Zrelli *et al. EURASIP Journal on Wireless Communications and Networking* (2015) 2015:43
DOI 10.1186/s13638-015-0294-5

EURASIP Journal on
Wireless Communications and Networking
a SpringerOpen Journal

RESEARCH Open Access

# Blind identification of code word length for non-binary error-correcting codes in noisy transmission

Yasamine Zrelli[1,2†], Roland Gautier[1,2*†], Eric Rannou[1,3†], Mélanie Marazin[1,2†] and Emanuel Radoi[1,2†]

**Abstract**

In cognitive radio context, the parameters of coding schemes are unknown at the receiver. The design of an intelligent receiver is then essential to blindly identify these parameters from the received data. The blind identification of code word length has already been extensively studied in the case of binary error-correcting codes. Here, we are interested in non-binary codes where a noisy transmission environment is considered. To deal with the blind identification problem of code word length, we propose a technique based on the Gauss-Jordan elimination in GF($q$) (Galois field), with $q = 2^m$, where $m$ is the number of bits per symbol. This proposed technique is based on the information provided by the arithmetic mean of the number of zeros in each column of these matrices. The robustness of our technique is studied for different code parameters and over different Galois fields.

**Keywords:** Cognitive radio; Blind identification; Non-binary error-correcting codes; Galois field

## Introduction

Error-correcting codes are frequently used in modern digital transmission systems in order to improve the communication quality. These codes are designed to achieve a good immunity against channel impairments by introducing redundancy in the informative data. Due to the complexity of both encoding and especially decoding procedures, the majority of research and practical implementations of real-time embedded systems were often restricted to encoders manipulating binary data, i.e., elements of the Galois field GF(2). Over the last decade, low-density parity check (LDPC) codes and turbo codes over GF(2) have attracted considerable interest of many researchers due to their excellent error correction capability. They have been generalized to finite fields GF($q$) [1,2], where $q = 2^m$, and are among the most widely used error-correcting codes in wireless communication standards. It has been shown in [1] that non-binary LDPC codes perform generally better than binary LDPC codes and turbo codes. However, the major drawback of these codes is their decoding complexity for a large Galois field order $q$ [3,4]. Low complexity decoding algorithms have recently been proposed [5,6], thus allowing the use of non-binary LDPC codes in practical implementations.

Our main research interests are focused on non-binary error-correcting codes in order to blindly identify their parameters. This topic is a part of a non-cooperative context like a military interception or cognitive radio applications. In this case, the receiver has no knowledge about the parameters used to encode the information at the transmitter. The solution is to design an intelligent receiver which is able to blindly identify the encoder parameters from the only knowledge of the received data stream. This blind identification function of the receiver permits to increase the data rate transmission, since it will be unnecessary to transmit supplementary information about the encoder parameters with the useful data. Such intelligent receiver is able to adapt automatically itself to the development of new high-performance coding schemes and the fast evolution of new communication standards without equipment change. In this work, we are only interested

*Correspondence: roland.gautier@univ-brest.fr
†Equal contributors
[1] Université Européenne de Bretagne, 5 Boulevard Laënnec, 35000 Rennes, France
[2] Université de Brest; CNRS, UMR 6285 Lab-STICC, 6 avenue Victor Le Gorgeu, 29238 Brest, France
Full list of author information is available at the end of the article

in blindly identifying the code word length of linear non-binary block codes. In the case of the interception, this parameter can not be transmitted. Likewise, if we want to change the encoder or get out of the list of possible choice of encoders, the code word length is not transmitted.

In this context, the published research results have been restricted so far to the blind recognition of the code word length of binary codes. To the best of our knowledge, this paper introduces, for the first time, an approach to blindly identify the code word length of non-binary codes in noisy conditions. In this work, the aim is to blindly identify the code word length from the only knowledge of received data. The authors in [7] proposed a technique of identification of non-binary LDPC parameters, but the identification is not blind because it is based on using a predefined candidate set of encoders which is known by both the transmitter and the receiver. Furthermore, this technique only works with LDPC codes unlike our proposed technique, which is general and suitable for all block codes. In our paper, the proposed blind identification technique is based on a generalization of an existing method used for binary codes. The principle of this generalization will be explained in this paper without specifying in details its detection performances. So, we present here state-of-the-art techniques to identify the code word length of binary linear block codes. The idea of these techniques is to find a basis of a dual code composed of parity check relations. For this purpose, an approach based on finding code words of small Hamming weight [8,9] was improved by Valembois [10] by using statistical hypothesis tests and recently by Cluzeau [11,12] and Côte [13]. A second approach based on linear algebra theory was introduced in [14] for noiseless channel. This approach permits to recover the length of code words by studying behaviors of the rank of matrices composed of received bits. However, the rank criterion was exploited without providing an algebraic and theoretical justification of such behavior. In [15], the use of this criterion was justified. In [16], the rank criterion approach was generalized to convolutional codes over GF($q$), where $q > 2$, assuming a noiseless transmission, but it was shown that this generalized technique can be also performed to non-binary linear block codes. In noisy transmissions, a technique based on the Gauss elimination in GF(2) was applied in [17-19] to matrices composed of noisy received bits in order to find the number of almost dependent columns permitting the identification of the code word length in the case of binary error-correcting codes. Indeed, an almost dependent column of a matrix composed of noisy received symbols corresponds to a column which may be a linear combination of some preceding columns without the presence of erroneous symbols and which leads to a column that contains more zero elements after the Gauss elimination.

Compared to previous works, we demonstrate here that it is possible to generalize the blind identification technique proposed in [17,18] to non-binary block codes provided that the Galois field parameters (the cardinality and the primitive polynomial) are known by the receiver. To identify the primitive polynomial, an algorithm of identification was proposed in [20]. To achieve our purpose, it is necessary to identify the number of almost dependent columns in the matrices composed of noisy symbols of GF($q$) by studying the probability of detection of these columns, denoted as $P_i$. In fact, the computation of $P_i$ is essential in order to determine an optimal detection threshold. Assuming a transmission over $q$-ary symmetric channel with an error probability $p_e$, the techniques based on finding a base of a dual code [18,19] for binary codes require the knowledge of $p_e$, where a hard decision demodulation is considered. For this reason, we propose here an approach which is more robust because it allows us the blind identification of the code word length of non-binary and binary block codes without using the error probability $p_e$. This approach is based on analyzing behaviors of the arithmetic mean of the number of zeros in the columns of the matrices constructed by the Gauss elimination in GF($q$). In this paper, the proposed method is a general method that should be applied to all non-binary block codes even though most examples of codes given here are non-binary LDPC codes. For this reason, the properties of LDPC codes are not exploited by our method.

