufficient to achieve optimality. In particular, as we will see, source and channel coding based on discrete codebooks are key ingredients of the achievability argument. In some sense, this is the consequence of the inherently discrete functionality we have required of the decoder with our formulation.

As a second remark, note that Theorem 1 can be contrasted with its information embedding counterpart, which as generalized from [19] in [36], states that a pair $(R, D_{\rm e})$, where R is the embedding rate, is achievable if and only if there exists a function $f(\cdot,\cdot)$ and a distribution

$$p(y,x,u,s) = p(s)p(u\,|\,s)p(x\,|\,s,u)p(y\,|\,x)$$

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

with X deterministic (i.e., $p(x\,|\,u,s) = 1_{x=f(s,u)}$) such that

$$I(U;Y) - I(S;U) \geq R \tag{10a}$$
$$E[d_{\rm e}(S, f(U,S))] \leq D_{\rm e}. \tag{10b}$$

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.

A. General System Structure

As developed in detail in Section V, an optimal authentication system can be constructed by choosing a codebook $\mathcal{C}$ with codewords appropriately distributed over the space of possible source outcomes. The elements of 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 editors.

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 the signal it obtains 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 attackers do not know $\mathcal{A}$, if an 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 should be sufficiently separated that they can be resolved at the output of the reference channel.

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 corresponds 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 encoding X^n. Since the

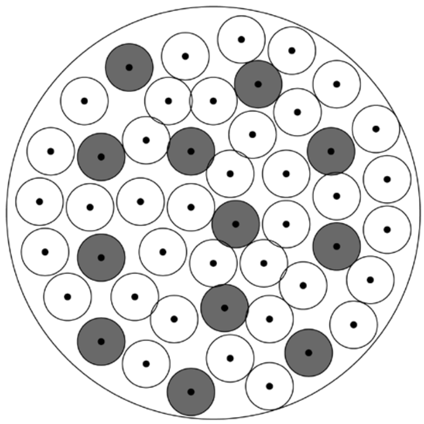

Fig. 3. Codebook construction for the Gaussian-quadratic scenario. The large sphere represents the space of possible source vectors and the small spheres representing the noise are centered on codewords. When the small spheres do not overlap, the codewords can be resolved at the output of the reference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

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 encoding X^n, and the noise spheres do not intersect since $r > \sigma_N$.

Furthermore, when the reference channel is not in effect and an attacker tampers with the signal such that the nearest codeword C is different from that chosen by the encoder 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 nonadmissible 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 the Appendix as well as [49], [50] for insights into the practical implementation of this class of systems including those designed based on a public key instead of a secret key.

V. Proofs

A. 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.

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.,

$$\gamma \to 0 \quad \text{and} \quad n\gamma \to \infty, \qquad \text{as } n \to \infty. \tag{12}$$

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.

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 selects 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 obtained 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 $\varnothing$.

3) System Failure Probabilities: We begin by analyzing the system failure probabilities.

a) Probability of successful attack: Suppose the attacker causes the codeword obtained by the decoder 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,

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

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.

b) Probability of distortion violation: The distortion violation events $\mathcal{E}_{D_e}$ and $\mathcal{E}_{D_r}$ defined in (6) and (7) can arise due to any of the following typicality failure events.

- $\mathcal{E}_{st}$: The source is not typical.
- $\mathcal{E}_{et}$: The encoder fails to find an admissible codeword that is jointly typical with its input.
- $\mathcal{E}_{ct}$: The channel fails to produce an output jointly typical with its input when the reference channel law is in effect.
- $\mathcal{E}_{dt}$: 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

$$\mathcal{E}_{tf} = \mathcal{E}_{st} \cup \mathcal{E}_{et} \cup \mathcal{E}_{ct} \cup \mathcal{E}_{dt} \quad (16)$$

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

$$\begin{aligned}\Pr[\mathcal{E}_{dv}] &= \Pr[\mathcal{E}_{dv} \mid \mathcal{E}_{tf}] \cdot \Pr[\mathcal{E}_{tf}] + \Pr[\mathcal{E}_{dv} \mid \mathcal{E}_{tf}^c] \cdot \Pr[\mathcal{E}_{tf}^c] \\ &\le \Pr[\mathcal{E}_{dv} \mid \mathcal{E}_{tf}^c] + \Pr[\mathcal{E}_{tf}] \\ &= \Pr[\mathcal{E}_{dv} \mid \mathcal{E}_{tf}^c] + \Pr[\mathcal{E}_{st}] + \Pr[\mathcal{E}_{et} \mid \mathcal{E}_{st}^c] \\ &\quad + \Pr[\mathcal{E}_{ct} \mid \mathcal{E}_{st}^c, \mathcal{E}_{et}^c] + \Pr[\mathcal{E}_{dt} \mid \mathcal{E}_{st}^c, \mathcal{E}_{et}^c, \mathcal{E}_{ct}^c]. \quad (17)\end{aligned}$$

First, according to well-known properties of typical sequences [43], by choosing n large enough we can make

$$\Pr[\mathcal{E}_{st}] \le \epsilon/4 \quad (18)$$

$$\Pr[\mathcal{E}_{ct} \mid \mathcal{E}_{st}^c, \mathcal{E}_{et}^c] \le \epsilon/4. \quad (19)$$

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

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

from standard joint typicality arguments.

Third,

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

Indeed, using standard joint typicality results, the probability that the sequence Y^n presented to the decoder 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

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

Fourth,

$$\Pr[\mathcal{E}_{dv} \mid \mathcal{E}_{tf}^c] = 0. \quad (22)$$

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

$$\frac{1}{n}\sum_{i=1}^{n} d_e(S_i, X_i) \le E[d_e(S,X)] + \delta \cdot \bar{d}_1$$

$$\frac{1}{n}\sum_{i=1}^{n} d_r(S_i, g_i(U_i)) \le E[d_r(S, g(U))] + \delta \cdot \bar{d}_2$$

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

$$\bar{d}_1 = \sup_{(s,x)\in\mathcal{S}\times\mathcal{X}} d_e(s,x) \quad (23)$$

$$\bar{d}_2 = \sup_{(s,\hat{s})\in\mathcal{S}\times\mathcal{S}} d_r(s,\hat{s}). \quad (24)$$

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)–(22) in (17) we obtain

$$\Pr[\mathcal{E}_{dv}] \le 3\epsilon/4 + 2 \cdot 2^{-n\gamma} \quad (25)$$

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

B. Converse Part: Necessity

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

$$R^*(D_e, D_r) \triangleq \sup_{\substack{p(U \mid S), f:\mathcal{U}\times\mathcal{S}\mapsto\mathcal{X}, g:\mathcal{U}\mapsto\mathcal{S}: \\ E[d_e(S,f(U,S))]\le D_e, E[d_r(S,g(U))]\le D_r}} I(U;Y) - I(S;U). \quad (26)$$

Note that $R^*(D_e, D_r) \ge 0$ 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 IV exist with $\lim_{n\to\infty} \Pr[\mathcal{E}_{sa}] = 0$, $\lim_{n\to\infty} \Pr[\mathcal{E}_{D_e}] = 0$, and—when the reference channel is in effect—$\lim_{n\to\infty} \Pr[\mathcal{E}_{D_r}] = 0$. We then show that these functions imply that $R^*(D_e, D_r) \ge 0$ 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 [19] or by directly applying [36, Lemma 2] it is possible to prove that allowing X to be nondeterministic has no advantage, i.e.,

$$R^*(D_e, D_r) \ge \sup_{\substack{p(U \mid S), p(X \mid U,S): \\ E[d_e(S,X)]\le D_e, E[d_r(S,g(U))]\le D_r}} I(U;Y) - I(S;U). \quad (27)$$

Arguments similar to those in [19] and [36, Lemma 1] show that $R^*(D_e, D_r)$ is monotonically nondecreasing and concave in (D_e, D_r). These properties will later allow us to make use of the following lemma, whose proof follows readily from that of Lemma 4 in [19].

