Optimum Perfect Steganography of Memoryless Sources as a Rate-Distortion Problem

Félix Balado and David Haughton

School of Computer Science and Informatics, University College Dublin
Belfield Campus, Dublin 4, Ireland
felix@ucd.ie, david.haughton@ucdconnect.ie

***Abstract*—Slepian's variant I permutation coding has been recently shown to be a fundamental steganographic tool, as it implements optimum perfect steganography of memoryless sources. Although real host signals are not memoryless, a decorrelating energy-preserving transform can always be applied before a method that assumes a memoryless source, as is usually done in the dual problem of source coding. A further constraint is needed in practice: the information-carrying signal must be close to the host, according to some distance measure. Thus steganography of memoryless sources using permutation coding is a rate-distortion problem. Here we delve deeper in the study of the embedding distortion of permutation coding, and we show that the rate-distortion tradeoff for partitioned permutation coding is near-optimum according to the Gel'fand and Pinsker capacity formula.**

I. Introduction

According to Cachin's criterion [1] perfect steganography is implemented by preserving the distribution of a host signal. If the warden is passive (i.e., the channel is noiseless), then fulfilling this criterion is all that is required by an ideal steganographic algorithm. In practice, Cachin's criterion raises an important question: what is the distribution to be preserved when using a real signal as the host? The current consensus in steganography research is that the distribution of real signals is incognisable [2]. Although this seems reasonable, taking this claim literally implies that only heuristic steganographic methods are feasible, at best loosely inspired in Cachin's criterion. In fact, a heuristic path based on exploiting embedding distortion functions that perform well under machine-learning detection has been widely followed in mainstream steganography research in recent times. But no matter how successful this approach may be in the short term, it will always leave fundamental questions unanswered. Most significantly, it cannot unambiguously engage with the theory of channel coding with side information at the encoder —which is essential to determine the optimum steganographic rate— nor with a wealth of key results from source coding.

It is actually source coding, which has been acknowledged as a dual problem of steganography since the early days of this discipline [1], that lays bare the trouble with taking the incognisability claim above at face value. If this claim were absolutely true, then the following issue would be faced by source coding: an optimum encoder must rely on the distribution of the signal to be compressed, but if the distribution of real signals were completely incognisable then they could only be compressed using heuristic algorithms. However, as evinced by universal (or distribution-free) source coding [3], the pragmatic answer to this question is that the empirical distribution of the signal can always act as a proxy for its underlying distribution. Similarly, universal steganography (i.e., oriented to preserving the empirical host distribution) is the matter-of-fact approach to perfect steganography, as already noted by Cachin [1]. Distribution-free steganography is particularly appealing from the embedder's viewpoint, because if the empirical distribution of the host is preserved then the warden can never exploit a better host model.

Developing near-optimum universal source coding or steganographic methods for signals with memory is a tall order, since this requires taking into account all higher-order statistics (intersample dependencies). However if the signal is memoryless (i.e., formed by independently drawn samples) then both optimum lossless source coding and optimum perfect steganography need only consider its first-order statistics. In the source coding field, arithmetic coding implements near-optimum lossless (i.e., with near-maximum compression rate and with perfect decompression) source coding of memoryless sources [3] using adaptive estimation of the empirical first-order statistics. In the steganography field, we showed in [4] that Slepian's variant I permutation coding [5] implements optimum perfect (i.e., with maximum embedding rate and histogram-preserving) steganography of memoryless sources, and, underlining the duality between steganography and source coding, we gave a near-optimum practical implementation of permutation coding based on arithmetic coding and adaptive estimation of the empirical first-order statistics. The compression rate and the embedding rate are in both cases the entropy of the signal to be compressed or to be used as a host.

Real sources are not memoryless. Conspicuously, it is well known that arithmetic coding does a poor job of compressing signals with memory when relying on a memoryless adaptive model. For this reason, lossless compression schemes proceed in two steps: 1) application of a reversible energy-preserving decorrelating transform; 2) compression in the transformed domain using a near-optimum lossless source coding algorithm for memoryless sources. The same procedure is essentially applicable in steganography, where step 2) now entails optimum

WIFS'2013, November 18-21, 2013, Guangzhou, China.
ISBN 978-1-4673-5593-3

perfect steganography of memoryless sources (permutation coding). The decorrelating transform must map integers to integers for reversibility; the reversible Karhunen-Loève transform (RKLT) [6] would appear to be the best choice, although context-based prediction may be a more practical approach to decorrelation. This two-step procedure decouples the problem of optimum perfect steganography (which is addressed here) from the problem of reversible decorrelation, thus enabling a systematic approach to steganography of real signals.

A further constraint is needed in practice in universal steganography: the information-carrying signal must be close to the host signal, according to some distance measure. This distortion constraint, not considered in Cachin's proposal for universal ϵ-secure steganography [1], is necessary in order to approximately preserve the semantics of the host, and thus it implicitly follows from universal steganography relying on an empirical estimate of the host distribution[1]. Hence steganography of memoryless sources using permutation coding is a rate-distortion problem—insofar as devoid of detectability concerns. In this paper we extend the study of the embedding distortion of permutation coding that was started in [4], taking a closer look at its asymptotics and at its geometric aspects. We also show that partitioned permutation coding (introduced in [4]) allows for a near-optimum rate-distortion tradeoff, according to the Gel'fand and Pinsker capacity formula.