This paper is organized as follows. In the 'Technical background' section, we present the encoding process of non-binary error-correcting codes. Then, the principle of the blind identification of code parameters in the noiseless case is described. The channel model used in this study is also defined and justified in this section. In the 'Blind identification of code word length in the noisy case' section, the blind identification method of the code word length in noisy environment is described. A comparison in terms of error probability and detection performances is shown in the 'Analysis and performances' section. Finally, some conclusions are drawn in the 'Conclusions' section and planned future work is pointed out.

## Technical background

### Non-binary error-correcting codes

The use of an efficient coding system in the transmitter as error-correcting codes is essential in order to fight disturbances present on the transmission channel. For a long time, cyclic codes such as BCH codes [21,22] and Reed-Solomon codes [23] have been the most commonly used as codes based on finite fields since they are characterized by large minimum distances for a hard decision decoding. The non-binary LDPC codes described by a sparse parity check matrix with elements in GF($q$) have been developed by Davey and MacKay in 1998 [1]. Significant works on the design and the decoding complexity reduction of these codes have shown that they have a great potential to replace Reed-Solomon codes in some applications of communication, such as space communications [24], and storage systems [25,26]. In this paper, we focus on the blind identification of code word length for the non-binary block codes, but this proposed method can also be applied to convolutional codes and concatenated codes.

Let us present the encoding process of these codes over GF($q$). Actually, the principle of a transmission chain is to send digital information from a source to one or more receivers. The information yielded by the source is binary data $\{0, 1\} = \mathrm{GF}(2)$. Each block of $m$ information bits are combined to generate a symbol of GF($q$). Then, the generated non-binary information, denoted as **d**, is encoded by one of the block codes over GF($q$) listed above. For most block error-correcting codes, a code word, denoted as **c**, composed of non-binary symbols is obtained by the multiplication of the information **d** and a non-binary generator matrix **G**:

$$\mathbf{c} = \mathbf{d} \cdot \mathbf{G} \tag{1}$$

In the case of LDPC codes, the encoding process needs the use of the parity check matrix, which is always sparse compared to the other codes.

In most of the standards, such as long-term evolution (LTE) standard [27], the encoding is performed in a systematic form in order to facilitate the decoding process without degrading performances of the error correction. For this reason, in the case of block codes, the required parameters to perform the decoding operation are the number of inputs, denoted as $k$, the code word length, denoted as $n$, and a parity check matrix, denoted as **H**. Indeed, the matrix **H** will be used by the decoder to detect or/and to correct the errors. The recovered information will be the first $k$ symbols of the recovered code word due to the systematic form used in the encoding. Our aim in this research work is to blindly identify the parameter $n$ from non-binary received symbols which are affected by noisy transmissions. In the noiseless context, we have already demonstrated in [16] that we can identify this parameter with the only knowledge of the received data, provided that the Galois field parameters are known. The principle of blind identification of the code parameter $n$ in the noiseless case is recalled in the following subsection.

### Principle of blind identification method of code word length in the noiseless case

In this part, we assume that the channel introduces no error. In [16], we have adapted the method proposed in [28] to identify the parameters of convolutional codes over GF($q$), where $q = 2^m$. We have shown that our method for the noiseless case can be applied to block codes. This method reshapes row-wise the received symbols, denoted as **r**, under a matrix form, denoted as $\mathbf{R}_l$, of size $(M \times l)$. Indeed, $\mathbf{R}_l$ is filled by received symbols from the top left corner to the bottom right as illustrated in Figure 1.

The number of columns $l$ varies between 1 and $l_{\max}$ and the number of rows $M$ which depends on $l$ is given by the integer part $\lfloor \frac{L}{l} \rfloor$ where $L$ is the length of a received symbol stream. Then, the rank over GF($q$) is calculated for each matrix $\mathbf{R}_l$. When all matrices $\mathbf{R}_l$ have full rank, it is impossible to detect the existence of a code. Nevertheless, the redundancy introduced by the code leads to rank deficiencies in some matrices $\mathbf{R}_l$. Henceforth, the rank behaviors of $\mathbf{R}_l$ allow us to detect the code and to identify its parameters, in particular the code word length. As demonstrated in [15] and studied in [16], there are two possible rank behaviors according to the number of columns $l$. If $l$ is a multiple of $n$ (i.e., $l = \alpha \cdot n$, $\alpha \in \mathbb{N}$), the ranks of the matrices $\mathbf{R}_l$ are proportional to the code rate $k/n$ (i.e., $\mathrm{rank}(\mathbf{R}_l) = l \cdot k/n$). Otherwise (i.e., $l \neq \alpha \cdot n$), $\mathbf{R}_l$ have full rank (i.e., $\mathrm{rank}(\mathbf{R}_l) = l$). Thus, the value of the rank deficiency depends on code parameters ($k$ and $n$). Indeed, only two consecutive rank deficiencies are necessary to determine all code parameters. The code word length $n$ can be determined by the difference between two values of $l$ corresponding to two consecutive rank deficiencies of $\mathbf{R}_l$. As shown in [16], the rank method gives good results in a noiseless environment. A theoretical and algebraic study of the behavior of the rank criterion, as well as particular cases which can occur for specific parameters of

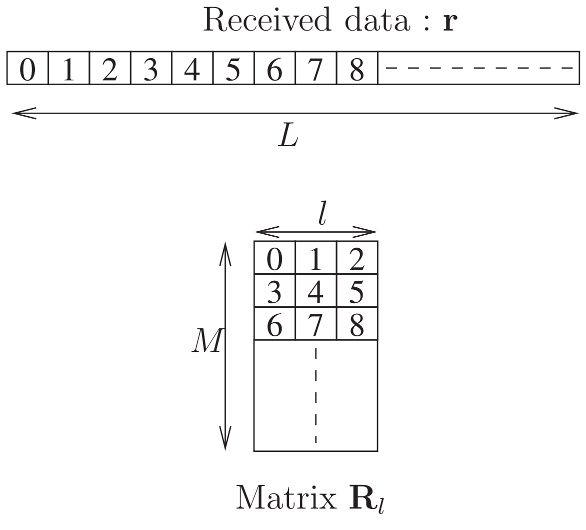

**Figure 1 Example of a matrix $\mathbf{R}_l$.** Example of a matrix construction from the received symbols.

codes, were presented in [15]. It was demonstrated that most matrices $\mathbf{R}_l$ have full rank when $l$ is not a multiple of $n$, except for some particular cases which depend on codes (generator matrix). In a noisy environment, the rank method can not be used, since all the matrices $\mathbf{R}_l$ have full rank in this case.