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

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

As demonstrated by the following lemma, a suitable U_i is

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

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

$$Y_i \leftrightarrow (S_i, X_i) \leftrightarrow U_i. \quad (30)$$

Proof: It suffices to note that

$$p(y_i \mid x_i, s_i) = p(y_i \mid x_i) = \frac{p\left(y_1^i \mid x^n\right)}{p\left(y_1^{i-1} \mid x^n\right)}$$
$$= \frac{p\left(y_1^i \mid x^n, s^n\right)}{p\left(y_1^{i-1} \mid x^n, s^n\right)} \quad (31)$$
$$= \frac{p\left(y_1^i \mid x^n, \hat{s}^n, s^n\right)}{p\left(y_1^{i-1} \mid x^n, \hat{s}^n, s^n\right)} = p\left(y_i \mid x^n, s^n, \hat{s}^n, y_1^{i-1}\right) \quad (32)$$

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

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

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_e, D_r) when the reference channel is in effect then

$$R^*(D_e, D_r) \geq 0. \quad (34)$$

Proof: Define $D_{e,i}$ and $D_{r,i}$ 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:

$$R^*(D_e, D_r) = R^*\left(\frac{1}{n}\sum_{i=1}^n D_{e,i}, \frac{1}{n}\sum_{i=1}^n D_{r,i}\right) \quad (35)$$
$$\geq \frac{1}{n}\sum_{i=1}^n R^*(D_{e,i}, D_{r,i}) \quad (36)$$
$$\geq \frac{1}{n}\sum_{i=1}^n [I(U_i; Y_i) - I(U_i; S_i)] \quad (37)$$
$$\geq \frac{1}{n}[I(\hat{S}^n; Y^n) - I(\hat{S}^n; S^n)] \quad (38)$$
$$= \frac{1}{n}[H(\hat{S}^n \mid S^n) - H(\hat{S}^n \mid Y^n)] \quad (39)$$
$$\geq -\frac{1}{n}H(\hat{S}^n \mid Y^n) \quad (40)$$
$$\geq -\frac{1}{n} - \Pr[\Phi_n(Y^n) \neq \hat{S}^n]\log|\mathcal{S}|. \quad (41)$$

The concavity of $R^*(D_e, D_r)$ 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

$$\Pr[\Phi_n(y^n) \neq \hat{S}^n]$$
$$= \Pr[\mathcal{E}_{sa}] + \Pr[\{\Phi_n(Y^n) \neq \hat{S}^n\} \cap \{\Phi_n(Y^n) = \varnothing\}] \quad (42)$$
$$\leq \Pr[\mathcal{E}_{sa}] + \Pr[\{\Phi_n(Y^n) = \varnothing\}] \quad (43)$$
$$\leq \Pr[\mathcal{E}_{sa}] + \Pr[\mathcal{E}_{D_r}]. \quad (44)$$

Therefore, exploiting that the system generates an authentic reconstruction ($\lim_{n\to\infty} \Pr[\mathcal{E}_{sa}] = 0$) of the right distortion ($\lim_{n\to\infty} \Pr[\mathcal{E}_{D_r}] = 0$) 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 $p(x \mid u, s)$ singular or over an alphabet $\mathcal{U}$ of cardinality at most $|\mathcal{S}| + |\mathcal{X}| + 3$ if $p(x \mid u, s)$ 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 $f_j(\boldsymbol{p})$ where $\boldsymbol{p}$ represents a distribution of the form $\Pr(S = s, X = x \mid U = u)$ and the $f_j(\cdot)$ functions capture the features of the distributions relevant to (9). According to Carathéodory's theorem [43, Theorem 14.3.4], [51], there exist $j_{\max} + 1$ distributions $\boldsymbol{p}_1$ through $\boldsymbol{p}_{j_{\max}+1}$ such that any vector of function values, $(f_1(\boldsymbol{p}'), f_2(\boldsymbol{p}'), \ldots, f_{j_{\max}}(\boldsymbol{p}'))$, achieved by some distribution $\boldsymbol{p}'$ can be achieved with a convex combination of the $\boldsymbol{p}_i$ 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 [52, bottom left of p. 634].

1) Replace $\Pr(X = x|U = u)$ with $\Pr(S = s, X = x|U = u)$ which is represented by the notation $\boldsymbol{p}$.
2) Choose
$$f_j(\boldsymbol{p}) = \sum_x \Pr(S = j, X = x \mid U = u) \quad (45)$$
for $j \in \{1, 2, \ldots, n\}$ where $n = |\mathcal{S}|$.
3) Choose
$$f_{n+1}(\boldsymbol{p}) = \sum_s \sum_x d_e(x, s) \Pr(S = s, X = x \mid U = u). \quad (46)$$
4) Choose
$$f_{n+2}(\boldsymbol{p}) = \sum_s \sum_x d_r(g(u), s) \Pr(S = s, X = x \mid U = u). \quad (47)$$
5) Choose
$$f_{n+3}(\boldsymbol{p}) = \sum_s \left[\sum_x \Pr(S = s, X = x \mid U = u) \cdot \log\left(\sum_x \Pr(S = s, X = x \mid U = u)\right)\right]. \quad (48)$$
6) Let
$$m(s, u, x, y) = \Pr(Y = y \mid X = x)\Pr(S = s, X = x \mid U = u)$$

and choose

$$f_{n+4}(\boldsymbol{p}) = \sum_y \left[\left(\sum_x \sum_s m(s,u,x,y) \right) \cdot \left(\sum_x \sum_s \log m(s,u,x,y) \right) \right]. \quad (49)$$

7) Choose

$$f_{n+5+j}(\boldsymbol{p}) = \sum_s \Pr(S=s, X=j \,|\, U=u) \quad (50)$$

for $j \in \{1, 2, \ldots, |\mathcal{X}|\}$.

Since the $f_j(\boldsymbol{p})$ determine $\Pr[S=s]$ (and, therefore, $H(S)$ as well), $D_{\mathrm{e}}, D_{\mathrm{r}}, H(S\,|\,U), H(Y\,|\,U)$, and $\Pr[X=x]$ (and, therefore, $\Pr[Y=y]$ and $H(Y)$ as well), they can be used to identify all points in the distortion region. According to [52, 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 [36, 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

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

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.

VI. Example: The Binary-Hamming Scenario

In some applications of authentication, the content of interest is 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. Or the alterations might be by the hand of human editor whose job it is to correct, refine, or otherwise enhance the exposition in preparation for its publication in a paper, journal, magazine, or book. Or the alternations may be the result of an attacker deliberately tampering with the text for the purpose of distorting its meaning and affecting how it will be interpreted.