Notation and framework. Boldface lowercase Roman letters are column vectors. The special symbols $\mathbf{1}$ and $\mathbf{0}$ are the all-ones vector and the null vector, respectively. Capital Greek letters denote matrices; the entry at row i and column j of matrix Π is $(\Pi)_{i,j}$. For the sake of keeping standard conventions, the only two exceptions to this notation are the identity matrix I and the exchange matrix J (defined later). $(\cdot)^t$ denotes a vector or matrix transpose. The 2-norm of a vector $\mathbf{u}$ is $\|\mathbf{u}\| = \sqrt{\mathbf{u}^t\mathbf{u}}$. Calligraphic letters are sets; $|\mathcal{X}|$ is the cardinality of $\mathcal{X}$. The indicator function is defined as $\mathbb{1}_{\{A\}} = 1$ if event A is true, and zero otherwise. All logarithms are base 2. Random variables are represented by capital letters. $\mathrm{E}\{X\}$ and $\mathrm{Var}\{X\}$ are the mean and variance of X, respectively, $H(X)$ is its entropy, and $I(X;Y)$ the mutual information between X and Y.

A host is denoted by the vector $\mathbf{x} = [x_1, x_2, \ldots, x_n]^t \in \mathcal{V}^n$ where $\mathcal{V} = \{v_1, v_2, \ldots, v_q\} \subset \mathbb{Z}$. We assume that $\mathbf{x} \neq \mathbf{0}$ and that $\mathbf{v} = [v_1, v_2, \ldots, v_q]^t$ gives the elements of $\mathcal{V}$ in increasing order, that is, $v_1 < v_2 < \cdots < v_q$. The histogram of $\mathbf{x}$ is a vector $\mathbf{h} = [h_1, h_2, \ldots, h_q]^t$ such that $h_k = \sum_{i=1}^{n} \mathbb{1}_{\{v_k = x_i\}}$, and then $\mathbf{h}^t\mathbf{1} = n$; $\mathbf{v}$ is therefore the vector containing the ordered histogram bins. Let $\mathcal{S}_n$ be the group of all permutations of $\{1, 2, \ldots, n\}$. We denote a permutation $\boldsymbol{\sigma} \in \mathcal{S}_n$ by means of a vector $\boldsymbol{\sigma} = [\sigma_1, \sigma_2, \ldots, \sigma_n]^t$ where $\sigma_i \in \{1, 2, \ldots, n\}$ and $\sigma_i \neq \sigma_j$ for all $i \neq j$. This vector can be used in turn to define a permutation matrix $\Pi_{\boldsymbol{\sigma}}$ with entries $(\Pi_{\boldsymbol{\sigma}})_{i,j} = \mathbb{1}_{\{\sigma_i = j\}}$. The rearrangement of $\mathbf{x}$ using $\boldsymbol{\sigma}$ is the vector $\mathbf{y} = \Pi_{\boldsymbol{\sigma}}\,\mathbf{x}$, for which $y_i = x_{\sigma_i}$ for $i = 1, 2, \ldots, n$. Notice that two or more different permutations may lead to the same rearrangement of $\mathbf{x}$; we will follow the convention that a rearrangement of $\mathbf{x}$ is a *unique* ordering of its elements. A special case is the rearrangement of $\mathbf{x}$ in nondecreasing order, which we denote by $\overrightarrow{\mathbf{x}}$. The rearrangement of $\mathbf{x}$ in nonincreasing order can be obtained from $\overrightarrow{\mathbf{x}}$ as $\overleftarrow{\mathbf{x}} = \mathrm{J}\overrightarrow{\mathbf{x}}$, where J is the exchange matrix with entries $(\mathrm{J})_{i,j} = \mathbb{1}_{\{j=n-i+1\}}$.

[1] A distortion constraint would be unnecessary if a distribution of the host exclusively modelling semantically meaningful outcomes were available.

II. Permutation Codes as Steganographic Codes

Any information-carrying vector $\mathbf{y}$ that preserves the histogram of $\mathbf{x}$ is a rearrangement of $\mathbf{x}$, and thus any first-order perfectly steganographic code must be chosen from the set of all Slepian's variant I permutation codes with base codeword $\mathbf{x}$ [5]. If $\mathbf{x}$ can be rearranged into r different vectors $\mathbf{y}^{(1)}, \mathbf{y}^{(2)}, \ldots, \mathbf{y}^{(r)}$ then there are at most r histogram-preserving watermarks given by $\mathbf{w}^{(m)} = \mathbf{y}^{(m)} - \mathbf{x}$ for $m = 1, 2, \ldots, r$ (and hence at most r different messages); we will drop the superindex m from $\mathbf{y}^{(m)}$ and $\mathbf{w}^{(m)}$ hereafter whenever this is unambiguous from the context. The number r of rearrangements of $\mathbf{x}$ only depends on its histogram $\mathbf{h}$, and it is given by the following multinomial coefficient:

$$r = \binom{n}{\mathbf{h}} = \frac{n!}{h_1!h_2!\cdots h_q!}. \tag{1}$$

In the remainder we will consider $\mathcal{S}_{\mathbf{x}} \subset \mathcal{S}_n$ to be any set of permutations leading to the $r = |\mathcal{S}_{\mathbf{x}}|$ rearrangements of $\mathbf{x}$. The steganographic embedding rate associated to a permutation code is $\rho \triangleq (1/n)\log r$ bits/host element. We showed in [4] that $\rho \approx H(X)$, where X is a random variable with distribution $\mathbf{p} = (1/n)\mathbf{h}$ (type of $\mathbf{x}$), and we also gave therein an efficient encoding/decoding method, based on adaptive arithmetic coding. This method allows us to obtain $\mathbf{y} = e(\mathbf{x}, m) = \Pi_{\boldsymbol{\sigma}}\mathbf{x}$ and $m = d(\mathbf{y})$ for $m = 1, 2, \cdots, 2^{\lfloor \log r \rfloor}$, sidestepping the exponential complexity of a naive table lookup encoder/decoder inherent in (1); therefore we assume in the remainder that any given permutation code is implementable.

