Exploiting Security Holes in Lattice Data Hiding

Luis Pérez-Freire and Fernando Pérez-González*

Signal Theory and Communications Department
University of Vigo, Vigo 36310, Spain
{lpfreire, fperez}@gts.tsc.uvigo.es

Abstract. This paper presents a security analysis for data hiding methods based on nested lattice codes, extending the analysis provided by previous works. The security is quantified in an information-theoretic sense by means of the information leakage between the watermarked signals seen by the attacker and the secret key used in the embedding process. The theoretical analysis accomplished in the first part of the paper addresses important issues such as the possibility of achieving perfect secrecy and the impact of the embedding rate and channel coding in the security level. In the second part, a practical algorithm for estimating the secret key is proposed, and the information extracted is used for implementing a reversibility attack on real images.

1 Introduction

Watermarking security has emerged in the last years as a new research topic, whose basics can be found in [1],[2],[3] and the references therein. The framework for security analysis adopted in these works follows a cryptanalytic approach: all the parameters of the watermarking scheme are assumed to be public, and the security relies only on a secret key, which is assumed to remain unchanged in the contents watermarked by the same user. The main target of the security analysis is to determine whether the watermarking scheme conceals properly the secret key; if it is not the case, then we are interested in assessing the security level of the scheme, defined as the number of observations needed to achieve an estimate of the secret key up to a certain accuracy [2].

In this paper we focus on the security analysis of data hiding schemes based on nested lattice codes [4], usually known as lattice DC-DM schemes. Specifically, the work in the present paper extends the theory and algorithms developed in [5] to a more general scenario. The analysis in [5] was mainly restricted to the so-called "Known Message Attack" (KMA) scenario, where the messages embedded

* This work was partially funded by *Xunta de Galicia* under projects PGIDT04 TIC322013PR, PGIDT04 PXIC32202PM, and "Competitive Research Units" program, Ref. 150/2006; MEC project DIPSTICK, Ref. TEC2004-02551/TCM, and European Commission through the IST Programme under Contract IST-2002-507932 ECRYPT. ECRYPT disclaimer: The information in this paper is provided as is, and no guarantee or warranty is given or implied that the information is fit for any particular purpose. The user thereof uses the information at its sole risk and liability.

T. Furon et al. (Eds.): IH 2007, LNCS 4567, pp. 159–173, 2007.

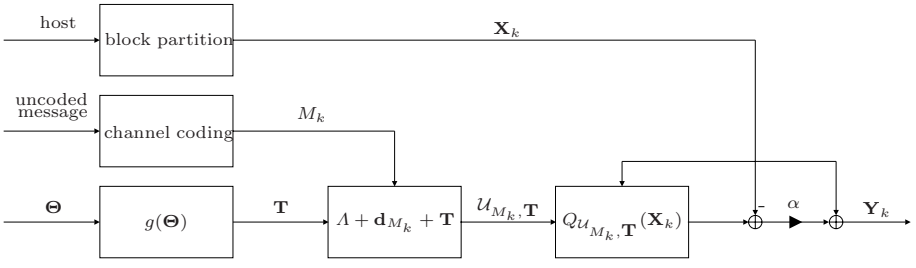


Fig. 1. Block diagram showing the lattice data hiding model

in each watermarked signal were assumed to be known by the attacker. This paper considers a general scenario (which encompasses most of the practical data hiding applications), termed "Watermarked Only Attack" (WOA), where the attacker no longer knows anything about the embedded messages. As in [5], the security level is measured by means of the mutual information (a.k.a. information leakage) between the watermarked signals and the secret key, which is related to the variance of the key estimation error. The first part of this paper measures the information leakage for lattice DC-DM schemes, paying special attention to the comparison between KMA and WOA scenarios, and considering also possible strategies that achieve good security levels. The second part shows how the information about the key provided by the observations can be extracted and used in practical scenarios, proposing a reversibility attack based on an estimate of the secret dither. The proposed estimation algorithm works with any arbitrary nested lattice code, and is applicable to high embedding rate scenarios.

The main notational conventions used in the paper are the following: Λ_f and Λ are the n-dimensional fine and coarse (shaping) lattices of the nested lattice code, respectively. The alphabet that encodes the messages to be transmitted is defined as $\mathcal{M} \triangleq \{0, 1, \ldots, p-1\}$, with p denoting its cardinality. Random variables are denoted by capital letters, and vectors are represented by boldface letters. $H(\cdot)$ and $h(\cdot)$ denote entropy and differential entropy [6], respectively.

2 Theoretical Model

The mathematical model for lattice data hiding considered in this paper is shown in Fig. 1. First, the host signal is partitioned into non-overlapping blocks $\mathbf{X}_k$ of length n. The message to be embedded may undergo channel coding, yielding the symbols $M_k \in \mathcal{M}$ which are assumed to be equiprobable, unless otherwise stated. Each symbol M_k is embedded in one block $\mathbf{X}_k$ by means of a randomized lattice quantizer yielding a watermarked signal $\mathbf{Y}_k$ as follows:

$$\mathbf{Y}_k = \mathbf{X}_k + \alpha(Q_\Lambda(\mathbf{X}_k - \mathbf{d}_{M_k} - \mathbf{T}) - \mathbf{X}_k + \mathbf{d}_{M_k} + \mathbf{T}), \tag{1}$$

where $Q_\Lambda(\mathbf{x})$ is a nearest neighbor quantizer whose centroids are distributed according to Λ, the coarse (shaping) lattice, $\alpha \in [0, 1]$ is the distortion

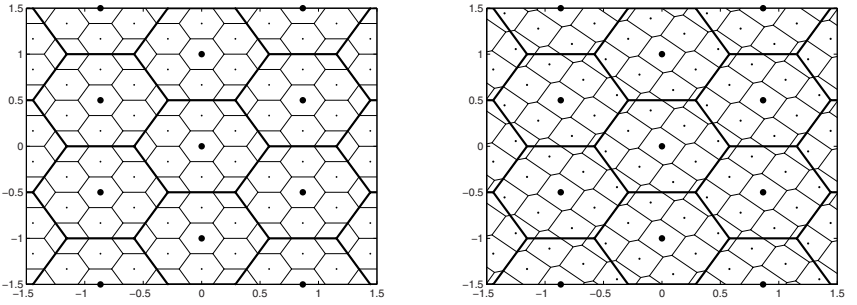


Fig. 2. Nested lattice codes of rate $R = \log(9)/2$ with hexagonal shaping lattice, obtained by means of self-similar construction (a) and Construction A with $\mathbf{g} = (1, 2)^T$ (b). Voronoi regions of Λ_f and Λ are represented by thin and thick lines, respectively.

compensation parameter, $\mathbf{d}_{M_k}$ is the coset leader associated to M_k, and $\mathbf{T} = g(\boldsymbol{\Theta})$ is the secret dither signal, which remains constant in each watermarked block. The embedding distortion per dimension in a mean-squared-error sense can be computed as $D_w = \frac{1}{n} E\left[||\mathbf{X}_k - \mathbf{Y}_k||^2\right] = \alpha^2 P(\Lambda)$, where $P(\Lambda)$ denotes the second-order moment per dimension of $\mathcal{V}(\Lambda)$.