### Non-binary channel

In order to evaluate our blind identification algorithm, we assume that the encoded sequences are transmitted through a $q$-ary (non-binary, for $q = 2^m > 2$) symmetric channel (QSC) which is the simplest channel. However, our proposed algorithms can work for every type of channel provided that the error probability $p_e$ computed at the output of the demodulator is known. Indeed, we consider that the blocks of the transmission chain, the modulator, the transmission channel, and the demodulator can be modeled by a non-binary channel, where a hard decision demodulation is considered. In a cognitive radio context, a multipath fading channel is used. This realistic channel leads to burst errors which can be corrected by using an interleaver and error-correcting codes. In this context, the errors at the output of a deinterleaver at the receiver side can be modeled by a QSC when a decoding process with hard decision will be used. The problem of a blind identification of the interleaver period, as well as a blind synchronization with the interleaver blocks was handled in [14,18].

Let us define the $q$-ary symmetric channel which is the generalization of the binary symmetric channel (BSC). In fact, it is a discrete memoryless channel with an error probability $p_e$ and composed of non-binary inputs and non-binary outputs belonging to the GF($q$), where $q = 2^m$. The symbols at the input of the channel are independent and distributed uniformly with a probability equal to $1/q$. A symbol $\delta \in \mathrm{GF}(q)$ at the channel input is received incorrectly with a probability $p_e/(q-1)$ [29]. In other words, it is replaced at the receiver by a different symbol $\beta$ of GF($q$). The probability of correctly receiving a symbol is equal to $1 - p_e$. The QSC channel is characterized by the conditional probabilities:

$$\begin{aligned} p(\tilde{r}_i = \beta | r_i = \delta) &= \frac{p_e}{q-1},\ \delta \neq \beta \\ p(\tilde{r}_i = \delta | r_i = \delta) &= 1 - p_e \end{aligned} \tag{2}$$

where the transmitted symbol is denoted $r_i$, i.e., $r_i = c_i$, for $i \in \{1, \cdots, L\}$, and the noisy received symbol is denoted $\tilde{r}_i$ such that $\tilde{r}_i = r_i + e_i$ with $e_i$ the transmission error introduced in the symbol $r_i$. An example of a non-binary symmetric channel for $q = 2^2$ is depicted in Figure 2.

In the following section, we present the blind identification method of the parameter $n$ in a noisy framework.

## Blind identification of code word length in the noisy case

In this part, we present the implementation method which allows us to identify the code word length of a non-binary code in a noisy environment. This method is based on the concept of finding the rank-deficient matrices among $\tilde{\mathbf{R}}_l, \forall l \in \{1, \ldots, l_{\max}\}$, corresponding to matrices having at least one almost dependent column. Indeed, the matrices $\tilde{\mathbf{R}}_l$ are reshaped in the same way as $\mathbf{R}_l$ using the noisy received symbols $\tilde{r}_i$. In [19], a method devoted to determine these matrices in the case of binary codes was presented. However, this method requires the knowledge of the error probability $p_e$. In order to avoid this constraint, we propose a method based on using the arithmetic mean criterion in order to detect the rank-deficient matrices which have some almost dependent columns without the need of the error probability $p_e$.

### Principle

In a noiseless case, the rank criterion is used to find the maximum number of linearly independent columns in the matrices $\mathbf{R}_l$. This allows us to derive the number of linearly dependent columns in $\mathbf{R}_l$ (columns which are linear combinations of other columns). The finite-field Gauss

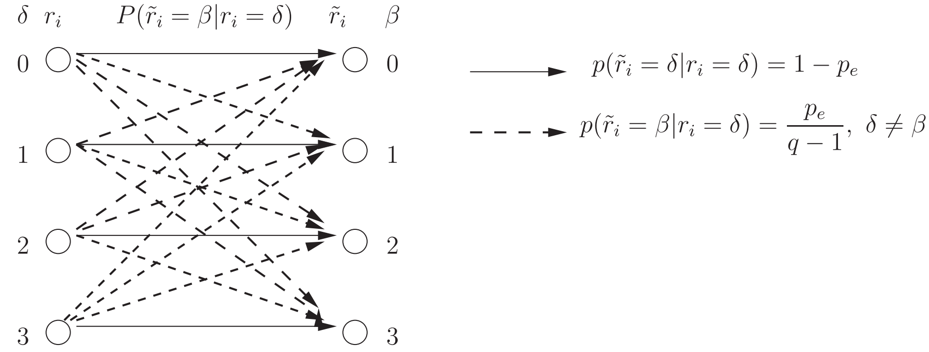
$\delta$
$r_i$
$P(\tilde{r}_i = \beta | r_i = \delta)$
$\tilde{r}_i$
$\beta$
0
1
2
3
0
1
2
3
$p(\tilde{r}_i = \delta | r_i = \delta) = 1 - p_e$
$p(\tilde{r}_i = \beta | r_i = \delta) = \frac{p_e}{q-1},\ \delta \neq \beta$