A. Embedding Distortion

A very useful way to measure the embedding distortion is by means of the squared Euclidean distance between a codeword $\mathbf{y}$ and the host $\mathbf{x}$, that is, $\|\mathbf{w}\|^2 = \|\mathbf{y} - \mathbf{x}\|^2$. It is important to remark that if $\mathbf{x}$ is a decorrelated signal, then this amount is the same in the original domain when the decorrelating transform is energy-preserving (unitary), and so the distortion analysis can be considered to be completely general. Using the fact that all histogram-preserving codewords $\mathbf{y}$ have the same norm $\|\mathbf{y}\| = \|\mathbf{x}\|$, the squared norm (or power) of a histogram-preserving watermark can be put as

$$\|\mathbf{w}\|^2 = 2\left(\|\mathbf{x}\|^2 - \mathbf{x}^t\mathbf{y}\right) = 2\left(\|\mathbf{x}\|^2 - \mathbf{x}^t\Pi_{\boldsymbol{\sigma}}\mathbf{x}\right) \tag{2}$$

for some $\boldsymbol{\sigma} \in \mathcal{S}_{\mathbf{x}}$. The two following embedding distortion parameters were obtained in [4]:

- *Average watermark power.* With equally likely messages this amount is $\overline{\|\mathbf{w}\|^2} \triangleq \frac{1}{r}\sum_{m=1}^{r}\|\mathbf{w}^{(m)}\|^2$, which yields

$$\overline{\|\mathbf{w}\|^2} = 2\left(\|\mathbf{x}\|^2 - \frac{1}{n}(\mathbf{x}^t\mathbf{1})^2\right) = 2n\,s_{\mathbf{x}}^2, \tag{3}$$

where $s_{\mathbf{x}}^2$ is the (biased) sample variance. The centrality of this amount in permutation coding for steganography will become apparent throughout this paper.

- *Maximum watermark power.* This amount can be put as $(\|\mathbf{w}\|^2)_{\max} \triangleq \max_{m\in\{1,2,\cdots,r\}} \|\mathbf{w}^{(m)}\|^2$, which yields

$$(\|\mathbf{w}\|^2)_{\max} = 2\left(\|\mathbf{x}\|^2 - \overrightarrow{\mathbf{x}}^t\overleftarrow{\mathbf{x}}\right).$$

As discussed in [4] the following two theoretical figures of merit for the embedding distortion of permutation codes can be put forward: 1) document-to-average watermark power ratio, $\underline{\xi} \triangleq \|\mathbf{x}\|^2/\overline{\|\mathbf{w}\|^2}$ and 2) document-to-worst-case watermark power ratio $\xi_{\min} \triangleq \|\mathbf{x}\|^2/(\|\mathbf{w}\|^2)_{\max}$. Obviously, $\underline{\xi} \geq \xi_{\min}$. In keeping with standard conventions and where convenient throughout the paper, we will also refer to ξ ratios in terms of decibels (dB), by which the amount $10\log_{10}\xi$ is understood.

B. Asymptotic Behaviour of the Embedding Distortion

We will study next the asymptotic behaviour, for large n, of the embedding distortion of a histogram-preserving codeword drawn at random. Our purpose is to quantify a condition under which (3) is a good predictor of the power of any watermark, via the weak law of large numbers. A way to do so is by obtaining Chebyshev's bound for the random variable $\|\mathbf{W}\|^2 = 2(\|\mathbf{x}\|^2 - \mathbf{x}^t\Pi\mathbf{x})$, which is just (2) assuming that Π is a random permutation matrix following a uniform distribution (for uniformly distributed messages). We know already that $\mathrm{E}\{\|\mathbf{W}\|^2\} = \overline{\|\mathbf{w}\|^2}$, and therefore we just need to obtain the second moment of $\|\mathbf{W}\|^2$ in order to compute its variance. Noting that $\mathbf{x}^t\Pi\mathbf{x} = \mathbf{x}^t\Pi^t\mathbf{x}$, this moment can be put as

$$\mathrm{E}\{\|\mathbf{W}\|^4\} = 4\left(\|\mathbf{x}\|^4 - 2\|\mathbf{x}\|^2\mathbf{x}^t\,\mathrm{E}\{\Pi\}\mathbf{x} + \mathbf{x}^t\,\mathrm{E}\{\Pi\mathbf{x}\mathbf{x}^t\Pi^t\}\,\mathbf{x}\right),$$

and hence the desired variance is

$$\mathrm{Var}\{\|\mathbf{W}\|^2\} = 4\left(\mathbf{x}^t\,\mathrm{E}\{\Pi\mathbf{x}\mathbf{x}^t\Pi^t\}\,\mathbf{x} - \left(\mathbf{x}^t\,\mathrm{E}\{\Pi\}\mathbf{x}\right)^2\right). \quad (4)$$