The coset leaders $\mathbf{d}_{M_k} \in \{\mathbf{d}_0, \ldots, \mathbf{d}_{p-1}\}$ that encode each symbol M_k are chosen so that $\bigcup_{k=0}^{p-1}(\mathbf{d}_k + \Lambda) = \Lambda_f$ and they coincide with the coset leaders of minimum norm of the nested lattice code. Nested lattice codes can be obtained in a number of ways; we consider in this paper self-similar lattice partitions and Construction A [7]. In self-similar lattice partitions, $\Lambda_f = p^{-\frac{1}{n}}\Lambda$, for $p^{\frac{1}{n}} \in \mathbb{N}$.[1] The lattice Λ is a sublattice of Λ_f, resulting in a "nesting ratio" $\frac{\mathrm{vol}(\mathcal{V}(\Lambda))}{\mathrm{vol}(\mathcal{V}(\Lambda_f))} = p$, and an embedding rate $R = \log(p)/n$. The coset leaders $\mathbf{d}_k$ can be obtained as $\Lambda_f \cap \mathcal{V}(\Lambda)$. Construction A is more flexible, and is summarized as follows:

1. Define a positive integer p. In order to construct a nested lattice code with good asymptotic properties, p must be prime.
2. Define a generating vector $\mathbf{g} \in \mathbb{Z}_p^n$ and compute the codebook $\mathcal{C} \triangleq \{\mathbf{c} \in \mathbb{Z}_p^n : \mathbf{c} = q \cdot \mathbf{g} \mod p,\ q = 0, \ldots, p-1\}$. Then, define the lattice $\Lambda' = p^{-1}\mathcal{C} + \mathbb{Z}^n$.
3. Define the generating matrix $\mathbf{G} \in \mathbb{R}^{n \times n}$ (where each row is a basis vector) of the coarse (shaping) lattice Λ. Apply the linear transformation $\Lambda_f = \Lambda'\mathbf{G}$. It follows that Λ is a sublattice of Λ_f and the nesting ratio is $\frac{\mathrm{vol}(\mathcal{V}(\Lambda))}{\mathrm{vol}(\mathcal{V}(\Lambda_f))} = p$, resulting in a coding rate $R = \log(p)/n$.
4. The coset leaders are given by $\Lambda_f \cap \mathcal{V}(\Lambda)$, or equivalently, $p^{-1}\mathcal{C}\mathbf{G} \mod \Lambda$.

Examples of 2-dimensional nested lattice codes are shown in Fig. 2.

With regard to the attacker's strategy, it is assumed that he manages to gather an ensemble of watermarked blocks $\{\mathbf{Y}_k,\ k = 1, \ldots, N_o\}$ (hereinafter,

[1] More general self-similar lattice partitions consider also rotations of Λ, but we will restrict our attention to those obtained through scaling.

observations), which may belong to different host signals, but all of them were watermarked with the same secret key $\boldsymbol{\Theta}$. He knows the parameters of the nested lattice code being used, i.e. Λ, $\{\mathbf{d}_i,\ i = 0, \ldots, p-1\}$, α, whereas he ignores the host blocks $\mathbf{X}_k$, the embedded symbols M_k, and $\mathbf{T} = g(\boldsymbol{\Theta})$. The objective of the attacker is to obtain an estimate of $\mathbf{T}$. The first step performed by him is the modulo reduction of the watermarked blocks as $\tilde{\mathbf{Y}}_k \triangleq \mathbf{Y}_k \mod \Lambda$, where the modulo operation is defined as $\mathbf{X} \mod \Lambda \triangleq \mathbf{X} - Q_\Lambda(\mathbf{X})$. Under the assumption (a.k.a. "flat-host assumption") that the variance of the components of $\mathbf{X}_k$ is much larger than the embedding distortion, such modulo reduction does not imply any loss of information for the attacker, as discussed in [5]. Let us define

$$f_0(\mathbf{x}) \triangleq \begin{cases} (\mathrm{vol}(\mathcal{Z}(\Lambda)))^{-1}, & \mathbf{x} \in \mathcal{Z}(\Lambda) \\ 0, & \text{otherwise,} \end{cases} \tag{2}$$

with $\mathcal{Z}(\Lambda) \triangleq (1-\alpha)\mathcal{V}(\Lambda)$, where $\mathcal{V}(\Lambda) \triangleq \{\mathbf{x} \in \mathbb{R}^n : Q_\Lambda(\mathbf{x}) = \mathbf{0}\}$ denotes the Voronoi region of Λ [8]. The probability density function of the signals seen by the attacker can be computed by taking into account that, under the flat-host assumption, $f(\tilde{\mathbf{y}}_k | m_k, \mathbf{t}) = f_0(\tilde{\mathbf{y}}_k - \mathbf{d}_{m_k} - \mathbf{t} \mod \Lambda)$. Finally, the function $g(\cdot)$ is assumed to yield a secret dither $\mathbf{T}$ uniformly distributed in the Voronoi region $\mathcal{V}(\Lambda)$, which turns out to be the worst case for the attacker [5].

3 Theoretical Security Analysis

The amount of information that leaks from the observations is quantified by means of the mutual information $I(\tilde{\mathbf{Y}}_1, \ldots, \tilde{\mathbf{Y}}_{N_o}; \mathbf{T})$. Making use of the chain rule for entropies [6], it can be written in a more illustrative manner as

$$\begin{aligned} &I(\tilde{\mathbf{Y}}_1, \ldots, \tilde{\mathbf{Y}}_{N_o}; \mathbf{T}) \\ &\quad = I(\tilde{\mathbf{Y}}_1, \ldots, \tilde{\mathbf{Y}}_{N_o}; \mathbf{T}, M_1, \ldots, M_{N_o}) - I(\tilde{\mathbf{Y}}_1, \ldots, \tilde{\mathbf{Y}}_{N_o}; M_1, \ldots, M_{N_o} | \mathbf{T}) \\ &\quad = I(\tilde{\mathbf{Y}}_1, \ldots, \tilde{\mathbf{Y}}_{N_o}; \mathbf{T} | M_1, \ldots, M_{N_o}) + I(\tilde{\mathbf{Y}}_1, \ldots, \tilde{\mathbf{Y}}_{N_o}; M_1, \ldots, M_{N_o}) \\ &\quad - I(\tilde{\mathbf{Y}}_1, \ldots, \tilde{\mathbf{Y}}_{N_o}; M_1, \ldots, M_{N_o} | \mathbf{T}). \end{aligned} \tag{3}$$