As perhaps the simplest model representative of such discrete problems, we now 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[6] 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

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

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

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

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

$$D_{\mathrm{e}} = \alpha\tau. \quad (54)$$

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

$$\hat{S} = U \bmod 2 \quad (55)$$

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

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

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

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

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.,

$$h(q) = -q \log q - (1-q)\log(1-q), \qquad \text{for } 0 \leq q \leq 1.$$

Similarly, $I(U;Y)$ takes the form

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

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),

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

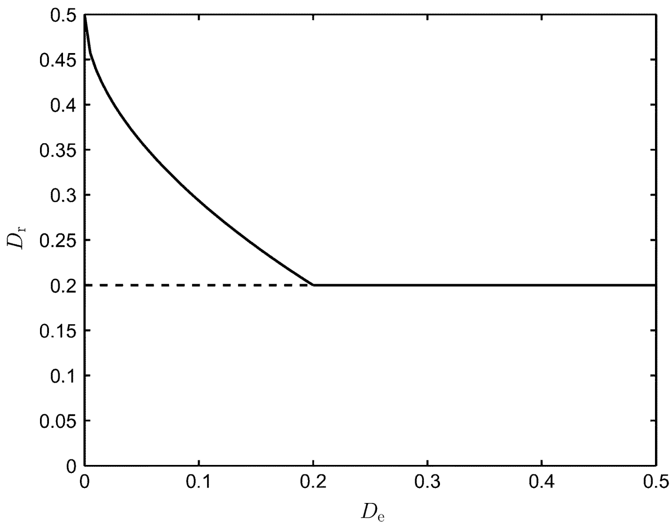

Fig. 4. The solid curve represents the frontier of the achievable distortion region for a binary-symmetric source and a binary-symmetric reference channel with crossover probability $p = 0.2$. This plot reflects the system behavior when the reference channel is in effect. The dashed line represents the boundary of the larger distortion region achievable when authentication is not required.

and (56) that the boundary point $D_{\rm e} = D_{\rm r} = p$, in particular, is obtained by the parameter values $\alpha = 1$ and $\tau = p$ (with any choice of ν). Numerical optimization over all $p(u \mid s)$ and all (not necessarily singular) $p(x \mid s, u)$ with the alphabet size $|\mathcal{U}| = 7$ chosen in accordance with Proposition 1 confirms that Fig. 4 captures all achievable distortion pairs.

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 [36], optimality can also be achieved without any coding in the binary-Hamming case and thus all points in the region $D_{\rm e} \geq 0$ and $D_{\rm r} \geq p$ 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.

VII. Example: The Gaussian-Quadratic Scenario

In some other applications of authentication, the content of interest is inherently continuous. Examples involve sources such as imagery, video, or audio. In addition to tampering attacks, such content may encounter degradations as a result of routine handling that includes compression, transcoding, resampling, printing, and scanning, as well as perturbations from editing to enhance the content.

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 V 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 it appears difficult to obtain a closed-form expression for the optimal distribution for U,[7] we instead develop good inner and outer bounds on the boundary of the achievable distortion region.

A. 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 constraint on the power P input to the reference channel, it is well known that the minimum possible source reconstruction distortion $D_{\rm r}$ achievable from the output of the channel can be achieved without either source or channel coding in this Gaussian scenario, and the resulting distortion is

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

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

$$\begin{aligned} P = E[X^2] &= E[(X - S + S)^2] \\ &= E[(X-S)^2] + E[S^2] + 2E[(X-S)S] \\ &\leq D_{\rm e} + \sigma_S^2 + 2\sqrt{D_{\rm e}\sigma_S^2} \end{aligned} \tag{61}$$

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

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

B. 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.[8] We examine two regimes: a low $D_{\rm e}$ regime in which we restrict our attention to the parameterization $(a,b,c,d) = (1,1,1/\alpha,1)$, and a high $D_{\rm e}$ regime in which we restrict our attention to the parameterization $(a,b,c,d) = (1,\beta,1,0)$. As we will see, time

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

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

sharing between these parameterizations yields almost the entire achievable distortion region for Gaussian codebooks.

1) 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 [20]. In the language of [26], 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

$$U = S + T/\alpha \tag{65}$$
$$X = U + (1-\alpha)(S-U) = S + T \tag{66}$$

where T is a Gaussian random variable with mean zero and variance σ_T^2 independent of both the source S and the reference 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)

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

and, in turn

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

To show that distortions (67) and (68) are achievable requires proving that (9a) holds. In [20], the associated difference of mutual informations is computed (using slightly different notation) as

$$\begin{aligned} &I(U;Y) - I(S;U) \\ &\quad = \frac{1}{2}\log\frac{\sigma_T^2\left(\sigma_T^2 + \sigma_S^2 + \sigma_N^2\right)}{\sigma_T^2\sigma_S^2(1-\alpha)^2 + \sigma_N^2\left(\sigma_T^2 + \alpha^2\sigma_S^2\right)} \end{aligned} \tag{69}$$

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

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

Collecting terms in powers of α yields

$$\begin{aligned} &\alpha^2\left(\sigma_T^2\sigma_S^2 + \sigma_N^2\sigma_S^2\right) - 2\alpha\sigma_T^2\sigma_S^2 - \sigma_T^4 \\ &\qquad = (\alpha - r_+)(\alpha - r_-) \leq 0 \end{aligned} \tag{71}$$

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

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

where $\alpha_\mathrm{ie} = \sigma_T^2/(\sigma_T^2 + \sigma_N^2)$ is the corresponding information embedding scaling parameter determined by Costa [20]. 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.

2) 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

$$D_\mathrm{e} = (1-\beta)^2\sigma_S^2 + \beta^2\sigma_T^2 \tag{77}$$
$$D_\mathrm{r} = \frac{\sigma_S^2\sigma_T^2}{\sigma_S^2 + \sigma_T^2}. \tag{78}$$

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

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

the mutual information constraint (9a) implies that

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

C. 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 SNRs. To see that the first and second outer bounds are asymptotically the best achievable for low and high D_e, respectively, we superimpose on these figures the best Gaussian codebook performance, as obtained by numerically optimizing the parameters in (63).