We showed in [4], as a step in the derivation of (3), that the second expectation in (4) is

$$\mathrm{E}\{\Pi\} = \frac{1}{n!}\sum_{\boldsymbol{\sigma}\in\mathcal{S}_n}\Pi_{\boldsymbol{\sigma}} = \frac{1}{n}\mathbf{1}\mathbf{1}^t, \quad (5)$$

whereas the first expectation in (4) can be obtained using the general procedure described by Daniels in [7] to evaluate $\mathrm{E}\{\Pi\Delta\Pi^t\}$. In the special case $\Delta = \mathbf{x}\mathbf{x}^t$ relevant to us, Daniels' result $\mathrm{E}\{\Pi\mathbf{x}\mathbf{x}^t\Pi^t\} = a\,\mathrm{I} + b\,\mathbf{1}\mathbf{1}^t$ uses $b = \frac{1}{n(n-1)}((\mathbf{x}^t\mathbf{1})^2 - \|\mathbf{x}\|^2)$ and $a + b = \frac{1}{n}\|\mathbf{x}\|^2$. Employing this result, after some algebraic manipulations (4) becomes

$$\mathrm{Var}\{\|\mathbf{W}\|^2\} = \frac{4}{n-1}\left(\|\mathbf{x}\|^2 - \frac{1}{n}(\mathbf{x}^t\mathbf{1})^2\right)^2 = \frac{\left(\overline{\|\mathbf{w}\|^2}\right)^2}{n-1}.$$

Finally, we just use $\mathrm{Var}\{\|\mathbf{W}\|^2\}$ and $\mathrm{E}\{\|\mathbf{W}\|^2\}$ to obtain Chebyshev's bound for $\|\mathbf{W}\|^2$. For any $\gamma > 0$, this yields

$$\Pr\left\{\left|\|\mathbf{W}\|^2 - \overline{\|\mathbf{w}\|^2}\right| \geq \gamma\overline{\|\mathbf{w}\|^2}\right\} \leq \frac{1}{\gamma^2(n-1)}. \quad (6)$$

This inequality can be informally read as saying that, independently of the host $\mathbf{x}$, the embedding distortion associated to a randomly drawn permutation codeword is not likely to be too different from the average $\overline{\|\mathbf{w}\|^2}$ for large n. Figure 1 compares (6) to Monte Carlo computations using three hosts randomly sampled from Lena 512×512, for different values of n. The actual values of $\mathbf{x}$ are largely irrelevant to verify (6), because of its lack of dependence with the host. Adaptive arithmetic decoding [4] is used to empirically generate rearrangements of $\mathbf{x}$ for messages generated uniformly at random. Chebyshev's bound is known to be loose, although it is completely general and it illustrates the asymptotic behaviour. Despite these considerations, one might still be concerned about the rare instances in which $\|\mathbf{W}\|^2$ is much larger than the average. We will see next that these concerns will be dispelled by the fact that the geometry of permutation coding strictly confines the worst-case distortion.

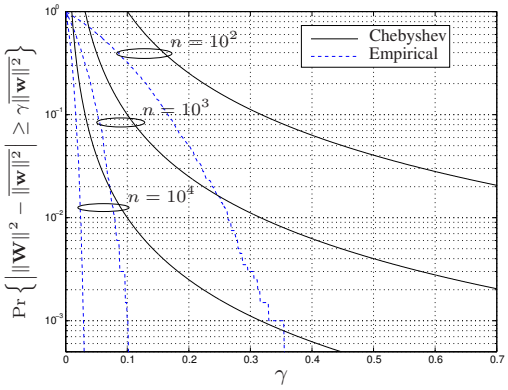

Fig. 1. Asymptotic behaviour of embedding distortion

C. Geometry and Embedding Distortion

As noted by Slepian [5] the two basic geometric properties of permutation codes are: 1) as $\|\mathbf{y}\| = \|\mathbf{x}\|$, all codewords lie on an n-dimensional *permutation sphere* with center $\mathbf{0}$ and radius $\|\mathbf{x}\|$; and 2) the codewords are really $n-1$ dimensional, as they also lie on the *permutation plane* $\mathbf{y}^t\mathbf{1} = \mathbf{x}^t\mathbf{1}$.

As we will show next, relevant geometric insights for the embedding distortion analysis can be obtained from the *covering sphere* of the permutation code. This is a sphere with centre $\mathbf{c}\in\mathbb{R}^n$ and minimum radius R_c such that for any codeword $\mathbf{y}$ it holds that $\|\mathbf{y}-\mathbf{c}\|^2 \leq R_c^2$. Since the intersection of the permutation plane and the permutation sphere is a sphere in $n-1$ dimensions that contains all histogram-preserving codewords, then this intersection must also be the intersection of the covering sphere with the permutation plane. In order to obtain $\mathbf{c}$ and R_c we first compute the average of all codewords $\overline{\mathbf{y}} = \frac{1}{r}\sum_{m=1}^r \mathbf{y}^{(m)}$, or, equivalently, the barycenter of the convex hull (polytope) spanned by all the rearrangements of $\mathbf{x}$. Using (5), this centroid is

$$\overline{\mathbf{y}} = \frac{1}{r}\sum_{\boldsymbol{\sigma}\in\mathcal{S}_{\mathbf{x}}}\Pi_{\boldsymbol{\sigma}}\mathbf{x} = \frac{1}{n!}\sum_{\boldsymbol{\sigma}\in\mathcal{S}_n}\Pi_{\boldsymbol{\sigma}}\mathbf{x} = \frac{1}{n}(\mathbf{x}^t\mathbf{1})\mathbf{1}. \quad (7)$$

Since all codewords lie on the permutation plane, so does $\overline{\mathbf{y}}$ ($\overline{\mathbf{y}}^t\mathbf{1} = \mathbf{x}^t\mathbf{1}$). Also, $\overline{\mathbf{y}}$ is always a point with equal coordinates,