**Figure 2 Non-binary symmetric channel.** An example of a non-binary symmetric channel for $q = 2^2$.

elimination method [30] has to be used to eliminate those linear-dependent columns to zero. In noisy transmissions, all matrices $\tilde{\mathbf{R}}_l$ have full rank. A matrix $\tilde{\mathbf{R}}_l$ can be expressed according to $\mathbf{R}_l$ by:

$$\tilde{\mathbf{R}}_l = \mathbf{R}_l + \mathbf{E}_l \tag{3}$$

where $\mathbf{E}_l$ is the error matrix of size $(M \times l)$ constructed in the same way as $\mathbf{R}_l$ using the errors induced by the channel. Therefore, the dependence of the columns is disturbed by the presence of errors in some received symbols. In such context, the authors in [17,18] proposed to look for the number of almost dependent columns in the matrices composed of noisy received bits by using the Gauss elimination over GF(2). Inspired by this idea, it is sufficient, in the case of non-binary error correcting codes, to apply the finite-field Gauss elimination in GF($q$) to $\tilde{\mathbf{R}}_l$ in order to obtain a new matrix $\tilde{\mathbf{T}}_l$ of size $(M \times l)$. This algorithm gives also at output a matrix of size $(l \times l)$, denoted $\tilde{\mathbf{A}}_l$, that describes the combination operations performed to the columns of the matrix $\tilde{\mathbf{R}}_l$ in order to obtain the transformation matrix $\tilde{\mathbf{T}}_l$. A recall of the finite-field Gauss elimination over GF($q$) is presented in Algorithm 1. To describe this algorithm, we denote $\mathbf{I}_l$ the identity matrix of size $(l \times l)$, $\mathbf{x}_i^{(l)}$ the $i$-th column of a given matrix $\mathbf{X}_l$ and $x_i^{(l)}(j)$ a coefficient of a matrix $\mathbf{X}_l$ placed in the $i$-th column and in the $j$-th row.

**Algorithm 1** The finite-field Gauss elimination over GF($q$).

```
Require: $\tilde{\mathbf{R}}_l$
Ensure: $\tilde{\mathbf{T}}_l$ and $\tilde{\mathbf{A}}_l$
  Initialization: $\tilde{\mathbf{T}}_l \leftarrow \tilde{\mathbf{R}}_l$ and $\tilde{\mathbf{A}}_l \leftarrow \mathbf{I}_l$
  for $i = 1$ to $l$ do
    if the $i$-th element of the $i$-th column $\tilde{t}_i^{(l)}(i) = 0$ then
      Permute the $i$-th column $\tilde{\mathbf{t}}_i^{(l)}$ with the first column $\tilde{\mathbf{t}}_{i'}^{(l)}$
      ($i' > i$) that has a non-zero on its $i$-th element
      Permute the column $\tilde{\mathbf{a}}_i^{(l)}$ with $\tilde{\mathbf{a}}_{i'}^{(l)}$
    end if
    Multiply the columns $\tilde{\mathbf{t}}_i^{(l)}$ and $\tilde{\mathbf{a}}_i^{(l)}$ by $\nu = 1/\tilde{t}_i^{(l)}(i)$ in order
    to have $\tilde{t}_i^{(l)}(i) = 1$
    for $j = i + 1$ to $l$ do
      Let $b = \tilde{t}_j^{(l)}(i)$. Apply the following operation to the
      columns $\tilde{\mathbf{t}}_i^{(l)}$ and $\tilde{\mathbf{t}}_j^{(l)}$ in order to have $\tilde{t}_j^{(l)}(i) = 0$:

        $\tilde{\mathbf{t}}_j^{(l)} = \tilde{\mathbf{t}}_j^{(l)} - b \cdot \tilde{\mathbf{t}}_i^{(l)}$

      Apply the following operations to the columns of the
      matrix $\mathbf{A}_l$:

        $\tilde{\mathbf{a}}_j^{(l)} = \tilde{\mathbf{a}}_j^{(l)} - b \cdot \tilde{\mathbf{a}}_i^{(l)}$

    end for
  end for
```

By means of this algorithm, the linear-dependent columns in the matrix will be eliminated to zeros. The whole matrix is considered in our proposed method instead of only the lower part of the matrix $\tilde{\mathbf{R}}_l$ as mentioned in [17]. It would be more accurate than assuming that errors do not occur in the upper part of the matrix, but it is not the real case.

We can note that the finite-field Gauss elimination over GF($q$) can be defined by a linear application given by:

$$\tilde{\mathbf{R}}_l \cdot \tilde{\mathbf{A}}_l = \tilde{\mathbf{T}}_l \tag{4}$$

In noiseless transmissions, the number of dependent columns in $\mathbf{R}_l$, for $l = \alpha \cdot n$, $\alpha \in \mathbb{N}$, corresponds to the number of the zero columns in the matrix $\mathbf{T}_l$ which is the result of the transformation of $\mathbf{R}_l$ by the finite-field Gauss elimination in GF($q$) ($\mathbf{R}_l \cdot \mathbf{A}_l = \mathbf{T}_l$). The matrix form of $\mathbf{T}_l$ is described in Figure 3.

In fact, the dimension identification of a vector space generated by a code $\mathcal{C}$ is equivalent to finding the