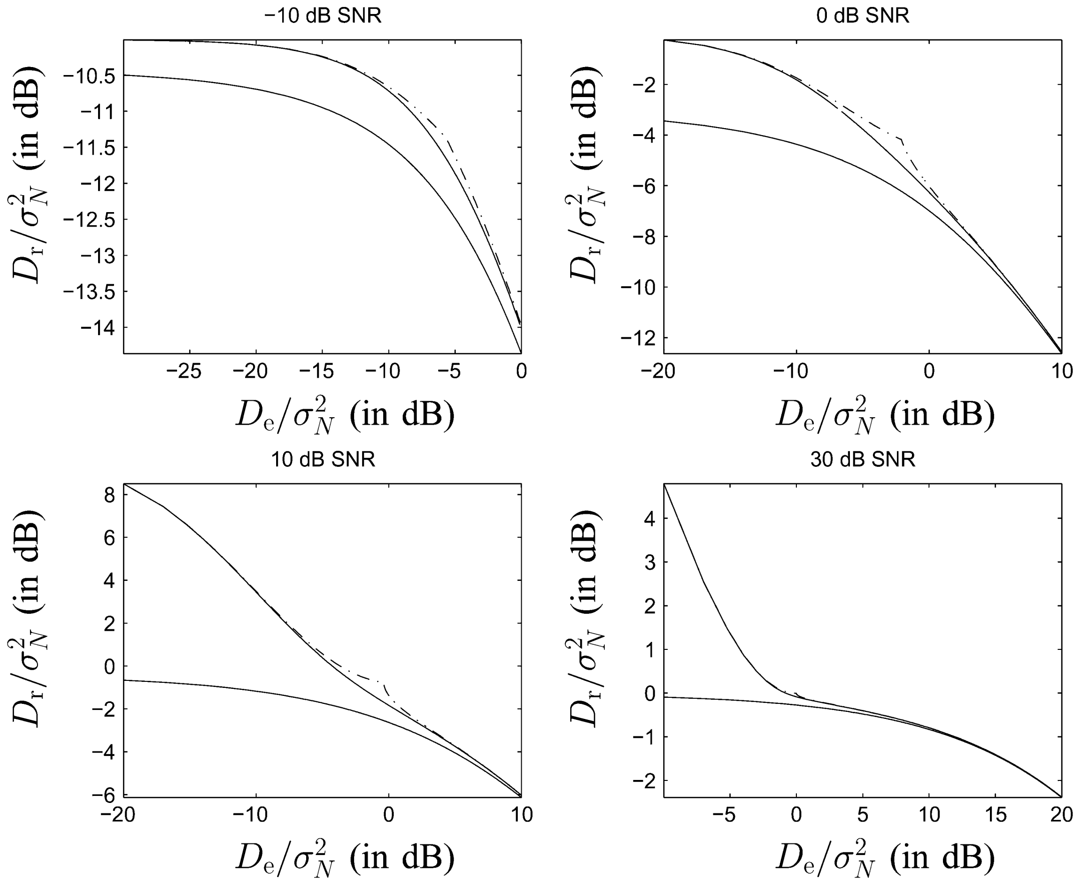

Fig. 5. Bounds on the achievable distortion region for the Gaussian-quadratic problem. The lowest solid curve is the inner bound corresponding to the boundary of the achievable region when reconstructions need not be authentic. The numerically obtained upper solid curve is the outer bound resulting from the use of Gaussian codebooks. The dashed curve corresponds to the lower convex envelope of the simple low and high D_e analytic outer bounds derived in the text.

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 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 $D_{\rm e}$ regimes of operation, requiring authenticity strictly reduces the achievable distortion region. This behavior is analogous to that observed in the binary-Hamming case.

VIII. Comparing Authentication Architectures

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

A. Authentication Systems Based on Robust Watermarking

The robust watermarking approach to encoding for authentication (see, e.g., [4], [10], [11], [15], [16]) takes the form of a quantize- and-embed strategy. The basic steps of the encoding are 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 $p(\hat{s} \mid s)$ and $p(u \mid s)$, 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 [43] and the capacity of information-embedding systems with distortion constraints as developed in [36] as an extension of [19].

Comparing (82) to (9) with $\hat{S} = g(U)$ we see that quantize-and-embed systems are unnecessarily constrained, which translates to a loss of efficiency relative to the optimum joint source–channel-authentication coding system constructions of Section V. This performance penalty can be quite severe in the typical regimes of interest, as we now illustrate. In particular, we quantify this behavior in the two example scenarios considered earlier: the binary-Hamming and Gaussian-quadratic cases.

1) Example: Binary-Hamming Case: In this scenario, the rate-distortion function is [43]

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

while the information embedding capacity is (see [36]) 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 VI, with the performance gap largest for the cleanest reference channels. Moreover, since as we saw in Section III-B clean reference channels correspond to ensuring small encoding and reconstruction distortions, this means that quantize-and-embed systems suffer the largest losses precisely in the regime one would typically want to operate in.

2) Example: Gaussian-Quadratic Case: In this scenario, the rate-distortion function is [43]

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

while the information embedding capacity is [20]

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

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 VII. 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

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

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

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

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

Hence, as in the binary-Hamming case, we see again that quantize-and-embed systems suffer the largest losses in the regime where one is most interested in operating—that where the editor is allowed to make only perturbations small enough that the corresponding encoding and reconstruction distortions are small.[9]

B. Authentication Systems Based on Fragile Watermarking

A fundamentally different approach to the authentication problems of this paper is based on constraining the semantic severity of the modifications the editor is allowed to make. In particular, given a distortion measure that captures the semantic impact of edits to the content, the decoder will declare the edited content authentic if and only if the distortion is below some predetermined threshold. We refer to these as authentication systems based on semantic thresholding.

[9]It should be emphasized that while one could argue that the quadratic distortion measure is a poor measure of semantic proximity in many applications, such reasoning confuses two separate issues. We show here that quantize-and-embed systems are quite poor when the quadratic measure corresponds *exactly* to the semantics of interest. For problems where it is a poor match, one can expect systems based on more accurate measures to exhibit the same qualitative behavior—that quantize-and-embed systems will be least attractive in regimes where the source encodings and reconstructions are constrained to be semantically close to the original source.

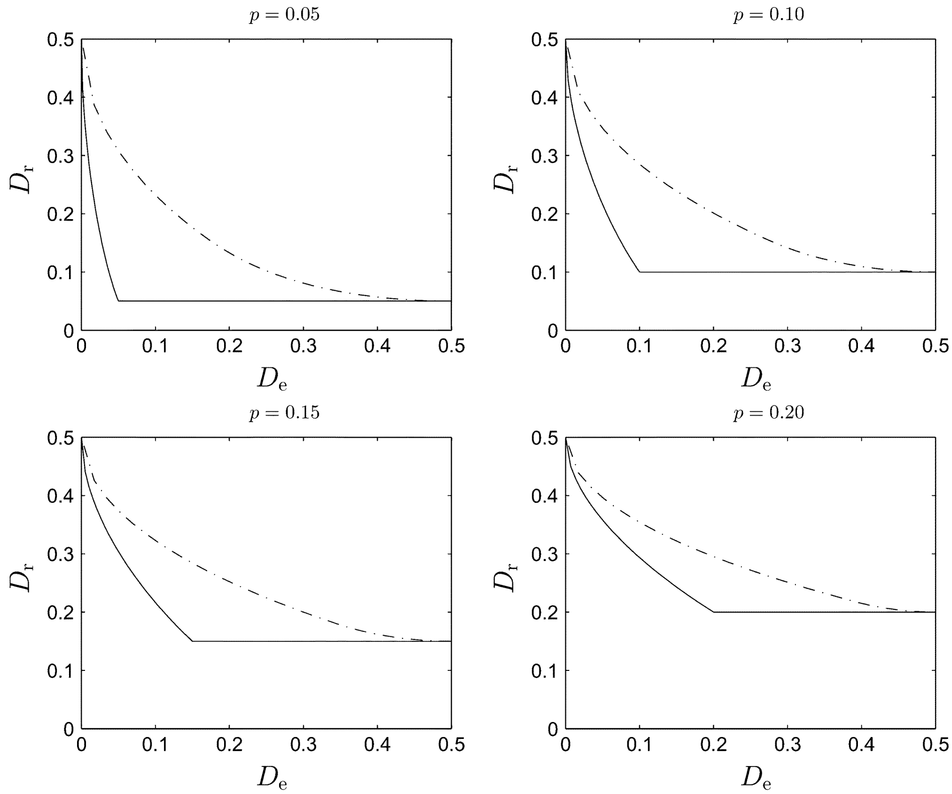

Fig. 6. Performance loss of quantize-and-embed systems for the binary-Hamming scenario with various reference channel crossover probabilities p. The solid curve depicts the boundary of the achievable regions for the optimum system; the dashed curve depicts that of the best quantize-and-embed system.