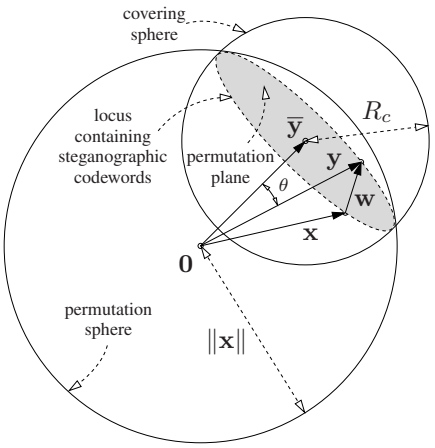

Fig. 2. Schematic of the geometry of first-order perfect steganography

in the positive or negative orthant depending on the sign of $\mathbf{x}^t\mathbf{1} \neq 0$. Now, the square of the Euclidean distance of an arbitrary codeword $\mathbf{y}$ to $\overline{\mathbf{y}}$ is $\|\mathbf{y}-\overline{\mathbf{y}}\|^2 = \|\mathbf{x}\|^2 - \|\overline{\mathbf{y}}\|^2$, where we have used $\mathbf{y}^t\mathbf{1} = \mathbf{x}^t\mathbf{1}$. As this squared distance is independent of $\mathbf{y}$, then it must also be the square of the covering radius, R_c^2, and $\overline{\mathbf{y}}$ must be the centre $\mathbf{c}$ of the covering sphere. Using (3) and (7) we can thus write

$$R_c^2 = \frac{1}{2}\overline{\|\mathbf{w}\|^2}. \tag{8}$$

Note that, in passing, we have also shown that when $\mathbf{x}^t\mathbf{1} \neq 0$ all codewords lie simultaneously on two different spheres: the covering sphere centered at $\overline{\mathbf{y}}$ with radius R_c, and the permutation sphere. Observing (3) and (8) we can see that $R_c \leq \|\mathbf{x}\|$, with equality when $\mathbf{x}^t\mathbf{1} = 0$. This is the reason why the covering sphere was not obtained in previous works applying permutation codes to channel/source coding: in these scenarios $\mathbf{x}$ can be chosen and $\mathbf{x}^t\mathbf{1} = 0$ is usually necessary for energy minimisation purposes (see [5, equation (19)]). Clearly, when $R_c < \|\mathbf{x}\|$ the two spheres intersect at the permutation plane $\mathbf{y}^t\mathbf{1} = n\sqrt{\|\mathbf{x}\|^2 - R_c^2} = \mathbf{x}^t\mathbf{1}$.

Using the triangle inequality we can verify next that

$$\|\mathbf{w}\| = \|(\mathbf{y}-\overline{\mathbf{y}}) - (\mathbf{x}-\overline{\mathbf{y}})\| \leq 2\|\mathbf{y}-\overline{\mathbf{y}}\| = 2R_c, \tag{9}$$

or, equivalently, that $\|\mathbf{w}\|$ cannot be greater than the diameter of the covering sphere. Combining (8), (9) and $R_c \leq \|\mathbf{x}\|$, we may then write the following inequalities for the worst-case scenario:

$$(\|\mathbf{w}\|^2)_{\max} \leq 2\overline{\|\mathbf{w}\|^2} \leq 4\|\mathbf{x}\|^2. \tag{10}$$

As it can be seen from (3), equality occurs in the second inequality in (10) when $\mathbf{x}^t\mathbf{1} = 0$. In this case, if there exist two antipodal codewords (i.e., $\mathbf{y} = -\mathbf{y}'$), then there is also equality in the first inequality in (10). This happens when $\overrightarrow{\mathbf{x}}$ is such that $\overrightarrow{x}_i = -\overrightarrow{x}_{n-i+1}$ for all $i = 1, 2, \cdots, n$, the antipodal codewords being $\overrightarrow{\mathbf{x}}$ and $\overleftarrow{\mathbf{x}}$. If $\mathbf{x}$ is not zero sum then there can only be equality in the first inequality in (10) when $\mathbf{x} = v\mathbf{1}$. This fact follows from the identity $(1/n)\overrightarrow{\mathbf{x}}^t\mathbf{1}\mathbf{1}\overrightarrow{\mathbf{x}}^t = \overrightarrow{\mathbf{x}}^t\mathrm{J}\overrightarrow{\mathbf{x}}^t$ that must hold in this case, in which we also have that $\overline{\|\mathbf{w}\|^2} = (\|\mathbf{w}\|^2)_{\max} = 0$ and null embedding rate.

If $\mathbf{x}$ lies in the nonnegative (or nonpositive) orthant then we can improve the second inequality in (10), as the greatest possible diameter of the covering sphere in this special case implies that $(\|\mathbf{w}\|^2)_{\max} \leq 2\|\mathbf{x}\|^2$ (with equality when $\mathbf{x} = [v, 0, \ldots, 0]^t$, or any of its n rearrangements), and thus we may replace (10) by $(\|\mathbf{w}\|^2)_{\max} \leq 2\min(\|\mathbf{x}\|^2, \overline{\|\mathbf{w}\|^2})$.

The two main consequences of the geometric analysis for the figures of merit are:

- The document-to-worst-case watermark power ratio is lower bounded as follows:

$$\xi_{\min} \geq \underline{\xi}/2 \geq 1/4. \tag{11}$$

  In decibels, the first inequality in (11) is $\xi_{\min} \gtrapprox \underline{\xi} - 3$ dB. Lastly, if $\mathbf{x}$ is in the nonnegative (or nonpositive) orthant then we can sharpen (11) using $\xi_{\min} \geq \max(1/2, \underline{\xi}/2)$.
- The document-to-average watermark power ratio can be expressed as a sole function of the angle θ between $\mathbf{x}$ and $\mathbf{1}$ (equivalently, between any codeword $\mathbf{y}$ and $\overline{\mathbf{y}}$). Since $\cos\theta = \mathbf{x}^t\mathbf{1}/(\|\mathbf{x}\|\|\mathbf{1}\|)$, we have from (3) that

$$\underline{\xi} = \frac{1}{2\sin^2\theta}.$$

The facts discussed in this section are schematically illustrated in Figure 2.