The first term of (3) is the information leakage in the KMA case, that was studied in [5]. One fundamental property of the KMA scenario is that, under the assumption $\mathbf{T} \sim U(\mathcal{V}(\Lambda))$, the conditional pdf of the dither signal is [5]

$$f(\mathbf{t} | \tilde{\mathbf{y}}_1, \ldots, \tilde{\mathbf{y}}_{N_o}, \mathbf{m}) = \begin{cases} (\mathrm{vol}(\mathcal{S}_{N_o}(\mathbf{m})))^{-1}, & \mathbf{t} \in \mathcal{S}_{N_o}(\mathbf{m}) \\ 0, & \text{otherwise,} \end{cases} \tag{4}$$

where $\mathbf{m} \triangleq (m_1, \ldots, m_{N_o})$,

$$\mathcal{S}_{N_o}(\mathbf{m}) \triangleq \bigcap_{j=1}^{N_o} \mathcal{D}_j(m_j), \tag{5}$$

$$\mathcal{D}_j(m_j) = (\tilde{\mathbf{y}}_j - \mathbf{d}_{m_j} - \mathcal{Z}(\Lambda)) \mod \Lambda. \tag{6}$$

Eq. (5) denotes the "feasible region" for the secret dither, conditioned on the observations and the message sequence $\mathbf{m}$. This property will be frequently used in the remaining of this paper. The third term of (3) represents the achievable rate for a fair user, i.e., knowing the secret dither $\mathbf{T}$, whereas the second term is the rate achievable by unfair users (which is not null, in general) that do not know $\mathbf{T}$. A similar reasoning to that followed in [5, Sect. II] shows that the mutual information in (3) is concave and increasing with N_o. Notice that

$$I(\tilde{\mathbf{Y}}_1, \ldots, \tilde{\mathbf{Y}}_{N_o}; M_1, \ldots, M_{N_o}) - I(\tilde{\mathbf{Y}}_1, \ldots, \tilde{\mathbf{Y}}_{N_o}; M_1, \ldots, M_{N_o} | \mathbf{T}) < 0, \quad (7)$$

so the information leakage in the WOA case never exceeds that in the KMA case, as expected. In order to compute the asymptotic gap (when $N_o \to \infty$) between the security level of KMA and WOA scenarios, the left hand side of (7) is rewritten as $N_o \cdot H(M_1|\tilde{\mathbf{Y}}_1, \mathbf{T}) - H(M_1, \ldots, M_{N_o}|\tilde{\mathbf{Y}}_1, \ldots, \tilde{\mathbf{Y}}_{N_o})$. Although the proof cannot be included here due to space limitations, it is possible to show that, for equiprobable message sequences,

$$\lim_{N_o \to \infty} (H(M_1, \ldots, M_{N_o}|\tilde{\mathbf{Y}}_1, \ldots, \tilde{\mathbf{Y}}_{N_o}) - N_o \cdot H(M_1|\tilde{\mathbf{Y}}_1, \mathbf{T})) \to \log(p). \quad (8)$$

Hence, the asymptotic gap between KMA and WOA scenarios in terms of information leakage per dimension is $R = \log(p)/n$, i.e., the embedding rate. This result has important implications, since in practical scenarios we usually resort to low embedding rates that allow to recover the embedded message without the use of complex channel coding schemes. The problem is that low embedding rates may yield a security level similar to that of the KMA scenario. In spite of this, the WOA scenario still provides one major advantage over the KMA in terms of security, because in the WOA case the attacker cannot aspire to acquire perfect knowledge of the secret dither vector (even for infinite N_o) unless he has information about the a priori probabilities of the message sequences (introduced by the specific channel coding scheme being applied, for instance). This is a consequence of the following property: [2]

$$\Pr(\mathbf{m}|\tilde{\mathbf{y}}_1, \ldots, \tilde{\mathbf{y}}_{N_o}) = \Pr((\mathbf{m} + j \cdot \mathbf{1}) \mod p | \tilde{\mathbf{y}}_1, \ldots, \tilde{\mathbf{y}}_{N_o}) \cdot \frac{\Pr(\mathbf{m})}{\Pr((\mathbf{m} + j \cdot \mathbf{1}) \mod p)}, \; j \in \mathcal{M}, \quad (9)$$

where the modulo operation is applied componentwise, and $\mathbf{1}$ denotes the n-dimensional vector with all components equal to 1. The obtention of Eq. (9) follows by combining equations (18) and (20) of the Appendix, and taking into account that the addition of a constant vector to the observations does not change the a posteriori probabilities of the embedded messages. This ambiguity makes impossible to reduce the uncertainty about $\mathbf{T}$ beyond a set of p discrete (equiprobable) points. However, such uncertainty can be further reduced by exploiting the statistical dependence between the symbols embedded in different blocks if a channel code has been applied.

[2] Eq. (9) holds directly for nested codes obtained through Construction A, and also for codes obtained through self-similar partitions if the coset leaders are properly arranged in $\mathcal{M}$.

3.1 Coding Strategies for Achieving Perfect Secrecy

Under certain assumptions it is possible to achieve null information leakage, a.k.a. "perfect secrecy", i.e., $I(\tilde{\mathbf{Y}}_1, \ldots, \tilde{\mathbf{Y}}_{N_o}; \mathbf{T}) = 0$. In the following lemma, two different strategies are considered.

Lemma 1. *Assuming equiprobable symbols and independence between the messages embedded in different blocks, the two following strategies achieve perfect secrecy:*

1. *Using self-similar lattice partitions with nesting ratio p and distortion compensation parameter $\alpha_k = 1 - kp^{-\frac{1}{n}}, k = 1, \ldots, p^{\frac{1}{n}} - 1$.*
2. *Making $|\mathcal{M}| \to \infty$ with the coset leaders $\mathbf{d}_k,\ k = 0, \ldots, \infty$ uniformly distributed in $\mathcal{V}(\Lambda)$.*

Outline of the proof: Due to the lack of space, a detailed proof is not included. The proof is based on the fact that, under the assumption of independence between the embedded messages, $h(\tilde{\mathbf{Y}}_1) = h(\tilde{\mathbf{Y}}_1|\mathbf{T})$ is a necessary and sufficient condition for achieving perfect secrecy. For proving the first part of the lemma, we have to prove that

$$f(\tilde{\mathbf{y}}_1|\mathbf{T} = \mathbf{t}) = \frac{1}{p}\sum_{i=0}^{p-1} f_0(\tilde{\mathbf{y}}_1 - \mathbf{t} - \mathbf{d}_i \mod \Lambda) = \frac{1}{\mathrm{vol}(\mathcal{V}(\Lambda))} \ \forall\ \tilde{\mathbf{y}}_1 \in \mathcal{V}(\Lambda), \quad (10)$$

where $f_0(\cdot)$ is given by (2). Intuitively, Eq. (10) turns out to be true for the considered values of α because in that case the union of p regions $\mathcal{Z}(\Lambda)$ (which are scaled versions of $\mathcal{V}(\Lambda_f)$) shifted by the corresponding coset leaders $\mathbf{d}_k, k \in \mathcal{M}$, perfectly packs in space, yielding a watermarked signal uniformly distributed in $\mathcal{V}(\Lambda)$. The proof of the second part of the lemma consists in showing that $f(\tilde{\mathbf{y}}_1|\mathbf{T} = \mathbf{t}) = (\mathrm{vol}(\mathcal{V}(\Lambda)))^{-1}\ \forall\ \tilde{\mathbf{y}}_1 \in \mathcal{V}(\Lambda)$, which is true due to the uniform distribution of the coset leaders. □

Some remarks to the results stated in Lemma 1 are in order:

1) The first strategy stated in Lemma 1 yields a finite and discrete set of values for α that permit to achieve perfect secrecy; however, the choice of these values may be in conflict with robustness requirements. Notice also that the second strategy is independent of α and the type of lattice partition.