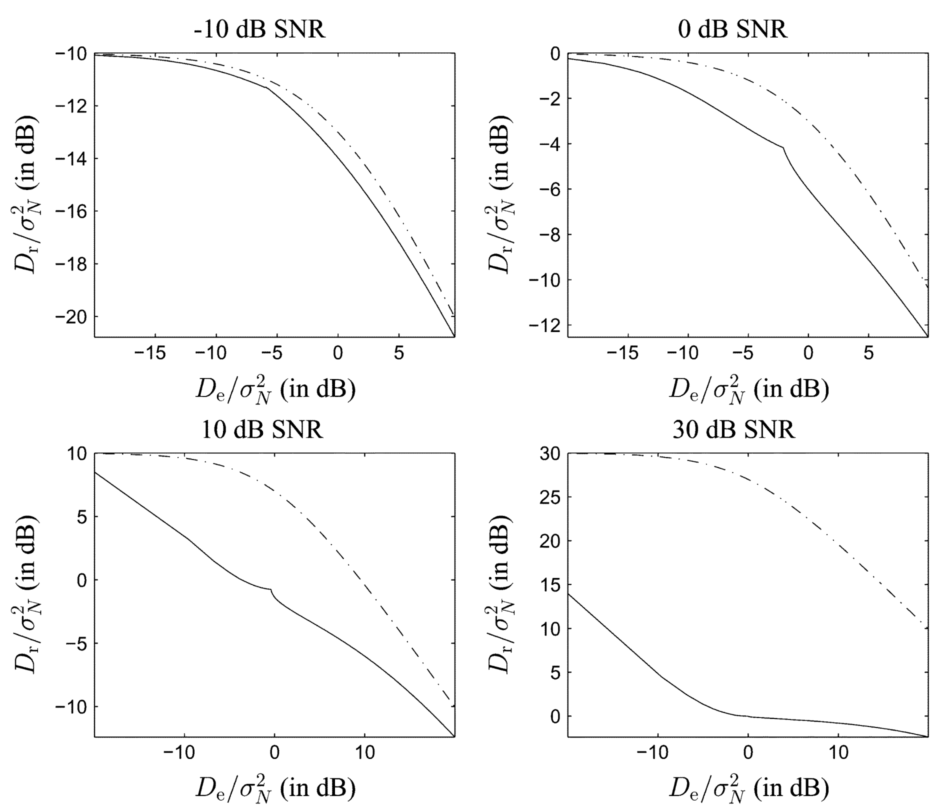

Fig. 7. Performance loss of quantize-and-embed systems for the Gaussian-quadratic scenario at various reference channel SNRs. The solid curve depicts the asymptotic outer bound of the achievable regions for the optimum system; the dashed curve depicts that of the best quantize-and-embed system.

It is important to appreciate that the manner in which the editor is constrained in systems based on semantic thresholding is qualitatively quite different from the way the editor is constrained in the systems developed in this paper. In particular, in our formulation, the editor is contrained according to a reference channel model that can be freely chosen—independently of any semantic model.

While in this section we are primarily interested in discussing the properties of such systems, we first briefly describe how such systems can be designed. We begin by noting that role of the encoder in such systems is to mark the original content so as to enable the eventual decoder to estimate the distortion between the edited content and that original content, despite not having direct access to the latter.

One approach to such a problem would be to use the self-embedding idea discussed in Section VIII-A. In particular, a compressed version of the original content would be embedded into that content so that it could be reliably extracted from the edited content by the decoder and used in the distortion calculation. In practice, such self-embedding can be somewhat resource inefficient, much as it was in the context of Section VIII-A. Instead, an approach based on so-called fragile watermarking is more typically proposed, which allows the decoder to measure the distortion without explicitly being given an estimate of the original content. With this approach, distortion in the known watermark that results from editing the content are used to infer the severity of distortion in the content itself.

Typical implementations of the fragile watermarking approach to encoding for authentication (see, e.g., [5], [7], [13], [14]) take the following form. A watermark message M known only to the encoder and decoder (and kept secret from the editor) is embedded into the source signal by the encoder. The editor's processing of the encoded content indirectly perturbs the watermark. A decoder extracts this perturbed watermark $\hat{M}$, measures the size of the perturbation (e.g., by computing the distortion between $\hat{M}$ and M with respect to some suitable measure), then uses the result to assess the (semantic) severity of the editing the content has undergone. If the severity is below some predetermined threshold, the decoder declares the signal to be authentic.

A detailed information-theoretic characterization of authentication systems based on semantic thresholding is beyond the scope of this paper. However, in the sequel, we emphasize some important qualitative differences in the security characteristics between such schemes and those developed in this paper. In particular, as we now develop, there is a fundamental vulnerability in semantic thresholding schemes that results from their inherent sensitivity to mismatch in the chosen semantic model.

To see this, consider a mismatch scenario in which the authentication system is designed with an incorrect semantic model (distortion measure). If the system is based on semantic thresholding, then an attacker who recognizes the mismatch can exploit this knowledge to make an edit that is semantically significant, but which the system will deem as semantically insignificant due to the model error, and thus accept as authentic. Thus, for such systems, a mismatch can lead to a security failure.

By contrast, for the authentication systems developed in this paper, designing the system based on the incorrect semantic model reduces the efficiency of the system, but does not impact its security. In particular, use of the incorrect semantic model leads to encodings and/or authentic reconstructions with unnecessarily high distortions (with respect to the correct model). However, attackers cannot exploit this to circumvent the security mechanism, since they are constrained by the reference channel, which is independent of the semantic model.

From such arguments, one might conclude that systems based on semantic thresholding might be preferable so long as care is taken to develop accurate semantic models. However, such a viewpoint fails to recognize that in practice some degree of mismatch is inevitable—the high complexity of accurate semantic models makes them inherently difficult to learn. Thus, in a practical sense, authentication systems based on semantic thresholding are intrinsically less secure than those developed in this paper.

IX. Layered Authentication: Broadcast Reference Channels

For many applications, one might be interested in an authentication system with the property that an authentic reconstruction is always produced, but that its quality degrades gracefully with the extensiveness of the editing the content has undergone. In this section, we show that discretized versions of such behavior are possible, and can be built as a natural extension of the formulation of this paper.

To develop this idea, we begin by observing that the systems developed thus far in the paper represent a first-order approximation to such behavior. In particular, for edits consistent with the reference channel model, an authentic reconstruction of fixed quality is produced. When the editing is not consistent with the reference channel, the only possible authentic reconstruction is the minimal quality one obtained from the *a priori* distribution for the content, since the edited version must be ignored altogether. In this section, we show that by creating a hierarchy of reference channels corresponding to increasing amounts of editing, one can create multiple authentication reconstructions. In this way, a graceful degradation characteristic can be obtained to any desired granularity.

Such systems can be viewed as layered authentication systems, and arise naturally out of the use of broadcast reference channel models. With such systems there is a fixed encoding of the source that incurs some distortion. Then, from edited content that is consistent with any of the constituent reference channels in the broadcast model, the 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-user memoryless degraded broadcast channel [43] as our reference channel. This corresponds to a two-layer authentication system. For convenience, we refer to the strong channel as the "mild-edit" one, and the weak channel, which is a degraded version of the strong one, as the "harsh-edit" one. Edits consistent with the mild-edit branch of the reference channel will allow higher quality authentic reconstructions, which we will call "fine," while edits consistent with the harsh-edit branch will allow lower quality authentic reconstructions, which we

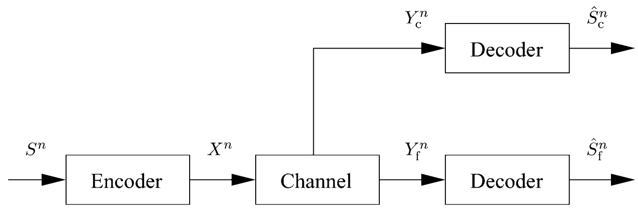

Fig. 8. Two-layer authentication system operation when the reference channel is in effect. From the outputs $Y_{\rm f}$ and $Y_{\rm c}$ of the degraded broadcast reference channel, corresponding to mild and harsh editing, the respective fine and coarse authentic reconstructions $\hat{S}_{\rm f}^n$ and $\hat{S}_{\rm c}^n$ are produced. The common encoding obtained from the source S^n is X^n.

will call "coarse." For edits inconsistent with either branch, the only authentic reconstruction will be one that ignores the edited data, which will be of lowest quality.