III. Rate-Distortion Tradeoff

A permutation code based on $\mathbf{x}$ may not directly meet a preestablished constraint $\underline{\xi}'$ on the minimum value of $\underline{\xi}$. For instance, (3) implies $\overline{\|\mathbf{w}\|^2} \leq 2\|\mathbf{x}\|^2$, and thus it can be seen that in the worst case $\underline{\xi} = 1/2$ (≈ -3 dB). Also from the second inequality in (11) we know that at worst $\xi_{\min} = 1/4$ (≈ -6 dB) —although we have seen that this case is vanishingly unlikely for large n. Therefore, in spite of the fact that a permutation code does implement first-order perfect steganography with maximum embedding rate, some form of embedding distortion control is needed in practice, which will lead to a decrease of the maximum rate $\rho \approx H(X)$.

A. Partitioned Permutation Coding

As described in [4], $\underline{\xi}$ can always be raised by restricting the codewords to a judiciously chosen subset from the ensemble of all histogram-preserving codewords as follows: 1) partition $\mathbf{x}$ into p disjoint subvectors $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_p$ with lengths $n_1, n_2, \cdots, n_p$ such that $\sum_{j=1}^p n_j = n$, following some *partitioning strategy*; and 2) undertake permutation coding within each $\mathbf{x}_j$ independently, that is, $\mathbf{y}_j = \Pi_{\boldsymbol{\sigma}_j}\mathbf{x}_j$ with $\boldsymbol{\sigma}_j \in \mathcal{S}_{\mathbf{x}_j}$ for $j = 1, 2, \cdots, p$. This strategy still preserves the histogram of $\mathbf{x}$, as trivially $\mathbf{y} = \Pi_{\boldsymbol{\sigma}}\mathbf{x}$ for some $\boldsymbol{\sigma} \in \mathcal{S}_{\mathbf{x}}$. Geometrically, the permissible codewords must now lie on the permutation ellipsoid $\sum_{j=1}^p \|\mathbf{y}_j\|^2/\|\mathbf{x}_j\|^2 = p$, as well as on the loci discussed in Section II-C.

The number of embeddable messages with partitioning is $r = \prod_{j=1}^p r_j$, where $r_j = \binom{n_j}{\mathbf{h}_j}$ is the multinomial coefficient associated to $\mathbf{x}_j$, and hence the embedding rate becomes

$$\rho = \sum_{j=1}^{p} \frac{n_j}{n}\rho_j, \tag{12}$$

where $\rho_j = (1/n_j)\log r_j$ is the embedding rate for the j-th partition. The average watermark power with partitioning is $\overline{\|\mathbf{w}\|^2} = (1/r)\sum_{m_1=1}^{r_1}\cdots\sum_{m_p=1}^{r_p}\sum_{j=1}^{p}\|\mathbf{w}_j^{(m_j)}\|^2$. This expression can be developed as follows: $\overline{\|\mathbf{w}\|^2} = \sum_{j=1}^{p}(1/r)\left(\prod_{\substack{i=1\\ i\neq j}}^{p} r_i\right)\sum_{m_j=1}^{r_j}\|\mathbf{w}_j^{(m_j)}\|^2 = \sum_{j=1}^{p}\overline{\|\mathbf{w}_j\|^2}$. Therefore

$$\overline{\|\mathbf{w}\|^2} = 2\left(\|\mathbf{x}\|^2 - \sum_{j=1}^{p}\frac{1}{n_j}(\mathbf{x}_j^t\mathbf{1})^2\right) = 2\sum_{j=1}^{p} n_j\ \mathbf{s}_{\mathbf{x}_j}^2, \quad (13)$$

where $s_{\mathbf{x}_j}^2 = \overline{\|\mathbf{w}_j\|^2}/(2n_j)$ is the sample variance of $\mathbf{x}_j$. The maximum watermark power is now $(\|\mathbf{w}\|^2)_{\max} = \max_{m_1,m_2,\cdots,m_p}\sum_{j=1}^{p}\|\mathbf{w}_j^{(m_j)}\|^2 = \sum_{j=1}^{p}(\|\mathbf{w}_j\|^2)_{\max}$. Thus, we have that

$$(\|\mathbf{w}\|^2)_{\max} = 2\left(\|\mathbf{x}\|^2 - \sum_{j=1}^{p}\overrightarrow{\mathbf{x}}_j^t\overleftarrow{\mathbf{x}}_j\right). \quad (14)$$

From expressions (13) and (14) one obtains $\underline{\xi}$ and $\xi_{\min}$ for the partitioned problem (see Section II-A). Importantly, the inequalities (11) still hold for partitioned permutation coding, because (10) holds for each of the p summands into which (13) and (14) can be decomposed. Hence, considering the fact that $\xi_{\min} \geq \underline{\xi}/2$ still holds with partitioning, it is reasonable to focus on the rate-distortion problem for the average distortion only, especially when recalling the asymptotic analysis in Section II-B.

