The Capacity of the Vector Gaussian Watermarking Game

Aaron S. Cohen[1]
Massachusetts Inst. of Technology
77 Mass. Ave., 35-303
Cambridge, MA 02139
e-mail: acohen@mit.edu

Amos Lapidoth
Swiss Federal Inst. of Technology
ETH-Zentrum
CH-8092, Zurich, Switzerland
e-mail: lapidoth@isi.ee.ethz.ch

Abstract — **We compute the coding capacity of the watermarking game for a vector Gaussian covertext and squared-error distortions. As with a scalar Gaussian covertext [1], the capacity does not depend on knowledge of the covertext at the decoder. Unlike the scalar version, an attacker based on the rate distortion solution (i.e. optimal compression) is suboptimal.**

I. Introduction

The watermarking game [1, 2] can model a situation where an original source sequence ("covertext") needs to be copyright-protected before it is distributed to the public. The copyright ("message") needs to be embedded in the distributed version ("stegotext") so that no "attacker" with access to the stegotext will be able produce a "forgery" that resembles the covertext and yet does not contain the embedded copyright message.

Different messages may correspond to different possible owners, versions, dates, etc. of the covertext, and it is thus of interest to study the number of distinct messages that can be embedded (subject to an encoding distortion constraint) if reliable decoding is required from any reasonable forgery. The highest exponential rate at which this number can grow in relation to the covertext size is the coding value of the game.

Our previous study [1] of IID scalar Gaussian covertexts is extended here to IID vector Gaussian covertexts. These results provide a stepping stone to the study of general Gaussian covertexts [3].

II. Watermarking model

The vector Gaussian watermarking game can be described as follows. A source emits a *covertext* $\boldsymbol{U}$ consisting of n IID zero-mean Gaussian random vectors, each with $m \times m$ covariance matrix $K_{\boldsymbol{u}}$. (The blocklength n is allowed to grow, while the vector size m stays fixed.) Independently of $\boldsymbol{U}$, a *message* W is drawn uniformly over the set $\{1, \ldots, \lfloor 2^{nR} \rfloor\}$, where R is the *rate* of the system.

Using a secret key Θ_1, which is independent of $\boldsymbol{U}$ and W, the *encoder* produces the *stegotext* $\boldsymbol{X} = \boldsymbol{X}(\boldsymbol{U}, W, \Theta_1) \in \mathbb{R}^{n \times m}$. We require the encoder to satisfy $\frac{1}{n}\sum_{i,j}(X_{ij} - U_{ij})^2 \leq \Delta_1$, a.s., where i and j range from 1 to n and m, respectively, $\Delta_1 > 0$ is a given constant called the *encoder distortion level*, and a.s. stands for "almost surely". The *attacker*, which is assumed to be ignorant of $\boldsymbol{U}$, W and Θ_1, produces a *forgery* $\boldsymbol{Y} = \boldsymbol{Y}(\boldsymbol{X}, \Theta_2) \in \mathbb{R}^{n \times m}$ based on $\boldsymbol{X}$ and its own attack key Θ_2. We similarly require the attacker to satisfy $\frac{1}{n}\sum_{i,j}(Y_{ij} - X_{ij})^2 \leq \Delta_2$, a.s., where $\Delta_2 > 0$ is a given constant called the *attacker distortion level.* The *decoder* produces an estimate of the message $\hat{W}$. In the *public version* of the game, the decoder only uses the encoder's secret key and the forgery, so that $\hat{W} = \hat{W}(\boldsymbol{Y}, \Theta_1)$. In the *private version* of the game, the decoder also uses the covertext, so that $\hat{W} = \hat{W}(\boldsymbol{Y}, \Theta_1, \boldsymbol{U})$. We consider the probability of error averaged over the covertext, message and both sources of randomness, which is written $\bar{P}_e(n) = \Pr(\hat{W} \neq W)$.

We adopt a conservative approach to the watermarking game and assume that the attacker knows the details of the encoder and decoder (but not the realizations of $\boldsymbol{U}$, W and Θ_1). Conversely, the encoder and decoder have no knowledge of the attacker, and in particular, how the attacker will distribute its distortion. We thus say that rate R is *achievable* if there exists a sequence of allowable rate-R encoder and decoder pairs such that for any sequence of allowable attackers, $\bar{P}_e(n)$ tends to zero as n tends to infinity. The *coding capacity* of the game is the supremum of all achievable rates.

Theorem. *The coding capacity of the vector Gaussian watermarking game (private and public versions) is given by*

$$\max_{\substack{\boldsymbol{D}_1 \geq 0, \\ \sum_{j=1}^m D_{1j} \leq \Delta_1}} \; \min_{\substack{\boldsymbol{D}_2 \geq 0, \\ \sum_{j=1}^m D_{2j} \leq \Delta_2}} \; \sum_{j=1}^m C^*(D_{1j}, D_{2j}, \sigma_j^2), \qquad (1)$$

where $C^*(D_1, D_2, \sigma^2)$ is the capacity of the scalar Gaussian watermarking game [1] and $\sigma_1^2, \ldots, \sigma_m^2$ are the eigenvalues of the covariance matrix $K_{\boldsymbol{u}}$.

To better understand our main result, let us assume that $K_{\boldsymbol{u}}$ is diagonal so that $\boldsymbol{U}$ consists of m streams, each a length-n sequence of IID zero-mean Gaussian random variables with respective variances $\sigma_1^2, \ldots, \sigma_m^2$. After choosing vectors $\boldsymbol{D}_1$ and $\tilde{\boldsymbol{D}}_2$, the encoder encodes stream j using the scalar encoder of [1] based on D_{1j} and $\tilde{D}_{2j}$. Every attacker is associated with a feasible $\boldsymbol{D}_2$ (not necessarily equal to $\tilde{\boldsymbol{D}}_2$), where D_{2j} describes the amount of distortion the attacker adds to stream j. For the optimal choice of $\tilde{\boldsymbol{D}}_2$, the attacker will choose $\boldsymbol{D}_2 = \tilde{\boldsymbol{D}}_2$ in order to minimize the achievable rates. We next note that the max and min in (1) cannot be switched and thus the order of the game remains critical, even if the encoder and attacker are constrained to use optimal scalar strategies on each of the streams. Finally, the attacker's optimal distortion distribution is not given by reverse water-filling on the powers of the components of $\boldsymbol{X}$. Thus, unlike in the scalar version, an attacker based on the rate distortion solution allows rates larger than capacity to be achieved.

References

[1] A. Cohen and A. Lapidoth, "The Gaussian watermarking game," to appear in *IEEE Trans. Inform. Theory.*

[2] J. O'Sullivan, P. Moulin, and J. Ettinger, "Information theoretic analysis of steganography," in *Proc. of ISIT*, 1998.

[3] A. Cohen and A. Lapidoth, "The watermarking capacity of Gaussian sources with memory," in preparation.

[1] This research was supported in part by a NSF Graduate Fellowship (A. Cohen) and by the NSF Faculty Early Career Development (CAREER) program (A. Lapidoth).

0-7803-7123-2/01/$10.00 ©2001 IEEE