In this scenario, for any prescribed level of encoding distortion $D_{\rm e}$, there is a fundamental tradeoff between the achievable distortions $D_{\rm r}^{\rm f}$ and $D_{\rm r}^{\rm c}$ of the corresponding fine and coarse authentic reconstructions, respectively. Of course, $D_{\rm r}^{\rm c} \geq D_{\rm r}^{\rm f}$ will always be satisfied. However, as we will see, achieving smaller values of $D_{\rm r}^{\rm c}$ in general requires accepting larger values of $D_{\rm r}^{\rm f}$ and *vice versa*. Using the ideas of this paper, one can explore the fundamental nature of such tradeoffs.

A. 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 eight-tuple

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

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

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

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

$$D_{\rm r}^{\rm c} = \frac{1}{n}\sum_{i=1}^{n} d_{\rm r}\left(S^n, \hat{S}_{\rm c}^n\right) \tag{93a}$$

$$D_{\rm r}^{\rm f} = \frac{1}{n}\sum_{i=1}^{n} d_{\rm r}\left(S^n, \hat{S}_{\rm f}^n\right). \tag{93b}$$

The following theorem develops tradeoffs 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 $g_{\rm f}(\cdot,\cdot)$, such that

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

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

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

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

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

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

In the interests of brevity, and since it closely parallels that for the single-layer case, we avoid a formal derivation of this result. Instead, we sketch the key ideas of the construction. We also leave determining the degree to which the distortion region can be further extended via more elaborate coding for future work.

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

The encoder first searches $\mathcal{C}_{\rm c}$ for an admissible codeword $c_{\rm c}$ jointly typical with the source and then searches $\mathcal{C}_{\rm f}(c_{\rm c})$ for a refinement $c_{\rm f}$ that is jointly typical with the source. The pair $(c_{\rm c}, c_{\rm f})$ 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 $R_{\rm c} > I(U;S)$ and $R_{\rm f} > I(T;S \,|\, U)$.

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

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 in effect meets the distortion targets provided that $R_{\rm c} < I(U;Y_{\rm c})$ and $R_{\rm f} < I(T;Y_{\rm f} \,|\, U)$. □

B. Example: Gaussian-Quadratic Case

The Gaussian-quadratic case corresponds to the mild- and harsh-edit outputs of the reference channel taking the forms $Y_{\rm f} = X + N$ and $Y_{\rm c} = Y_{\rm f} + V$, 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 Fig. 9, which generalizes that of the single-layer systems developed in Section VII. 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 encoding X^n. If the channel output stays within the darkly shaded sphere centered about T^n, e.g., producing $Y_{\rm f}^n$ as shown, the decoder produces a fine-grain authentic reconstruction from T^n. If the channel output is outside the darkly shaded

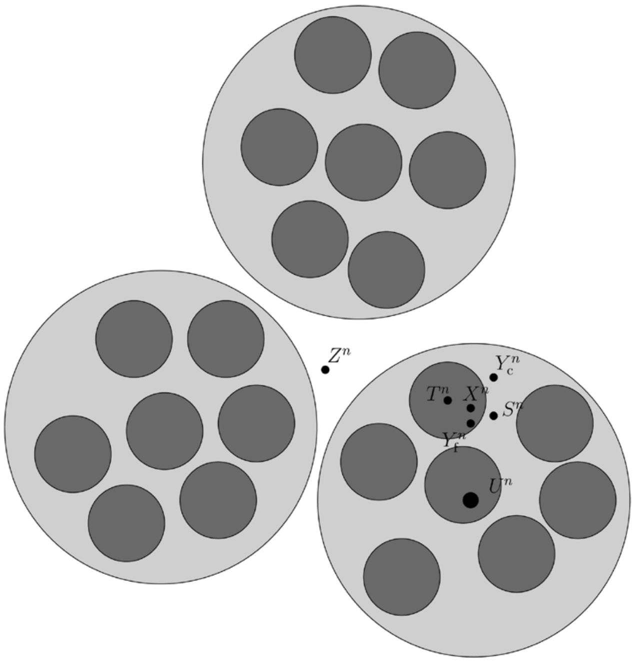

Fig. 9. Illustration of the nested codebook geometry associated with a two-layer authentication system for the Gaussian-quadratic scenario. The centers of large and small shaded spheres correspond to admissible coarse and fine authentic reconstructions, respectively.

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.

An achievable distortion region for this layered authentication scenario is obtained from Theorem 2 with the auxiliary random variables chosen 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

$$D_\mathrm{e} = \sigma_A^2 + \sigma_B^2 \tag{98}$$
$$D_\mathrm{r}^\mathrm{c} = \sigma_S^2\left(1 - \frac{E[SU]^2}{E[S^2]E[U^2]}\right) = \frac{\sigma_S^2\sigma_A^2}{\sigma_A^2 + \alpha^2\sigma_S^2} \tag{99}$$
$$\begin{aligned} D_\mathrm{r}^\mathrm{f} &= \sigma_S^2 - \Lambda_{S,[UT]}\Lambda_{[UT]}^{-1}\Lambda_{[UT],S} \\ &= \frac{\sigma_S^2\sigma_A^2\sigma_B^2}{\beta^2\sigma_S^2\sigma_A^2 + \sigma_A^2\sigma_B^2 + \alpha^2\sigma_S^2\sigma_B^2} \end{aligned} \tag{100}$$

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 immediately apply the arguments from Section VII-B to obtain

$$\begin{aligned} &I(U;Y_\mathrm{c}) - I(S;U) \\ &\quad = \frac{1}{2}\log\frac{\sigma_A^2\left(\sigma_A^2+\sigma_S^2+\sigma_N^2+\sigma_V^2+\sigma_B^2\right)}{\sigma_A^2\sigma_S^2(1-\alpha)^2 + \left(\sigma_N^2+\sigma_V^2+\sigma_B^2\right)\left(\sigma_A^2+\alpha^2\sigma_S^2\right)}. \end{aligned} \tag{101}$$

From this we can solve for α as in the single-layer case of Section VII-B 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

$$\begin{aligned} &I(S;T\,|\,U) - I(T;Y_\mathrm{f}\,|\,U) \\ &\quad = H(T\,|\,U,Y_\mathrm{f}) - H(T\,|\,U,S) \\ &\quad = H(T,U,Y_\mathrm{f}) - H(U,Y_\mathrm{f}) - H(T,U,S) + H(U,S) \end{aligned} \tag{102}$$

we see that (94b) implies

$$\frac{\det(\Lambda_{[TUY_\mathrm{f}]})}{\det(\Lambda_{[UY_\mathrm{f}]})} \leq \frac{\det(\Lambda_{[TUS]})}{\det(\Lambda_{[US]})}. \tag{103}$$

By varying σ_A^2, σ_B^2, and β such that (103) 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_r^f and D_r^c for various values of the encoding distortion D_e. Note that it follows from our single-layer inner bounds that for a particular choice of encoding distortion D_e, the achievable tradeoffs between D_r^c and D_r^f are contained within the region

$$D_\mathrm{r}^\mathrm{c} \geq \frac{\sigma_S^2\left(\sigma_N^2+\sigma_V^2\right)}{\sigma_N^2+\sigma_V^2+\left(\sqrt{D_\mathrm{e}}+\sigma_S\right)^2} \tag{104}$$
$$D_\mathrm{r}^\mathrm{f} \geq \frac{\sigma_S^2\sigma_N^2}{\sigma_N^2+\left(\sqrt{D_\mathrm{e}}+\sigma_S\right)^2} \tag{105}$$

where obviously the lower bound of (105) is smaller than that of (104).

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 harsh-edit reference channel, and for the remaining time uses a codebook appropriate for the mild-edit reference channel. When the harsh-edit 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 mild-edit 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). However, as Fig. 10 also illustrates, this approach is in general quite inefficient: the use of such time-sharing results in a substantially smaller achievable region.

