On the Gaussian Watermarking Game

Aaron Cohen
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 value of the watermarking game for a Gaussian covertext and squared-error distortions. Both the public version of the game (covertext known to neither attacker nor decoder) and the private version of the game (covertext unknown to attacker but known to decoder) are treated. Surprisingly, the two versions yield identical values.

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 to produce a "forgery" that resembles the covertext and yet does not contain the embedded copyright message. The watermarking process ("encoding") should, of course, introduce little distortion so as to guarantee that the stegotext closely resembles the original covertext.

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 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. A precise statement of this problem and some proofs can be found in [3].

II. WATERMARKING MODEL

The watermarking game can be described as follows. A source emits the zero-mean variance-$\sigma^2$ IID length-$n$ covertext sequence $U$. Independently of $U$, a copyright message $W$ is drawn uniformly over the set $W_n = \{1, \ldots, [2^nA]\}$, where $R$ is the rate of the system.

Using a secret key $\Theta_1$, which is independent of $U$ and $W$, the encoder produces the stegotext $X = X(U, W, \Theta_1) \in \mathbb{R}^n$. We require the encoder to satisfy $\frac{1}{n} \|X - U\|^2 \leq D_1$, a.s., where $D_1 > 0$ 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 $U$ and $\Theta_1$, produces a forgery $Y = Y(X, \Theta_2) \in \mathbb{R}^n$ based on $X$ and its own attack key $\Theta_2$. We similarly require the attacker to satisfy $\frac{1}{n} \|Y - X\|^2 \leq D_2$, a.s., where $D_2 > 0$ 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 $\hat{W} = \hat{W}(Y, \Theta_1)$.

In the private version of the game, the decoder also uses the covertext, so that $\hat{W} = \hat{W}(Y, \Theta_1, U)$. We consider the probability of error averaged over the covertext, message and both sources of randomness, which is written $\hat{P}_e(n) = \Pr(\hat{W} \neq W)$.

We adopt a conservative approach to the watermarking system and assume that once the watermarking system is employed, its details are made available to the attacker. The attacker can thus optimize for the encoder and decoder. This precludes the decoder from using the maximum-likelihood decoding rule. 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, $\hat{P}_e(n)$ tends to zero as $n$ tends to infinity.

The value of the game is called the coding capacity, and it is the supremum of all achievable rates. We write the coding capacity as $C_{\text{priv}}(D_1, D_2, \sigma^2_1)$ and $C_{\text{pub}}(D_1, D_2, \sigma^2_2)$ for the private and public versions of the game, respectively.

Theorem 1. For the Gaussian watermarking game,

$$C_{\text{pub}}(D_1, D_2, \sigma^2_2) = C_{\text{priv}}(D_1, D_2, \sigma^2_2).$$

If the interval

$$A(D_1, D_2, \sigma^2_2) = \left[ \max \left\{ D_2, (\sigma_1 - \sqrt{D_1})^2 \right\}, (\sigma_1 + \sqrt{D_1})^2 \right],$$

is empty, then $C_{\text{priv}}(D_1, D_2, \sigma^2_2)$ is zero. Otherwise,

$$C_{\text{priv}}(D_1, D_2, \sigma^2_2) = \max_{A \in A(D_1, D_2, \sigma^2_2)} \frac{1}{2} \log \left( 1 + \frac{1 - \frac{1}{A}}{\frac{1}{A} - \frac{1}{D_2}} \right) \left( \frac{1}{A} - \frac{(\sigma_1^2 + D_1)^2}{4\sigma^2_2} \right).$$

If expected rather than a.s. distortion constraints are used, then the coding capacity for both versions is zero.

Note that the optimal $A$ is a root of a cubic equation and hence a closed form solution for the capacity exists. Different capacity results for yet another version of this game with expected distortion constraints and a decoder that knows the attack strategy (ML decoder) have been recently reported in [1].

REFERENCES