B. Near-optimality of Partitioned Permutation Coding

The optimum rate-distortion tradeoff in a problem of communications with side information at the encoder, such as steganography, is given by Gel'fand and Pinsker's formula [8]. In a noiseless channel (passive warden) the best achievable rate is $\rho^* = \max_{p(y,\tilde{u}|x)} I(Y;\widetilde{U}) - I(\widetilde{U};X)$ bits/host sample subject to an embedding distortion constraint between Y and X, and where $\widetilde{U}$ is an auxiliary random variable. As shown in [8], it is enough to consider $Y = g(X,\widetilde{U})$, where $g(\cdot,\cdot)$ is a deterministic function. An additional constraint in perfect steganography is that the distribution of Y must be identical to the distribution of X. Hence the difference of mutual informations in Gel'fand and Pinsker's formula can be developed for this problem as follows:

$$I(Y;\widetilde{U}) - I(\widetilde{U};X) = H(X|\widetilde{U}) - H(Y|\widetilde{U}) \leq H(X|\widetilde{U}) \quad (15)$$

where the equality is because $H(Y) = H(X)$, and the inequality because the discrete entropy is nonnegative. Equality is achieved when $Y = g(\widetilde{U})$, as in this case $H(Y|\widetilde{U}) = 0$.

Let us next analyse the embedding rate (12) of partitioned permutation coding. As we did in [4] to approximate the embedding rate of the unpartitioned problem, we can use Stirling's formula for the factorial to write $\rho_j \approx H(X|\,U=j)$ bits/host sample, where U is a random variable with support set $\{1,2,\cdots,p\}$ and such that $p(U=j) = n_j/n$ (which is the probability that X belongs to the j-th partition), and $X|(U=j)$ is a random variable with probability mass function $\mathbf{p}_j = (1/n_j)\mathbf{h}_j$. We can then approximate (12) as

$$\rho \approx \sum_{j=1}^{p} p(U=j)\ H(X|U=j) = H(X|U).$$

Hence, the theoretical rate of partitioned permutation coding has the same mathematical form as the upper bound (15). Moreover, as in the condition for equality in the inequality in (15), the codeword sample y_i is chosen nearly independently of x_i, and only depending on the partitioning strategy, which is modelled by U. For the reasons above, U plays the same role as $\widetilde{U}$ in (15), and hence a careful choice of the partitioning strategy can lead to a near-optimum rate-distortion tradeoff.

To conclude this section, note that Comesaña and Pérez-González [9] previously pointed out that $H(X)$ is the absolute upper limit to the embedding rate of perfect steganography. This is correct, since $H(X|U) \leq H(X)$, but the observation in [9] does not consider an embedding distortion constraint. Thus in [9] a high embedding distortion follows from enforcing the rate $H(X)$: in fact, when $E\{X\} = 0$, the same worst case of the average embedding distortion for unpartitioned permutation coding given at the start of this section ($\underline{\xi} = 1/2$), which applies when $\mathbf{x}^t\mathbf{1} = 0$. Naturally, if one obviates the distortion constraint in permutation coding then $\rho \approx H(X|U) = H(X)$ is achievable by using one single partition (or, equivalently, unpartitioned permutation coding).

C. Upper Bound on Rate-distortion Function

A closed-form approximate upper bound to the optimum rate ρ^* is possible using the Djackov-Massey-Willems differential entropy upper bound on the discrete entropy [10][2], which is $H(X) < (1/2)\log(2\pi e(\sigma_X^2 + 1/12))$ for a discrete random variable with support set $\mathbb{Z}$ and variance σ_X^2. The utility of such a bound is in letting us know the goodness of a given partitioning strategy with respect to ρ^*, for a given constraint on $\overline{\|\mathbf{w}\|^2}$. For unpartitioned permutation coding, using the entropy approximation to the embedding rate, and observing that the support set of $X \sim \mathbf{p}$ is $\mathbb{Z}$ and that, from (3), its variance is $s_{\mathbf{x}}^2 = \overline{\|\mathbf{w}\|^2}/(2n)$, we have that the Djackov-Massey-Willems bound yields

$$\rho \lessapprox \rho_u \triangleq \frac{1}{2}\log\left(2\pi e\left(\frac{\overline{\|\mathbf{w}\|^2}}{2n} + \frac{1}{12}\right)\right). \quad (16)$$

For unpartitioned permutation coding $\overline{\|\mathbf{w}\|^2}$ is fixed, which would appear to limit the interest of (16). However we will verify next that (16) also holds for partitioned permutation coding (i.e. for (12) and (13)), in which $\overline{\|\mathbf{w}\|^2}$ can be tuned almost at will by choosing a suitable partitioning. To see this, one just needs to apply (16) individually to each of the partition rates ρ_j in (12), and then use the concavity of the

[2]This result is incorrectly cited as unpublished in [3, Problem 8.7].

logarithm and Jensen's inequality, which yields

$$\rho \lessapprox \sum_{j=1}^{p} \frac{n_j}{n} \frac{1}{2} \log\left(2\pi e \left(s^2_{\mathbf{x}_j} + \frac{1}{12}\right)\right) \leq \frac{1}{2} \log\left(2\pi e \sum_{j=1}^{p} \frac{n_j}{n} \left(s^2_{\mathbf{x}_j} + \frac{1}{12}\right)\right). \quad (17)$$

Using next (13) in (17) we recover (16), and hence $\rho^* \lessapprox \rho_u$. Importantly, (16) holds for any host, even if not memoryless if we take into account energy-preserving decorrelation, and thus is a fundamental rate-distortion limit in perfect steganography.

D. Optimum Partitioning Selection

The remaining question is the choice of an optimum partitioning. If encoder and decoder share a partitioning strategy, then all theoretical predictions for the partitioned problem given in Section III-A are achievable in practice. In order to do so the encoder partitions $\mathbf{x}$ and then, for $j = 1, 2, \cdots, p$, produces $\mathbf{y}_j$ by undertaking adaptive arithmetic decoding of $\lfloor \log r_j \rfloor$ bits of the message to be embedded relying on the histogram $\mathbf{h}_j$ of $\mathbf{x}_j$, as discussed in [4]. The decoder partitions $\mathbf{y}$ and then, for $j = 1, 2, \cdots, p$, undertakes adaptive arithmetic encoding of each subvector $\mathbf{y}_j$, thus retrieving the message.