2) Lemma 1 suggests that, for achieving good security levels, the codewords (coset leaders) must be uniformly distributed over $\mathcal{V}(\Lambda)$ in order to completely fill the space (also with help of the self-noise introduced when $\alpha < 1$). Thus, simple coding schemes (as repetition coding, see Section 3.2) do not necessarily yield good security levels, even for high embedding rates.

3) The condition of mutual independence between the symbols embedded in different observations is key to guarantee perfect secrecy. To see this, note that the conditional pdf of the dither signal can be written as

$$f(\mathbf{t}|\tilde{\mathbf{y}}_1, \ldots, \tilde{\mathbf{y}}_{N_o}) = \sum_{\mathbf{m} \in \mathcal{M}^{N_o}} f(\mathbf{t}|\mathbf{m}, \tilde{\mathbf{y}}_1, \ldots, \tilde{\mathbf{y}}_{N_o}) \cdot \Pr(\mathbf{m}|\tilde{\mathbf{y}}_1, \ldots, \tilde{\mathbf{y}}_{N_o}),$$

where $\mathcal{M}^{N_o}$ denotes the whole message space. When perfect secrecy is achieved, the probability distribution $\Pr(\mathbf{m}|\tilde{\mathbf{y}}_1, \ldots, \tilde{\mathbf{y}}_{N_o})$ makes the conditional pdf of the dither uniform over $\mathcal{V}(\Lambda)$. If the symbols m_i are not mutually independent, then the a posteriori distribution of the messages is changed, so the conditional pdf of the dither is no longer uniform. Hence, the strategies proposed in Lemma 1 in conjunction with channel coding across different blocks will provide perfect secrecy only if the attacker ignores the channel code being applied and the dependencies between symbols that it introduces.

4) The proof of the lemma resorts to the flat-host assumption to show null information leakage. In practice, small information leakages may exist due to the finite variance of the host signal, which causes the host distribution to not be strictly uniform in each quantization cell. However, this information leakage seems to be hardly exploitable in practical attacks.

3.2 Theoretical Results for Cubic Lattices with Repetition Coding

One of the most popular schemes for lattice data hiding is DC-DM with repetition coding [9], which can be seen as a particular case of Construction A using $\mathbf{g} = (1, \ldots, 1)^T$ and $\Lambda = \Delta\mathbb{Z}^n$. In order to obtain the information leakage for this scheme, Eq. (3) is rewritten using the chain rule for mutual informations [6] and the results in [5, Sect. III] as (assuming equiprobable message sequences)

$$\begin{aligned}\tfrac{1}{n} I(\tilde{\mathbf{Y}}_1, \ldots, \tilde{\mathbf{Y}}_{N_o}; \mathbf{T}) \\ &= \frac{1}{n} N_o \cdot H(M_1|\tilde{\mathbf{Y}}_1, \mathbf{T}) - \frac{1}{n} H(M_1, \ldots, M_{N_o}|\tilde{\mathbf{Y}}_1, \ldots, \tilde{\mathbf{Y}}_{N_o}) \\ &+ \sum_{i=2}^{N_o} \frac{1}{i} - \log(1-\alpha),\ N_o \geq 2. \end{aligned} \tag{11}$$

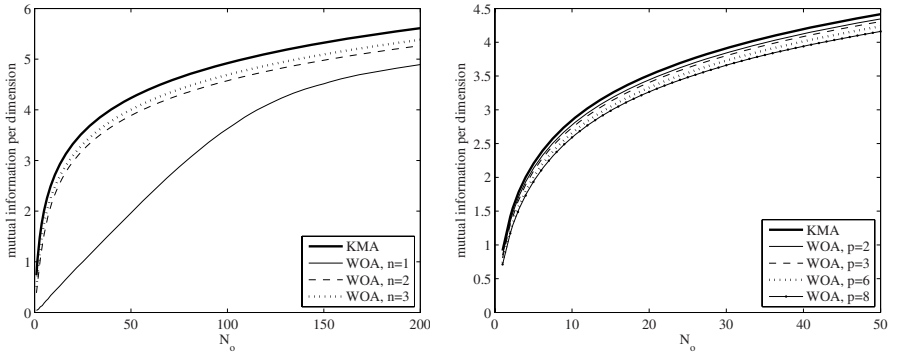
Eq. (11) does not admit a closed-form expression, although it is possible to obtain the entropies of interest numerically. The second term of Eq. (11) is

$$E_{\tilde{\mathbf{Y}}_1, \ldots, \tilde{\mathbf{Y}}_{N_o}} \left[H(M_1, \ldots, M_{N_o}|\tilde{\mathbf{Y}}_1 = \tilde{\mathbf{y}}_1, \ldots, \tilde{\mathbf{Y}}_{N_o} = \tilde{\mathbf{y}}_{N_o}) \right], \tag{12}$$

which can be computed through the a posteriori probability distribution of the message sequences, that can be obtained according to the Appendix, arriving at (assuming equiprobable message sequences again)

$$\Pr(m_1, \ldots, m_{N_o}|\tilde{\mathbf{y}}_1, \ldots, \tilde{\mathbf{y}}_{N_o}) = \frac{\mathrm{vol}(\mathcal{S}_{N_o}(m_1, \ldots, m_{N_o}))}{\sum_{\mathbf{m} \in \mathcal{M}^{N_o}} \mathrm{vol}(\mathcal{S}_{N_o}(\mathbf{m}))}, \tag{13}$$

where $\mathcal{M}^{N_o}$ denotes the whole message space for N_o observations (actually, only the message sequences with non-null probability need to be taken into account). The feasible region is always a hypercube, and as such it can be computed componentwise. Finally, the entropy (12) is obtained by averaging over the realizations



(a) (b)

Fig. 3. Information leakage per dimension for cubic Λ using repetition coding. Impact of the repetition rate (n) for $\alpha = 0.52$ and $p = 2$ (a), and impact of the alphabet size (p) for $\alpha = 0.6$ and $n = 10$ (b).