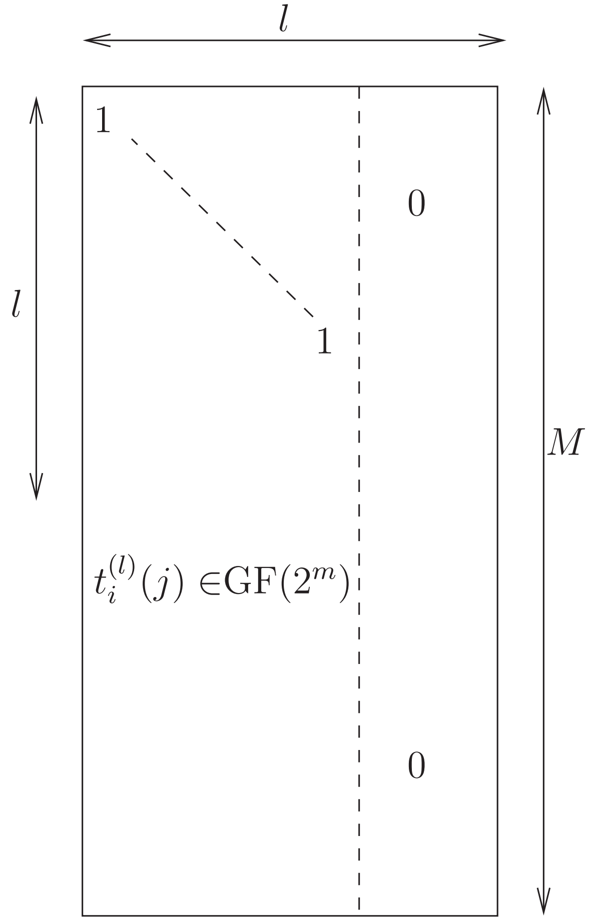

**Figure 3 Matrix form of $\mathbf{T}_l$.** An example of a matrix form of $\mathbf{T}_l$ for $l = \alpha \cdot n$.

dimension of a vector space generated by its dual code $\mathcal{C}^{\perp}$. For any vector $\mathbf{h}$ belonging to $\mathcal{C}^{\perp}$ and for any code word $\mathbf{r}$ of $\mathcal{C}$, the relation between both is defined by $\mathbf{r} \cdot \mathbf{h}^T = 0$. In noiseless conditions, the matrix $\mathbf{R}_n$, for $l = n$, which is composed of $M$ code words of length $n$, should satisfy:

$$\mathbf{R}_n \cdot \mathbf{h}^T = \mathbf{0} \tag{5}$$

We can note that $\mathbf{h}$ belongs to the kernel of $\mathbf{R}_n$, denoted as $\ker(\mathbf{R}_n)$. So, we have $\mathcal{C}^{\perp} \subset \ker(\mathbf{R}_n)$. Since the dependent columns in $\mathbf{R}_l$ multiplied by the columns $\mathbf{a}_i^{(l)}$ permit to have the zero columns in the matrix $\mathbf{T}_l$, the corresponding columns $\mathbf{a}_i^{(n)}$ will belong to $\ker(\mathbf{R}_n)$ in which the dual code $\mathcal{C}^{\perp}$ is contained. Therefore, finding the dependent columns in $\mathbf{R}_l$ is equivalent to finding the columns $\mathbf{a}_i^{(l)}$ which belong to the dual code $\mathcal{C}^{\perp}$.

Due to the presence of errors induced by the channel in $\tilde{\mathbf{R}}_l$, for $l = \alpha \cdot n$, the columns of $\tilde{\mathbf{T}}_l$ corresponding to the almost dependent columns in $\tilde{\mathbf{R}}_l$ will contain some non-zero symbols. Assuming that the first $l$ rows and the pivots of the matrix $\tilde{\mathbf{T}}_l$ do not contain transmission errors, using (3) and (4) allows us to write the matrix $\tilde{\mathbf{T}}_l$ as:

$$\tilde{\mathbf{T}}_l = \mathbf{T}_l + \mathbf{E}_l \cdot \mathbf{A}_l \tag{6}$$

In this case, a vector $\mathbf{h}$ is a parity check relation (i.e., $\mathbf{h} \in \mathcal{C}^{\perp}$) with high probability if the relation $\tilde{\mathbf{R}}_l \cdot \mathbf{h}^T$ has a low Hamming weight [11]. However, the opposite is not necessarily true. We can conclude that $\tilde{\mathbf{a}}_i^{(l)}$ belongs to $\mathcal{C}^{\perp}$ if the corresponding $\tilde{\mathbf{t}}_i^{(l)} = \tilde{\mathbf{R}}_l \cdot \tilde{\mathbf{a}}_i^{(l)}$ has a small Hamming weight. In GF($q$), the Hamming weight of a vector is the number of non-zero elements in this vector. So, our aim is to determine the columns $\tilde{\mathbf{t}}_i^{(l)}$ which have a high number of zeros. The idea is to study the number of zeros in the columns of the $\tilde{\mathbf{T}}_l$ in order to detect the almost dependent columns in $\tilde{\mathbf{R}}_l$.

### *Behaviors of the number of zeros in the columns of* $\tilde{\mathbf{T}}_l$

Let $B_l(i)$ be the number of zeros in the $i$-th column of $\tilde{\mathbf{T}}_l$, $\tilde{\mathbf{t}}_i^{(l)}$. Hence, the variable $B_l(i)$ has two behaviors depending on whether the column $\tilde{\mathbf{a}}_i^{(l)}$ belongs to the dual code $\mathcal{C}^{\perp}$ or not. This variable will be studied as a function of $\tilde{\mathbf{a}}_i^{(l)}$ assuming that the bits that represent an element of the GF($q$), where $q = 2^m$, are uniformly distributed and independent from each other.

- If the column $\tilde{\mathbf{a}}_i^{(l)}$ does not belong to the dual code $\mathcal{C}^{\perp}$, the variable $B_l(i)$, for all $i \in [\![1, l]\!]$, will follow a binomial distribution of parameters $M$ and $1/q$ with a mean equal to $M/q$, denoted as $\mathcal{B}(M, 1/q)$.

- If the column $\tilde{\mathbf{a}}_i^{(l)}$ belongs to the dual code $\mathcal{C}^{\perp}$, the variable $B_l(i)$ will follow a binomial distribution with parameters $M$ and $P_i$, denoted as $\mathcal{B}(M, P_i)$. The parameter $P_i$ corresponds to the probability that a coefficient $\tilde{t}_i^{(l)}(j)$ of the column $\tilde{\mathbf{t}}_i^{(l)}$ is equal to 0 $\left(\text{i.e.,}\, P_i = Pr\left[\tilde{t}_i^{(l)}(j) = 0 \mid \tilde{\mathbf{a}}_i^{(l)} \in \mathcal{C}^{\perp}\right]\right)$.