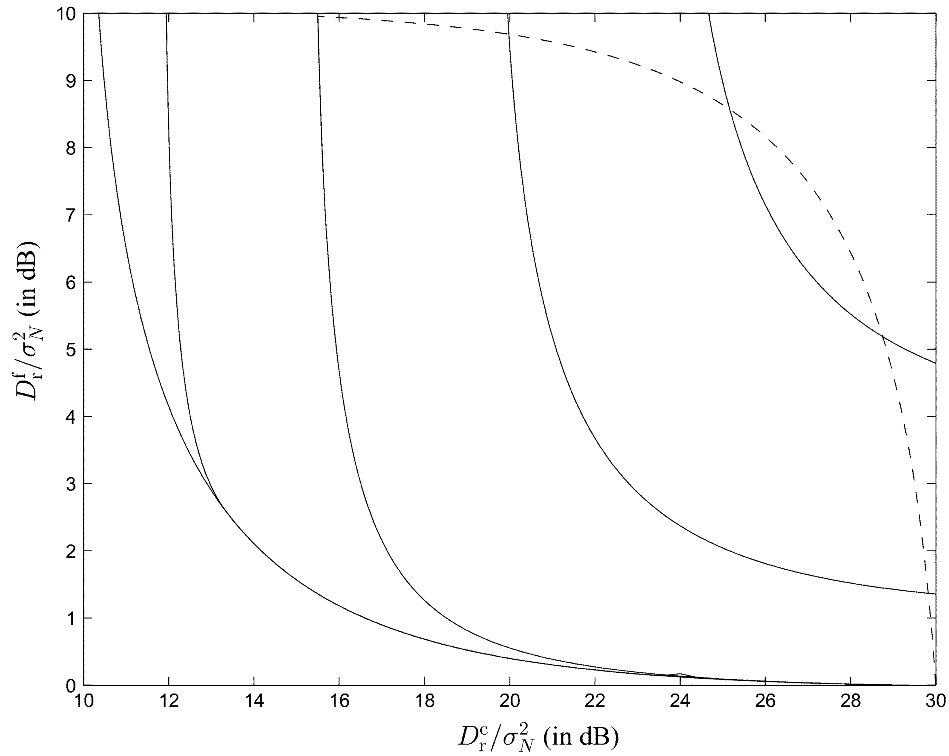

Fig. 10. Achievable fine and coarse quality reconstruction distortion pairs (D_r^f, D_r^c) in a layered authentication system for the Gaussian-quadratic case with $\sigma_S^2/\sigma_N^2 = 30$ dB, $\sigma_V^2/\sigma_N^2 = 10$ dB, and $\sigma_N^2 = 1$. From left to right, the curves are the boundaries of achievable distortion regions corresponding to encoding distortions of $D_e/\sigma_N^2 =$ 10, 5, 0, -5, -10 dB. The dashed curve corresponds to time-sharing between two operating points on the $D_e/\sigma_N^2 = 0$ dB curve.

X. Concluding Remarks

This paper develops one meaningful formulation for authentication problems in which the content may undergo a variety of types of legitimate editing prior to authentication. As part of this formulation, we adopt a particular formal notion of security in such settings. For such a formulation, and with the simplest classes of models, we establish that secure authentication systems can be constructed, and subsequently analyze their fundamental performance limits. From these models, we further develop how such systems offer significant advantages over other proposed solutions.

Many opportunities for further research remain. For example, extensions of the main results to richer content, semantic, and edit models may provide additional insights into the behavior of such systems. It would also be useful to understand the degree to which robust and/or universal solutions exist for the problem; such approaches seek to avoid requiring accurate prior model knowledge during system design.

There are additional opportunities to further refine the analysis even for the existing models. For example, characterizing the manner in which asymptotic limits are approached—for example, via error exponents—would provide useful engineering insights. Likewise, further analyzing public-key formulations, in which edits are more generally subject to computational constraints, could also be revealing. From this perspective, the Appendix represents but a starting point.

More generally, identifying and relating other meaningful notions of security for such problems would be particularly useful in putting the results of this paper in perspective. For example, a broader unifying framework for characterizing and comparing different notions of security could provide a mechanism for selecting a formulation best matched to the social needs and/or engineering constraints at hand.

Finally, there are many interesting questions about how to best approach the development of practical authentication systems based on these ideas. These include questions of customized code design and implementation, but also architectural issues concerning the degree to these systems can be built from interconnections of existing and often standardized components—i.e., existing compression systems, error-control codes, and public-key cryptographic tools.

Appendix
A Public-Key Adaptation of the Private-Key Authentication System Model

To simplify the analysis, we have focussed on private-key systems where the encoder and decoder share a secret key θ, which is kept hidden from editors. In most practical applications, however, it is more convenient to use public-key systems where a public key θ_p is known to all parties (including editors) 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 appendix, we briefly describe how a secret key authentication system can be combined with a generic digital signature scheme to yield a public-key system. Some additional aspects of such an implementation are discussed in, e.g., [49], [50].

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 sublinear in n or even finite.

Modified Encoder:

1) The public key of the digital signature scheme is published, and there is no secret key (equivalently, the secret key in the our original formulation is simply published).
2) The encoder uses the original authentication system to map the source S^n to $\tilde{X}^n = \Upsilon_n(S^n)$.
3) For a system like the one described in Section V-A, 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 bitwise representation $c(\hat{S}^n)$, from which the encoder computes the digital signature tag $\tau = \mathcal{S}(c(\hat{S}^n), \theta_s)$ using the digital signature algorithm.
4) Finally, the signature τ is embedded into $\tilde{X}^n$, producing X^n, using an information-embedding (data-hiding) algorithm. The chosen algorithm can be quite crude since τ only requires a sublinear number of bits. The algorithm parameters are chosen to that the embedding incurs asymptotically negligible additional distortion to the overall encoding process.

Modified Decoder:

1) The decoder extracts from Y^n an estimate $\hat{\tau}$ of the embedded signature τ. Since the size of τ is sublinear, the embedding algorithm parameters can be further chosen so that $\hat{\tau} = \tau$ with arbitrarily high probability when the reference channel is in effect.
2) Next, the decoder uses the original authentication system to produce $\tilde{S}^n = \Phi_n(Y^n)$, and then, in turn, its bitwise representation $c(\tilde{S}^n)$.
3) The decoder checks whether the digital signature verifying algorithm $\mathcal{V}(c(\tilde{S}^n), \hat{\tau}, \theta_p)$ accepts the $\tilde{S}^n$ as valid.
4) If so, then the decoder produces the authentic reconstruction $\hat{S}^n = \tilde{S}^n$. Otherwise, the decoder produces the special symbol $\emptyset$, declaring that it is unable to authenticate.

With this construction, we see that the security of such a system is determined by the security of the underlying public-key digital signature scheme used. Specifically, the only way an attacker can defeat the system is to find a matching $\hat{S}^n$ and τ accepted by the digital signature-verifying algorithm. All other performance aspects of the system are effectively unchanged.