of $\tilde{\mathbf{Y}}_1, \ldots, \tilde{\mathbf{Y}}_{N_o}$. For repetition coding, the problem of computing $H(M_1|\tilde{\mathbf{Y}}_1, \mathbf{T})$ is the dual of the computation of $h(\mathbf{T}|\tilde{\mathbf{Y}}_1, \ldots, \tilde{\mathbf{Y}}_{N_o}, M_1, \ldots, M_{N_o})$, which was addressed in [5]. For the entropy of interest, we have

$$H(M_1|\tilde{\mathbf{Y}}_1, \mathbf{T}) = H(M_1|\tilde{\mathbf{Y}}_1, \mathbf{T} = \mathbf{0}) = E_{\tilde{\mathbf{Y}}_1}\left[H(\mathbf{d}_{M_1}|\tilde{\mathbf{Y}}_1 = \mathbf{y}, \mathbf{T} = \mathbf{0})\right]. \quad (14)$$

and $H(\mathbf{d}_{M_1}|\tilde{\mathbf{Y}}_1 = \mathbf{y}, \mathbf{T} = \mathbf{0}) = \log(\sum_{k=0}^{p-1} I_{\mathcal{I}}((\Delta \cdot k/p) \mod \Delta))$, where $I_{\mathcal{I}}(\cdot)$ is the indicator function:

$$I_{\mathcal{I}}(x) \triangleq \begin{cases} 1, x \in \mathcal{I} \\ 0, \text{otherwise,} \end{cases}$$

and $\mathcal{I} \triangleq [\max_{i=1,\ldots,n}\{\tilde{y}_i\} - (1-\alpha)\Delta/2,\ \min_{i=1,\ldots,n}\{\tilde{y}_i\} + (1-\alpha)\Delta/2]$,with $y_i,\ i = 1, \ldots, n$, the components of $\tilde{\mathbf{y}}_1$. The expectation (14) is obtained by averaging over the realizations of $\tilde{\mathbf{Y}}_1$.

The results are illustrated in Fig. 3 and compared to the results obtained for the KMA scenario, supporting some of the conclusions given in sections 3 and 3.1. Specifically, Fig. 3(a) shows the negative impact in the security level of increasing the dimensionality while keeping constant the embedding rate, whereas Fig. 3(b) shows the security improvement brought about by the increase of the alphabet size. Fig. 3(a) shows that the gap in the information leakage between KMA and WOA tends asymptotically to $\log(2)/n$. On the other hand, Fig. 3(b) shows that increasing p does not yield a significant improvement; in fact, it can be shown that with repetition coding is not possible to achieve perfect secrecy in any way (but for $n = 1$, which yields a self-similar partition), because the distribution of the coset leaders (in a diagonal of the n-dimensional hypercube) does not allow to fulfill the condition of perfect secrecy ($h(\tilde{\mathbf{Y}}_1) = h(\hat{\mathbf{Y}}_1|\mathbf{T})$).

4 A Practical Dither Estimator

4.1 Dither Estimator Based on Set-Membership Theory for KMA

In case the embedded symbols $\mathbf{m} = \{m_1, \ldots, m_{N_o}\}$ are known by the attacker, the algorithm proposed in [5, Sect. IV] gives an accurate estimate of the secret dither. This estimator exploits the fact that each observation defines a bounded feasible region for $\mathbf{T}$, according to Eq. (5). It works under the assumption that $\alpha > 0.5$, in order to assure convergence. The feasible region corresponding to the ith observation (Eq. (6)) is redefined as $\mathcal{D}_i(m_i) \triangleq \tilde{\mathbf{v}}_i + \mathcal{Z}(\Lambda),\ i = 1, \ldots, N_o$,where $\tilde{\mathbf{v}}_i \triangleq (\tilde{\mathbf{y}}_i - \mathbf{d}_{m_i} - \tilde{\mathbf{y}}_1 + \mathbf{d}_{m_1}) \mod \Lambda$. By introducing the offset $-\tilde{\mathbf{y}}_1 + \mathbf{d}_{m_1}$ in every observation, we get a convex $\mathcal{S}_k(\mathbf{m})$ for all k and $\mathbf{m}$, as discussed in [5]. Obviously, this offset must be removed from the final dither estimate.

Since the exact computation of $\mathcal{S}_{N_o}(\mathbf{m})$ is, in general, computationally prohibitive, the algorithm proposed in [5, Sect. IV] computes an outer bound of $\mathcal{S}_{N_o}(\mathbf{m})$ in order to an keep an affordable computational complexity. We will consider in this paper the "inner polytope" algorithm [5], where $\mathcal{S}_{N_o}(\mathbf{m})$ is described by means of an n-dimensional ellipsoid. This allows to describe the feasible region with a reduced and constant number of parameters, independently of its complexity.

4.2 Joint Bayesian and Set-Membership Estimation for WOA

The uncertainty about the embedded symbols m_k invalidates the straightforward application of the estimation algorithm described in Section 4.1 to the WOA scenario. A possible solution would be to consider all the possible sequences of embedded messages so as to transform the WOA problem into p^{N_o} parallel KMA problems. Obviously, this brute force approach is not practical due to the huge number of possible message sequences, which grows exponentially with the number of observations. However, the a priori search space for the correct sequence of embedded symbols can be dramatically reduced if one considers their a posteriori probability, since certain message sequences have null or negligible probability of occurrence. This is the approach that will be followed here.

From the Appendix, we know that the a posteriori probability of a certain message sequence $\mathbf{m}$ (hereinafter, a "path") reads as

$$\Pr(\mathbf{m}|\tilde{\mathbf{y}}_1, \ldots, \tilde{\mathbf{y}}_{N_o}) = \frac{\mathrm{vol}(\mathcal{S}_{N_o}(\mathbf{m})) \cdot \Pr(\mathbf{m})}{(\mathrm{vol}(\mathcal{Z}(\Lambda)))^{N_o} \cdot \mathrm{vol}(\mathcal{V}(\Lambda)) \cdot f(\tilde{\mathbf{y}}_1, \ldots, \tilde{\mathbf{y}}_{N_o})}. \quad (15)$$

In the following we consider a priori equiprobable paths (either because no coding across different blocks takes place or because we do not know the actual coding scheme being applied), which represents the worst case for the attacker.[3] Under this assumption, the only term of (15) that depends on the hypothesized path is $\mathrm{vol}(\mathcal{V}(\Lambda))$. In practical terms, the most probable paths are those with the

[3] If the attacker had knowledge about the coding scheme being applied, he could consider the a priori probability of each path in order to simplify the estimation.

largest feasible region, $\mathcal{S}_{N_o}(\mathbf{m})$. Hence, we can define the "score" of a path $\mathbf{m}$ as $\lambda(\mathbf{m}) \triangleq \text{vol}(\mathcal{S}_{N_o}(\mathbf{m}))$, which can be used to compare the probabilities of different paths as long as they have the same length. It follows that, given N_o observations, maximum-likelihood (ML) estimation of the most probable path is given by $\hat{\mathbf{m}} = \arg\max_{\mathbf{m}} \lambda(\mathbf{m})$, and the ML estimate (if $\mathbf{T} \sim \mathcal{U}(\Lambda)$, as assumed in this paper) of the secret dither would be given by any point in $\mathcal{S}_{N_o}(\hat{\mathbf{m}})$.