It is possible to limit the two behaviors of the variable $B_l(i)$ by computing an optimal threshold $\hat{\eta}_{\text{opt}}$ such that:

$$\begin{cases} \text{If } B_l(i) > \hat{\eta}_{\text{opt}} \text{ then } \tilde{\mathbf{a}}_i^{(l)} \in \mathcal{C}^{\perp} \\ \text{If } B_l(i) \le \hat{\eta}_{\text{opt}} \text{ then } \tilde{\mathbf{a}}_i^{(l)} \notin \mathcal{C}^{\perp} \end{cases} \tag{7}$$

where $\hat{\eta}_{\text{opt}} = \frac{M}{q} \cdot \eta_{\text{opt}}$ is a real in the interval $[0, M]$. The optimal threshold $\eta_{\text{opt}}$ is able to minimize the probability of wrong detection of a column $\tilde{\mathbf{a}}_i^{(l)} \in \mathcal{C}^{\perp}$, denoted as $P_{\text{wd}}$, which corresponds to the sum of the false alarm probability, denoted as $P_{\text{fa}}$, and the probability of not detecting a theoretical dependent column, denoted as $P_{\text{nd}}$. The optimal threshold is determined by:

$$\begin{aligned} \eta_{\text{opt}} &= \arg\min_{\eta} (P_{\text{wd}}) = \arg\min_{\eta} (P_{\text{nd}} + P_{\text{fa}}) \\ &= \arg\min_{\eta} \left(1 + \sum_{j=\left\lfloor \frac{M}{q} \cdot \eta \right\rfloor + 1}^{M} \binom{M}{j} \cdot \right. \\ &\quad \left. \left[q^{-M} \cdot (q-1)^{M-j} - P_i^j \cdot (1-P_i)^{M-j}\right]\right) \end{aligned} \tag{8}$$

The normal distribution can be used to approximate the binomial probabilities of $B_l(i)$ when $M$ is large:

- If $\tilde{\mathbf{a}}_i^{(l)} \in \mathcal{C}^{\perp}$:

$$B_l(i) \to \mathcal{N}\left(\mu_0, \sigma_0^2\right) \tag{9}$$

- If $\tilde{\mathbf{a}}_i^{(l)} \notin \mathcal{C}^{\perp}$:

$$B_l(i) \to \mathcal{N}\left(\mu_1, \sigma_1^2\right) \tag{10}$$

where $\mathcal{N}\left(\mu_0, \sigma_0^2\right)$ is the normal distribution of parameters $\mu_0 = M \cdot P_i$ and $\sigma_0^2 = M \cdot P_i \cdot (1 - P_i)$ and $\mathcal{N}\left(\mu_1, \sigma_1^2\right)$ corresponds to the normal distribution of parameters $\mu_1 = M/q$ and $\sigma_1^2 = M \cdot (q-1)/q^2$.

Henceforth, the optimal value of the threshold $\hat{\eta}$ minimizing the probability of wrong detection $P_{\text{wd}}$ can be computed by:

$$\hat{\eta}_{\text{opt}} = \arg\min_{\hat{\eta}} \left(1 - \phi\left(\frac{\hat{\eta} - \mu_1}{\sigma_1}\right) + \phi\left(\frac{\hat{\eta} - \mu_0}{\sigma_0}\right)\right) \tag{11}$$

where $\phi(x)$ is the cumulative density function of the standard normal distribution:

$$\phi(x) = \frac{1}{\sqrt{2 \cdot \pi}} \cdot \int_{-\infty}^{x} e^{-\frac{t^2}{2}} \cdot dt \tag{12}$$

We can note that the optimal threshold $\hat{\eta}_{\text{opt}}$ depends on the parameters: $M$, $q$, and $P_i$. So, in order to delimit the two behaviors of the variable $B_l(i)$, it is necessary to compute the probability $P_i$.

### *Computation of the probability $P_i$*

In the case of binary codes, the probability $P_i$ has been calculated in [11]. But, it has never been studied in the general case of codes over GF($2^m$). In fact, the computation of the parameter $P_i$ is essential in order to detect the almost dependent columns in $\tilde{\mathbf{R}}_l$ by delimiting the two behaviors of the variable $B_l(i)$. Our aim is to investigate this probability in the case of non-binary codes. In the following, the theoretical study of $P_i$ is presented.

For $l = n$ and $i$ a position of a column $\tilde{\mathbf{a}}_i^{(l)}$ contained in $\mathcal{C}^{\perp}$, a coefficient $\tilde{t}_i^{(l)}(j)$ of the column $\tilde{\mathbf{t}}_i^{(l)}$ can be obtained, using (6), by:

$$\tilde{t}_i^{(l)}(j) = t_i^{(l)}(j) + \sum_{k=1}^{n} a_i^{(l)}(k) \cdot e_k^{(l)}(j) = \sum_{k=1}^{n} a_i^{(l)}(k) \cdot e_k^{(l)}(j) \tag{13}$$

where $t_i^{(l)}(j) = 0$ in the case of noiseless transmissions as explained previously. Indeed, the sum $\sum_{k=1}^{n} a_i^{(l)}(k) \cdot e_k^{(l)}(j)$ is null in this case because $e_k^{(l)}(j) = 0, \forall k \in \{1, \cdots, n\}$, and $\forall j \in \{1, \cdots, M\}$. However, in the case of noisy transmissions, the coefficients $e_i^{(l)}(j) \in \text{GF}(q)$ corresponds to the errors introduced by the noisy channel in the symbols $r_i^{(l)}(j) \in \text{GF}(q)$ in order to generate the noisy symbols $\tilde{r}_i^{(l)}(j) \in \text{GF}(q)$. Our aim is to determine $P_i$ the probability of detecting a zero coefficient in the column $\tilde{\mathbf{t}}_i^{(l)}$ corresponding to having $\sum_{k=1}^{n} a_i^{(l)}(k) \cdot e_k^{(l)}(j) = 0$:

$$P_i = Pr\left[\sum_{k=1}^{n} a_i^{(l)}(k) \cdot e_k^{(l)}(j) = 0\right] \tag{14}$$

Let $N_i(l)$ be the minimum number of linear combinations of columns required to obtain $\tilde{\mathbf{t}}_i^{(l)}$. This number corresponds also to the Hamming weight of the column $\tilde{\mathbf{a}}_i^{(l)}$. Then, there could be positions among $N_i(l)$ where $e_i^{(l)}(j) = 0$. Thus, $P_i$ can be defined as the probability of having $\sum_{k=1}^{s} a_i^{(l)}(k) \cdot e_k^{(l)}(j) = 0$ such that $s$ is the number of positions among $N_i(l)$ where $e_i^{(l)}(j) \neq 0$:

$$P_i = Pr\left[X = 0\right] + \sum_{s=1}^{N_i(l)} Pr\left[X = s, \sum_{k=1}^{s} a_i^{(l)}(k) \cdot e_k^{(l)}(j) = 0\right] \tag{15}$$

where $X$ is a random variable of the erroneous positions number among $N_i(l)$. Indeed, we show in Appendix that the probability $P_i$ of having $\tilde{t}_i^{(l)}(j) = 0$ can be determined by:

$$P_i = \frac{1 + (q-1) \cdot \left(1 - \frac{p_e \cdot q}{q-1}\right)^{N_i(l)}}{q} \tag{16}$$

In the case of GF(2) (i.e., $q = 2$), this probability can be written as:

$$P_i = \frac{1 + (1 - 2 \cdot p_e)^{N_i(l)}}{2} \tag{17}$$

This expression corresponds to that used in [11].