As it will be seen later, the class of *histogram-induced* partitionings is of practical utility in the optimisation problem. The condition for a partitioning to be histogram-induced is that the subvectors $\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_p$, which contain the bins associated with the histograms $\mathbf{h}_1, \mathbf{h}_2, \cdots, \mathbf{h}_p$, be pairwise disjoint. Before discussing optimum partitioning, see that the simplest option for encoder and decoder is to preagree a *static* partitioning. In general this implies a suboptimum rate-distortion tradeoff. However, particular static partitionings can be practical if they almost surely lead to guaranteed rates and small distortions. In fact, we showed in [4] that simple histogram-induced static partitionings suffice for permutation coding to outdo popular steganographic techniques such as LSB matching (± 1 steganography) and model-based steganography, in the sense of essentially providing the same embedding rate, distortion and efficiency as these techniques while, unlike them, exactly preserving the first-order statistics.

An optimum rate-distortion tradeoff requires *adaptive* (host-dependent) partitioning. This can be achieved as follows: 1) encoder and decoder preagree a embedding distortion target $\underline{\xi}'$; 2) the encoder chooses the partitioning with optimum rate among all histogram-induced partitionings for which $\underline{\xi} \geq \underline{\xi}'$ when applied to $\mathbf{x}$, and uses it to produce $\mathbf{y}$; and 3) the decoder chooses a partitioning as in 2) but relying on $\mathbf{y}$ instead of $\mathbf{x}$ (i.e., as if $\mathbf{y}$ were the host), and uses it to decode the information embedded in $\mathbf{y}$. If the optimum partitioning is unique, both parties will agree on it through the procedure above: crucially, the theoretical analysis yields identical results when the host is either $\mathbf{x}$ or any rearrangement $\mathbf{y}$ for any tentative histogram-induced partitioning —even if $\mathbf{y}$ was not obtained from $\mathbf{x}$ through that tentative partitioning (see (12-13)). If the optimum is not unique, both parties will still find the same partitioning by following the same sequence of optimisation steps. Notice that, in any case, the partitioning used by the encoder never needs to be sent to the decoder.

The two issues with this procedure are: 1) achieving ρ^* might require a general partitioning (i.e. not histogram-induced); and 2) both encoder and decoder must solve a combinatorial optimisation problem of the generalised assignment class, but this is known to be NP-hard (cf. the optimal embedding function in [1]). Fortunately, an important clue to optimum partitioning is concealed in plain sight in (17): Jensen's inequality is met with equality if and only if $s^2_{\mathbf{x}_j} = \overline{\|\mathbf{w}\|^2}/(2n)$ for all $j = 1, 2, \cdots, p$, that is to say, when all partitions have the same sample variance. The adaptive partitioning strategy proposed at the end of Section 3.2 in [4] approximates this principle, and thus yields a rate not far away from ρ_u (which is a strict upper bound, and thus unattainable). For $s = 2$, assuming any host $\mathbf{x}$ with slowly varying histogram, this strategy can be shown to yield $\rho \approx 1$ bits/host sample and $\overline{\|\mathbf{w}\|^2} \approx n/2$, for which $\rho_u = 1.2546$ bits/host sample.

IV. Conclusion

We have extended the embedding distortion analysis of permutation coding for perfect steganography of memoryless sources started in [4]. We have also shown the near-optimality of partitioned permutation coding in terms of its rate-distortion tradeoff. All the facts considered, permutation coding shows the potential to be the basis of a systematic solution to near-optimum perfect steganography of real signals, in conjunction with invertible decorrelation.

Acknowledgement

This work has been financially supported by Science Foundation Ireland under grant 09/RFP/CMS2212.

References

[1] C. Cachin, "An information-theoretic model for steganography," in *Procs. of the 2nd Int. Workshop on Information Hiding*, ser. LNCS, vol. 1525. Portland, USA: Springer-Verlag, April 1998, pp. 306–318.

[2] R. Böhme, "An epistemological approach to steganography," in *Procs. of the 11th Workshop on Information Hiding*, ser. Lecture Notes in Computer Science. Springer, 2009, vol. 5806, pp. 15–30.

[3] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. John Wiley & Sons, 2006.

[4] F. Balado and D. Haughton, "Permutation codes and steganography," in *38th IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Vancouver, Canada, May 2013, pp. 2954–2958.

[5] D. Slepian, "Permutation modulation," *Procs. of the IEEE*, vol. 53, no. 3, pp. 228–236, 1965.

[6] H. Pengwei and Q. Shi, "Reversible integer KLT for progressive-to-lossless compression of multiple component images," in *Procs. of the 10th IEEE Int. Conf. on Image Processing (ICIP)*, vol. 1, Barcelona, Spain, September 2003, pp. 633–636.

[7] H. E. Daniels, "Processes generating permutation expansions," *Biometrika*, vol. 49, no. 1/2, pp. 139–149, June 1962.

[8] S. I. Gel'fand and M. S. Pinsker, "Coding for channel with random parameters," *Problems of Control and Information Theory*, vol. 9, no. 1, pp. 19–31, 1980.

[9] P. Comesaña and F. Pérez-González, "On the capacity of stegosystems," in *Procs. of the 9th ACM Workshop on Multimedia & Security*, Dallas, USA, September 2007, pp. 15–24.

[10] J. Massey, "On the entropy of integer-valued random variables," in *Procs. of Int. Workshop on Information Theory*, Beijing, China, July 1988, pp. C1.2–C1.4.