A possible implementation of the proposed estimator is by means of a tree search where each branch of the tree represents a hypothesized path with a secret dither estimate associated. The tree search can be accomplished iteratively, discarding those paths with null probability, thus producing a subexponential increase in the number of feasible paths. Nevertheless, this tree search cannot be directly applied as is, in general, due to the some computational issues: 1) despite the subexponential increase in the number of feasible paths, the computational requirements may still become unaffordable; 2) as mentioned in Section 4.1, the exact computation of the feasible regions may be unfeasible in practice, except for some simple lattices. In order to overcome these computational restrictions, the following strategies are proposed.

1) Outer bounds of the feasible regions can be computed by means of the "inner polytope" algorithm [5], as mentioned in Section 4.1, providing a huge reduction of the computational complexity. However, this approximation may impact negatively the performance of the estimation algorithm because it modifies the actual scores of the paths.

2) A fast algorithm for checking null intersections (without computing the outer bound to the feasible region) can be used for speeding up the estimation procedure. An algorithm based on the OVE algorithm [10] for set-membership estimation is suited to our purposes.

3) In order to limit the number of feasible paths in each iteration, we resort to a "beam search" strategy: let $\lambda(\mathbf{m}_0)$ be the score of the most probable path. In each iteration, those paths $\mathbf{m}_i$ for which $\lambda(\mathbf{m}_0)/\lambda(\mathbf{m}_i) > \beta$ are discarded from the tree search. The parameter $\beta > 0$ is termed "beam factor" and causes a prunning of the tree by keeping only the branches with the highest probabilities. Besides the beam search strategy, an additional prunning criterion is implemented by limiting the maximum number of allowable feasible paths.

4) The a priori path space, given by $\mathcal{M}^{N_o}$, can be divided into equivalence classes (with p elements each) defined by the relation

$$\mathbf{m}_1 \sim \mathbf{m}_2 \text{ if } \mathbf{m}_2 = (\mathbf{m}_1 + j \cdot \mathbf{1}) \bmod p, \text{ for any } j = 0, \ldots, p-1, \text{ and } \mathbf{m}_1, \mathbf{m}_2 \in \mathcal{M}^{N_o}.$$

Since the paths belonging to the same equivalence class have the same a posteriori probability (recall Eq. (9)), the search space can be reduced to one representative per equivalence class, thus reducing the cardinality of the search space by a factor p without incurring in any loss of performance.

For the sake of clarity, the steps of the proposed estimation algorithm are summarized here. The input data are the observations $\{\tilde{\mathbf{y}}_i, i = 1 \ldots, N_o\}$ and the parameters of the nested lattice code.

1. Initialization: $\mathbf{m}_0 = 0$, $\mathcal{D}_1(0) = (1-\alpha)\mathcal{V}(\Lambda)$, and $K_1 = 1$, with K_1 denoting the number of feasible paths for the first observation (1 in our case). This

initialization takes into account the offset introduced in Section 4.1 and the division of $\mathcal{M}^{N_o}$ into equivalence classes.

2. For $i = 2, \ldots, N_o$
   (a) Let $\{\mathbf{m}_k,\ k = 1, \ldots, K_{i-1}\}$ be the set of feasible paths for the $i-1$ first observations. Construct a set of candidate paths as $\{\mathbf{m}_{k,l} = [\mathbf{m}_k\ l],\ k = 1, \ldots, K_{i-1},\ l = 0, \ldots, p-1\}$.
   (b) Compute the regions $\mathcal{S}_i(\mathbf{m}_{k,l})$ using $\tilde{\mathbf{v}}_r = (\tilde{\mathbf{y}}_r - \mathbf{d}_{m_{k,l}(r)} - \tilde{\mathbf{y}}_1) \mod \Lambda$, $r = 1, \ldots, i$, where $m_{k,l}(r)$ denotes the rth element of $\mathbf{m}_{k,l}$. If the inner polytope algorithm is applied, this step yields the ellipsoids that bound the true feasible regions. Prior to this step, the algorithm that checks null intersections may be applied for saving computational resources.
   (c) Compute the score $\lambda(\mathbf{m}_{k,l})$ of each path as $\mathrm{vol}(\mathcal{S}_i(\mathbf{m}_{k,l}))$. The paths with non-null score are added to the tree. If a prunning criterion is being applied, retain only those paths that fulfill the requirements. This step yields K_i paths $\{\mathbf{m}_0, \ldots, \mathbf{m}_{K_i-1}\}$ with non-null probability, termed "surviving paths".
3. The dither estimate is computed as the center of $\mathcal{S}_{N_o}(\mathbf{m}_0)$ (or its bounding region), where $\mathbf{m}_0$ is the path with the highest score (hence, the most likely) among the K_{N_o} surviving branches of the tree. The p paths belonging to the equivalence class $[\mathbf{m}_0]$ can be computed as $\mathbf{m}_k = (\mathbf{m}_0 + k\cdot\mathbf{1}) \mod p,\ k \in \mathcal{M}$, and the p corresponding dither estimates are given by $\hat{\mathbf{t}}_k = (\hat{\mathbf{t}}_0 + \mathbf{d}_k + \tilde{\mathbf{y}}_1) \mod \Lambda,\ k \in \mathcal{M}$, where $\hat{\mathbf{t}}_0$ is the dither estimate associated to the path $\mathbf{m}_0$. Note that $\tilde{\mathbf{y}}_1$ is added for canceling the offset introduced in Step 2-b.

4.3 Experimental Results

This section presents the results of applying the estimation algorithm proposed in 4.2 over some practical schemes. The experiments have been carried out under the following assumptions: the host signals follow a Gaussian distribution with zero mean and variance $\sigma_X^2 = 10$, and the DWR is 30 dB in all cases ($\mathrm{DWR} \triangleq 10\log_{10}(\sigma_X^2/D_w)$); the embedded messages are equiprobable (i.e., no coding is applied along different blocks), and the attacker knows the parameters of the nested lattice code being used, as stated in Section 2. In all cases, a beam factor $\beta = 10^{45/10}$ has been used, and the maximum number of feasible paths was limited to 250. The performance of the estimator is measured in terms of the mean squared error (MSE) per dimension between the dither estimate and the actual dither. In order to compute the MSE without ambiguities (due to the existence of p equiprobable paths), it is assumed that the message conveyed by the first observation corresponds to the symbol 0.