In Figure 4, we represent the wrong detection probability $P_{\text{wd}}$ as a function of $\hat{\eta}/M$ and $p_e$ assuming $q = 2^3$, $w\left(\tilde{\mathbf{a}}_i^{(l)}\right) = 20$ and $M = 2,000$. For each value of $p_e$, the optimal threshold $\hat{\eta}_{\text{opt}}$ corresponding to a root of (11) is computed. From Figure 4, we can deduce that the threshold interval satisfying $P_{\text{wd}} \approx 0$ decreases when the value of $p_e$ increases.

We can conclude that studying the behaviors of $B_l(i)$ in order to identify $n$ is based on the calculation of the optimal threshold $\hat{\eta}_{\text{opt}}$. However, this threshold depends on the value of the error probability $p_e$ which is unknown for the receiver. So, the need to estimate this parameter is a blocking step in the almost dependent columns method and also leads to a lack of robustness.

In order to address these problems, we propose a new iterative method based on the arithmetic mean of the variable $B_l(i)$ which do not depend on $p_e$ and where the iterative process permits to improve the detection probability.

## New iterative method based on the arithmetic mean of the variable $B_l(i)$

In this part, the proposed method based on the arithmetic mean of the number of zeros in the columns of the matrix $\tilde{\mathbf{T}}_l$ is described. We recall that the Gauss elimination described in Algorithm 1 should be applied in order to obtain $\tilde{\mathbf{T}}_l$. We show here that the identification of the parameter $n$ by our proposed method does not depend on the error probability $p_e$. In this method, in order to improve the detection probability of $n$, an iteration process is introduced. We consider the idea of the iterative process proposed in [18,19]. The principle of this process is to perform random permutations on the rows of the matrix $\tilde{\mathbf{R}}_l$ in order to obtain a new virtual realization of the received data. These permutations permit to increase the probability to obtain non-erroneous pivots during the Gauss elimination.

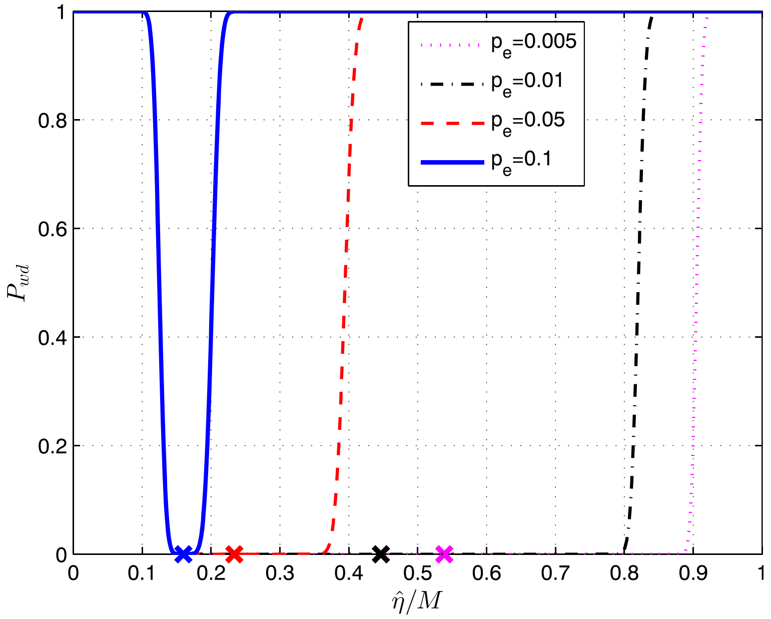

**Figure 4 Probability $P_{\mathrm{wd}}$ according to $\hat{\eta}/M$ and $p_e$.** The probability of wrong detection of almost dependent columns $P_{\mathrm{wd}}$ according to $\hat{\eta}/M$ and $p_e$ is represented for $q = 2^3$, $M = 2,000$ and $w\left(\tilde{\mathbf{a}}_i^{(l)}\right) = 20$.

The arithmetic mean of the variables $B_l(i)$, $\forall i \in [\![ 1, l ]\!]$, denoted $E_l$ is defined by:

$$E_l = \frac{\sum_{i=1}^{l} B_l(i)}{l} \tag{18}$$

**Property 1.** *If $X_1, X_2, \cdots, X_m$ are independent random variables respectively following:*

$$\mathcal{N}\left(\mu_1, \sigma_1^2\right), \mathcal{N}\left(\mu_2, \sigma_2^2\right), \ldots, \mathcal{N}\left(\mu_m, \sigma_m^2\right)$$

*the mean defined by $\frac{(X_1+X_2+\cdots+X_m)}{m}$ follows:*

$$\mathcal{N}\left(\frac{\mu_1+\mu_2+\cdots+\mu_m}{m}, \frac{\sigma_1^2+\sigma_2^2+\cdots+\sigma_m^2}{m^2}\right) \tag{19}$$

We recall that the variable $B_l(i)$ which is the number of zeros in the $i$-th column of the matrix $\tilde{\mathbf{T}}_l$ has two possible behaviors depending on $l$:

- If $l \neq \alpha \cdot n$, for $\alpha \in \mathbb{N}$, the variable $B_l(i)$ follows a normal distribution $\mathcal{N}\left(\mu_1, \sigma_1^2\right)$ for all columns $i$ of $\tilde{\mathbf{T}}_l$. In this case, using the property 1, the mean $E_l$ will follow:

$$E_l \to \mathcal{N}\left(\mu_1, \frac{\sigma_1^2}{l}\right) \tag{20}$$

We can note that the mean $E_l$ will be close to $M/q$.