ACKNOWLEDGMENT

The authors are grateful to Prof. Ram Zamir for many helpful suggestions including improvements to the proof of the converse part of Theorem 1. The authors would also like to thank the reviewers and Associate Editor 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. Inf. Theory*, vol. IT-22, no. 6, pp. 644–654, Nov. 1976.
[2] F. A. P. Petitcolas, R. J. Anderson, and M. G. Kuhn, "Information hiding—A survey," *Proc. IEEE*, vol. 87, no. 7, pp. 1062–1078, Jul. 1999.
[3] J. Fridrich, "Methods for tamper detection in digital images," in *Proc. Multimedia and Security Workshop at ACM Multimedia*, Orlando, FL, Oct. 1999, pp. 19–23.
[4] C. Rey and J.-L. Dugelay, "Blind detection of malicious alterations on still images using robust watermarks," in *Proc. IEE Seminar on Secure Images and Image Authentication*, Apr. 2000, pp. 7/1–7/6.
[5] R. B. Wolfgang and E. J. Delp, "A watermark for digital images," in *Proc. Int. Conf. Image Processing (ICIP)*, vol. 3, Sep. 1996, pp. 219–222.
[6] G. L. Friedman, "The trustworthy digital camera: Restoring credibility to the photographic image," *IEEE Trans. Consumer Electron.*, vol. 39, no. 4, pp. 905–910, Nov. 1993.
[7] D. Kundur and D. Hatzinakos, "Digital watermarking for telltale tamper proofing and authentication," *Proc. IEEE*, vol. 87, no. 7, pp. 1167–1180, Jul. 1999.
[8] P. W. Wong, "A public key watermark for image verification and authentication," in *Proc. Int. Conf. Image Processing (ICIP)*, vol. 1, Oct. 1998, pp. 445–459.
[9] M. Wu and B. Liu, "Watermarking for image authentication," in *Proc. Int. Conf. Image Processing (ICIP)*, vol. 2, Oct. 1998, pp. 437–441.
[10] M. P. Queluz, "Toward robust, content based techniques for image authentication," in *Proc. Workshop Multimedia Signal Processing (MMSP)*, Los Angeles, CA, Dec. 1998, pp. 297–302.
[11] S. Bhattacharjee and M. Kutter, "Compression tolerant image authentication," in *Proc. Int. Conf. Image Processing (ICIP)*, vol. 1, Oct. 1998, pp. 435–439.
[12] B. Macq and J.-L. Dugelay, "Watermarking technologies for authentication and protection of images," *Ann. Telecommun.*, vol. 55, no. 3–4, pp. 92–100, Mar.–Apr. 2000.
[13] J. J. Eggers and B. Girod, "Blind watermarking applied to image authentication," in *Proc. Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, vol. 3, Salt Lake City, UT, May 2001, pp. 1977–1980.
[14] M. M. Yeung and F. Mintzer, "An invisible watermarking technique for image verification," in *Proc. Int. Conf. Image Processing (ICIP)*, vol. 2, Oct. 1997, pp. 680–683.
[15] M. Schneider and S. Chang, "A robust content based digital signature for image authentication," in *Proc. Int. Conf. Image Processing (ICIP)*, vol. 3, Sep. 1996, pp. 227–230. .
[16] 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.
[17] L. Me and G. R. Arce, "A class of authentication digital watermarks for secure multimedia communication," *IEEE Trans. Image Process.*, vol. 10, no. 11, pp. 1754–1764, Nov. 2001.
[18] C.-S. Lu and H. Liao, "Multipurpose watermarking for image authentication and protection," *IEEE Trans. Image Process.*, vol. 10, no. 10, pp. 1579–1592, Oct. 2001.
[19] S. I. Gel'fand and M. S. Pinsker, "Coding for channel with random paramters," *Probl. Contr. Inf. Theory*, vol. 9, no. 1, pp. 19–31, 1983.
[20] M. H. M. Costa, "Writing on dirty paper," *IEEE Trans. Inf. Theory*, vol. IT-29, no. 3, pp. 439–441, May 1983.
[21] C. Heegard and A. A. El Gamal, "On the capacity of computer memory with defects," *IEEE Trans. Inf. Theory*, vol. IT-29, no. 5, pp. 731–739, Sep. 1983.
[22] J. A. O'Sullivan, P. Moulin, and J. M. Ettinger, "Information-theoretic analysis of steganography," in *Proc. IEEE Int. Symp. Information Theory*, Cambridge, MA, Aug. 1998, p. 297.
[23] B. Chen and G. W. Wornell, "Quantization index modulation: A class of provably good methods for digital watermarking and information embedding," in *Proc. IEEE Int. Symp. Information Theory*, Sorrento, Italy, Jun. 2000, p. 46.
[24] A. Cohen and A. Lapidoth, "On the Gaussian watermarking game," in *IEEE Int. Symp. Inform. Theory*, Sorrento, Italy, Jun. 2000, p. 48.

[25] P. Moulin and J. O'Sullivan, "Information-theoretic analysis of information hiding," in *IEEE Int. Symp. Information Theory*, Sorrento, Italy, Jun. 2000, p. 19.

[26] B. Chen and G. W. Wornell, "Quantization index modulation: A class of provably good methods for digital watermarking and information embedding," *IEEE Trans. Inf. Theory*, vol. 47, no. 4, pp. 1423–1443, May 2001.

[27] P. Moulin and J. A. O'Sullivan, "Information-theoretic analysis of information hiding," *IEEE Trans. Inf. Theory*, vol. 49, no. 3, pp. 563–593, Mar. 2003.

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

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

[30] M. D. Swanson, M. Kobayashi, and A. H. Tewfik, "Multimedia data-embedding and watermarking technologies," *Proc. IEEE*, vol. 86, no. 6, pp. 1064–1087, Jun. 1998.

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

[32] 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.

[33] J. Chou, S. S. Pradhan, and K. Ramchandran, "On the duality between distributed source coding and data hiding," in *Proc. Asilomar Conf. Signals, Systems, and Computers*, vol. 2, Pacific Grove, CA, Oct. 1999, pp. 1503–1507.

[34] S. S. Pradhan, J. Chou, and K. Ramchandran, "Duality between source and channel coding and its extension to the side information case," *IEEE Trans. Inf. Theory*, vol. 49, no. 5, pp. 1181–1203, May 2003.

[35] 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, and Computers*, vol. 2, Pacific Grove, CA, Nov. 2000, pp. 1841–1845.

[36] R. J. Barron, B. Chen, and G. W. Wornell, "The duality between information embedding and source coding with side information and some applications," *IEEE Trans. Inf. Theory*, vol. 49, no. 5, pp. 1159–1180, May 2003.

[37] 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. IEEE Int. Symp. Information Theory*, Washington, DC, Jun. 2001, p. 300.

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

[39] M. Chiang and T. M. Cover, "Unified duality of channel capacity and rate distortion with state information," in *Proc. IEEE Int. Symp. Information Theory*, Washington, DC, Jun. 2001, p. 301.

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

[41] R. Zamir, S. Shamai (Shitz), and U. Erez, "Nested linear/lattice codes for structured multiterminal binning," *IEEE Trans. Inf. Theory*, vol. 49, no. 6, pp. 1250–1278, Jun. 2002.

[42] A. Sutivong, T. Cover, M. Chiang, and Y.-H. Kim, "Rate vs. distortion trade-off for channels with state information," in *Proc. IEEE Int. Symp. Information Theory*, Lausanne, Switzerland, Jun./Jul. 2002, p. 226.

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

[44] S. Verdú and T. S. Han, "A general formula for channel capacity," *IEEE Trans. Inf. Theory*, vol. 40, no. 4, pp. 1147–1157, Jul. 1994.

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

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

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

[48] S. Shamai (Shitz), S. Verdú, and R. Zamir, "Systematic lossy source/channel coding," *IEEE Trans. Inf. Theory*, vol. 44, no. 2, pp. 564–579, Mar. 1998.

[49] E. Martinian, "Authenticating multimedia in the presence of noise," Master's thesis, MIT, Cambridge, MA, 2000.

[50] 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)*, San Jose, CA, Jan. 2001.

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

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