Fig. 4 shows the results obtained for a scheme using a cubic shaping lattice in 10 dimensions and repetition coding (see Section 3.2) with $\alpha = 0.6$. In this case, the simplicity of the feasible regions allows to compute them exactly. It can be seen that for $p = 4$ is still possible to attain the same accuracy as in the KMA scenario, whereas for $p = 7$ and $p = 10$ a significant degradation of the MSE is observed. This degradation is a consequence of the fact that, as p is increased, the probability of decoding the correct path decreases. In the experiments, the

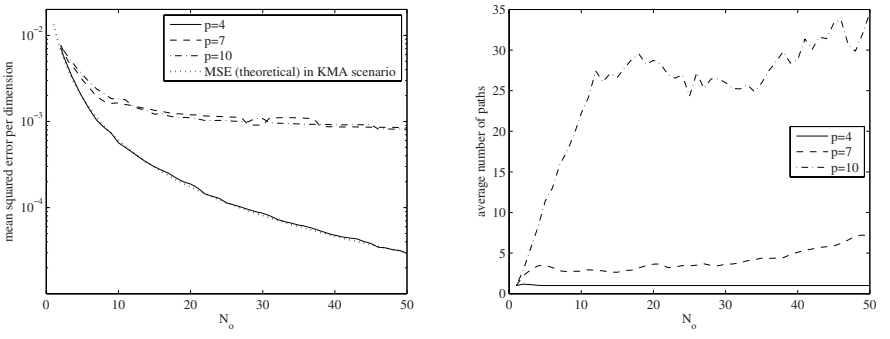


Fig. 4. MSE per dimension (a) and average number of surviving paths in the tree search (b) using cubic Λ with repetition coding ($n = 10$), $\alpha = 0.6$, and different embedding rates

probability of choosing an incorrect path has been shown to be around 0.05 and 0.1 for $p = 7$ and $p = 10$, respectively. The average number of surviving paths in the tree search is plotted in Fig. 4 for illustrating the complexity of the search procedure. In this regard, it can be seen that even in a difficult case as $p = 10$ with $\alpha = 0.6$, the tree search can still be performed with low complexity.

Fig. 5(a) shows the results obtained for a hexagonal shaping lattice and $\alpha = 0.7$, using the inner polytope algorithm in order to compute the approximate feasible regions. Notice that, although α is higher than in the former case, the maximum embedding rate considered now is substantially larger: $\frac{1}{2}\log_2(9)$ bits vs. $\frac{1}{10}\log_2(10)$ bits (the case with $p = 9$ corresponds to the lattice code shown in Fig. 2(b)). Similar comments as above apply in this case: increasing p degrades the MSE, and the spurious peaks in the plots are due to incorrect decisions about the actual path. Finally, Fig. 5(b) shows the results obtained for the E_8 shaping lattice [8], the best lattice quantizer in 8 dimensions.

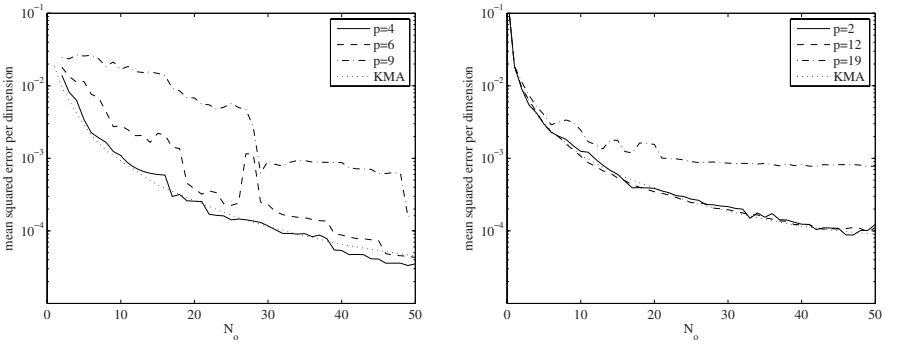
An accurate dither estimate (subjected to an unknown modulo-Λ shift, as the one obtained here) allows to implement a number of harmful attacks. We are going to focus on a reversibility attack as follows: based on a dither estimate $\hat{\mathbf{t}}$ and an estimated path $\hat{\mathbf{m}}$, the host vector corresponding to the kth watermarked block can be computed as

$$\hat{\mathbf{x}}_k = \mathbf{y}_k - \frac{\alpha}{1-\alpha}(Q_\Lambda(\mathbf{y}_k - \mathbf{d}_{\hat{m}_k} - \hat{\mathbf{t}}) - \mathbf{x}_k + \mathbf{d}_{\hat{m}_k} + \hat{\mathbf{t}}). \tag{16}$$

It is interesting to notice that the ambiguity in the estimated message does not affect negatively the host estimation whenever the estimated path $\hat{\mathbf{m}}$ fulfills

$$\hat{\mathbf{m}} = (\mathbf{m} + k \cdot \mathbf{1}) \mod p, \quad \text{for any } k \in \mathcal{M}, \tag{17}$$

being $\mathbf{m}$ the actual embedded path. The reason is that the dither estimate associated to any of those paths yields the same fine lattice Λ_f, and thus it is



(a) (b)

Fig. 5. MSE per dimension for $\alpha = 0.7$ and different embedding rates. Results for hexagonal ($n = 2$) (a) and E_8 ($n = 8$) (b) shaping lattices.

(a) (b)

Fig. 6. Illustration of a reversibility attack based on dither estimate according to Eq. (16). Image watermarked using $\Lambda = E_8$, $\alpha = 0.7$, $p = 10$ and PSNR = 38.2 dB (a), and estimate of the original image with PSNR = 55.9 dB (b).

valid for performing a successful reversibility attack. Fig. 6 shows the result of implementing this attack on a real watermarked image. The parameters of the watermarking algorithm are the same as above, and the watermark is embedded in the low frequency coefficients of 8×8 non-overlapping DCT blocks, yielding a PSNR = 38.2 dB. The resulting host estimate, shown in Fig. 6(b), presents a PSNR of approximately 56 dB. Nevertheless, if each pixel value of this host estimate is rounded off to the closest integer, then the PSNR goes to ∞.

5 Conclusions

We have presented in this paper an investigation of the security provided by data hiding schemes based on nested lattice codes randomized by means of secret dithering. It has been shown that, although it is theoretically possible to achieve perfect secrecy, the security level of many practical scenarios (i.e., simple shaping lattices, low embedding rates) can be fairly low. In fact, the security holes of the data hiding schemes studied in this paper have been shown to be exploitable in practice with affordable complexity, allowing for instance to reverse the watermarking process with high fidelity. In general, the information leakage about the secret dither can be reduced by increasing the embedding rate, but this solution demands for more powerful error correcting codes (ECC) if one wants to guarantee reliable transmission. A possible drawback, as noted in this paper, is that the use of ECCs introduces statistical dependence between different observations that could be exploited by an attacker, specially for simple ECCs. The complexity of exploiting the information leakage provided by channel coding deserves further attention in future works.

References

1. Cayre, F., Fontaine, C., Furon, T.: Watermarking security: theory and practice. IEEE Trans. Signal Processing 53 (2005)
2. Comesaña, P., Pérez-Freire, L., Pérez-González, F.: Fundamentals of data hiding security and their application to spread-spectrum analysis. In: Barni, M., Herrera-Joancomartí, J., Katzenbeisser, S., Pérez-González, F. (eds.) IH 2005. LNCS, vol. 3727, Springer, Heidelberg (2005)
3. Pérez-Freire, L., Comesaña, P., Troncoso-Pastoriza, J.R., Pérez-González, F.: Watermarking security: a survey. Transactions on Data Hiding and Multimedia Security I 4300, 41–72 (2006)
4. Moulin, P., Koetter, R.: Data hiding codes. Proceedings of IEEE 93, 2083–2126 (2005)
5. Pérez-Freire, L., Pérez-González, F., Furon, T., Comesaña, P.: Security of lattice-based data hiding against the known message attack. IEEE Transactions on Information Forensics and Security 1, 421–439 (2006)
6. Cover, T.M., Thomas, J.A.: Elements of Information Theory. Wiley series in Telecommunications (1991)
7. Erez, U., Zamir, R.: Achieving 1/2log(1+SNR) over the Additive White Gaussian Noise channel with lattice encoding and decoding. IEEE Transactions on Information Theory 50, 2293–2314 (2004)
8. Conway, J.H., Sloane, N.J.A.: Sphere packings, lattices and groups. third edn, 3rd edn. Comprehensive Studies in Mathematics, vol. 290. Springer, New York (1999)
9. Comesaña, P., Pérez-González, F., Balado, F.: On distortion-compensated dither modulation data-hiding with repetition coding. IEEE Transactions on Signal Processing 54, 585–600 (2006)
10. Cheung, M.F., Yurkovich, S., Passino, K.M.: An optimal volume ellipsoid algorithm for parameter set estimation. IEEE Transactions on Automatic Control 38, 1292–1296 (1993)

Appendix

A Posteriori Probability of the Message Sequences

In order to compute the probability a posteriori of a message sequence $\mathbf{m} = (m_1, \ldots, m_{N_o})$ (hereinafter, a "path"), this probability is first rewritten using Bayes' rule:

$$\Pr(m_1, \ldots, m_{N_o} | \tilde{\mathbf{y}}_1, \ldots, \tilde{\mathbf{y}}_{N_o}) = \frac{f(\tilde{\mathbf{y}}_1, \ldots, \tilde{\mathbf{y}}_{N_o} | m_1, \ldots, m_{N_o}) \cdot \Pr(m_1, \ldots, m_{N_o})}{f(\tilde{\mathbf{y}}_1, \ldots, \tilde{\mathbf{y}}_{N_o})}. \tag{18}$$

The a posteriori probability of the observations can be factored as:

$$\begin{aligned} f(\tilde{\mathbf{y}}_1, \ldots, \tilde{\mathbf{y}}_{N_o} | m_1, \ldots, m_{N_o}) &= \prod_{k=1}^{N_o} f(\tilde{\mathbf{y}}_k | m_1, \ldots, m_k, \tilde{\mathbf{y}}_1, \ldots, \tilde{\mathbf{y}}_{k-1}) \\ &= \prod_{k=1}^{N_o} \int_{\mathcal{V}(\Lambda)} f(\tilde{\mathbf{y}}_k | m_k, \mathbf{t}) \cdot f(\mathbf{t} | \tilde{\mathbf{y}}_1, \ldots, \tilde{\mathbf{y}}_{k-1} | m_1, \ldots, m_{k-1}) d\mathbf{t}. \end{aligned} \tag{19}$$

In order to compute each factor of (19), we will resort to the flat-host assumption, which implies that $f(\tilde{\mathbf{y}}_k | m_k, \mathbf{t}) = f_0(\tilde{\mathbf{y}}_k - \mathbf{d}_{m_k} - \mathbf{t} \mod \Lambda)$. Thus, each factor of (19) can be seen as a circular convolution over $\mathcal{V}(\Lambda)$:

$$f(\tilde{\mathbf{y}}_1, \ldots, \tilde{\mathbf{y}}_{N_o} | m_1, \ldots, m_{N_o}) = \prod_{k=1}^{N_o} f_0(\tilde{\mathbf{y}}_k - \mathbf{d}_{m_k} \mod \Lambda) \circledast f(\mathbf{t} | \tilde{\mathbf{y}}_1, \ldots, \tilde{\mathbf{y}}_{k-1}, m_1, \ldots, m_{k-1})$$

Furthermore, under the assumption that $\mathbf{T} \sim U(\mathcal{V}(\Lambda))$, we have that the conditional pdf of the dither is given by Eq. (4). By combining (2) and (4), it can be seen that the integrand of the kth factor in (19) is given by

$$\begin{cases} (\mathrm{vol}(\mathcal{Z}(\Lambda)) \cdot \mathrm{vol}(\mathcal{S}_{k-1}(\mathbf{m})))^{-1}, & \mathbf{t} \in \mathcal{S}_{k-1}(\mathbf{m}) : (\tilde{\mathbf{y}}_k - \mathbf{d}_{m_k} - \mathbf{t}) \mod \Lambda \in \mathcal{Z}(\Lambda) \\ 0, & \text{otherwise.} \end{cases}$$

The condition on $\mathbf{t}$ in the equation above is equivalent to $\mathbf{t} \in S_{k-1}(\mathbf{m}) : \mathbf{t} \in (\tilde{\mathbf{y}}_k - \mathbf{d}_{m_k} - \mathcal{Z}(\Lambda)) \mod \Lambda$, so each factor in (19) is proportional to the volume of $\mathcal{S}_k(\mathbf{m}) = \mathcal{S}_{k-1}(\mathbf{m}) \cap \mathcal{D}_k(m_k)$. Finally, Eq. (19) can be succinctly expressed as

$$\begin{aligned} f(\tilde{\mathbf{y}}_1, \ldots, \tilde{\mathbf{y}}_{N_o} | m_1, \ldots, m_{N_o}) &= \prod_{k=1}^{N_o} \frac{\mathrm{vol}(\mathcal{S}_k(m_1, \ldots, m_k))}{\mathrm{vol}(\mathcal{Z}(\Lambda)) \cdot \mathrm{vol}(\mathcal{S}_{k-1}(m_1, \ldots, m_{k-1}))} \\ &= \frac{\mathrm{vol}(\mathcal{S}_{N_o}(m_1, \ldots, m_{N_o}))}{(\mathrm{vol}(\mathcal{Z}(\Lambda)))^{N_o} \cdot \mathrm{vol}(\mathcal{V}(\Lambda))}. \end{aligned} \tag{20}$$