- If $l = \alpha \cdot n$, for $\alpha \in \mathbb{N}$:
  - If the $i$-th column is an almost dependent column, the variable $B_l(i)$ will follow the normal distribution of parameters $\mathcal{N}\left(\mu_0, \sigma_0^2\right)$.
  - If the $i$-th column is not an almost dependent column, the variable $B_l(i)$ will follow the normal distribution of parameters $\mathcal{N}\left(\mu_1, \sigma_1^2\right)$.

Thereby, the mean $E_l$ is given by:

$$E_l \to \mathcal{N}\left(\frac{Q(l)\cdot\mu_0 + k_l\cdot\mu_1}{l}, \frac{Q(l)\cdot\sigma_0^2 + k_l\cdot\sigma_1^2}{l^2}\right) \tag{21}$$

where $Q(l)$ is the number of almost dependent columns in the matrix $\tilde{\mathbf{R}}_l$ such that:

$$Q(l) = \mathrm{Card}\left\{i \in [\![ 0, l ]\!], B_l(i) > \hat{\eta}_{\mathrm{opt}}\right\} \tag{22}$$

where Card$(x)$ is the cardinal function which returns the set size. $k_l = l - Q(l)$ is the number of independent columns in the same matrix. In the noiseless environment, the mean $E_l$ is stable at:

$$E_l = \frac{M\cdot(q\cdot(n-k)+k)}{q\cdot n} \tag{23}$$

We note two behaviors of $E_l$ with respect to $l = \alpha \cdot n$ or $l \neq \alpha \cdot n$:

$$\begin{cases} \text{If } l \neq \alpha \cdot n \text{ then } E_l \leq \dfrac{M}{q} \\ \text{If } l = \alpha \cdot n \text{ then } E_l > \dfrac{M}{q} \end{cases} \tag{24}$$

The gap between these behaviors allows us to find the matrices which have the number of columns $l = \alpha \cdot n$.

Let $\mathcal{J}$ be a set of $l$-values where the gap $E_l - \frac{M}{q} > 0$:

$$\mathcal{J} = \left\{ l = 1, \cdots, l_{\max} | E_l - \frac{M}{q} > 0 \right\} \tag{25}$$

Thereby, the identified length of the code words will be such that:

$$\tilde{n} = \text{mode}(\text{diff}(\mathcal{J})) \tag{26}$$

where the functions diff($\mathbf{x}$) and mode($\mathbf{x}$) are defined by:

- Function diff($\mathbf{x}$): the output of this function is a vector of size $s - 1$ and it corresponds to the difference between two consecutive elements of the vector $\mathbf{x} = \big(x(1)\ x(2)\ \cdots\ x(s)\big)$:

$$\text{diff}(\mathbf{x}) = \big(x(2) - x(1)\ \cdots\ x(s) - x(s-1)\big) \tag{27}$$

- Function mode($\mathbf{x}$): this operation provides the value which has the highest occurrence in the vector $\mathbf{x}$.

The proposed iterative method of the code word length identification is summarized in the Algorithm 2.

**Algorithm 2** The algorithm based on the arithmetic mean calculation

```
Require: $\tilde{\mathbf{r}}$, $M$, $q$ and maximum number of iterations $\text{it}_{\max}$
Ensure: Identified code word length $\tilde{n}$
  Initialize the number of iterations $\text{it} = 1$
  Initialize the stop criterion $\text{end}_{\text{it}} = 0$
  while $\text{end}_{\text{it}} = 0$ do
    for $l = 1$ to $l_{\max}$ do
      Build matrix $\tilde{\mathbf{R}}_l$ of size $(M \times l)$
      $\tilde{\mathbf{R}}_l \rightarrow \tilde{\mathbf{T}}_l = \tilde{\mathbf{R}}_l \cdot \tilde{\mathbf{A}}_l$
      for $i = 1$ to $l$ do
        Count $B_l(i)$
      end for
      Compute $E_l$
    end for
    Create the set $\mathcal{J}$ (25)
    if $\mathcal{J}$ is empty then
      if $\text{it} < \text{it}_{\max}$ then
        $\text{it} = \text{it} + 1$
        Permute randomly the rows of $\tilde{\mathbf{R}}_l$
      else
        $\text{end}_{\text{it}} = 1$
      end if
    else
      Determine $\tilde{n}$ (26)
      $\text{end}_{\text{it}} = 1$
    end if
  end while
```

**Example 1.** Let us consider the Reed-Solomon code, denoted $RS(15, 11)$, over GF($2^4$) which is defined by: $n = 15$ and $k = 11$. The mean $E_l$ normalized by $M$, which is set to 1,000, is represented in Figures 5 and 6. In Figure 5, a zero probability of error (i.e., $p_e = 0$) is considered. For

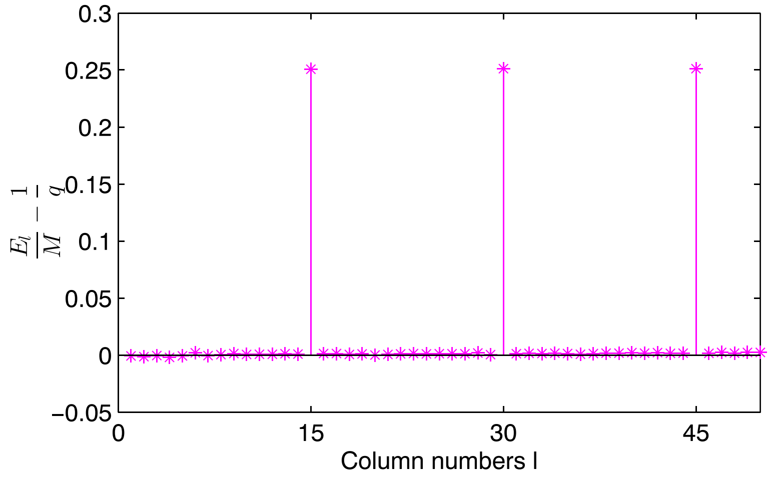

**Figure 5 Gap between the mean $E_l/M$ and $1/q$ of $RS(15, 11)$ over GF($2^4$) for $p_e = 0$.** The mean $E_l$ normalized by $M = 1{,}000$ is represented in the case of $p_e = 0$.

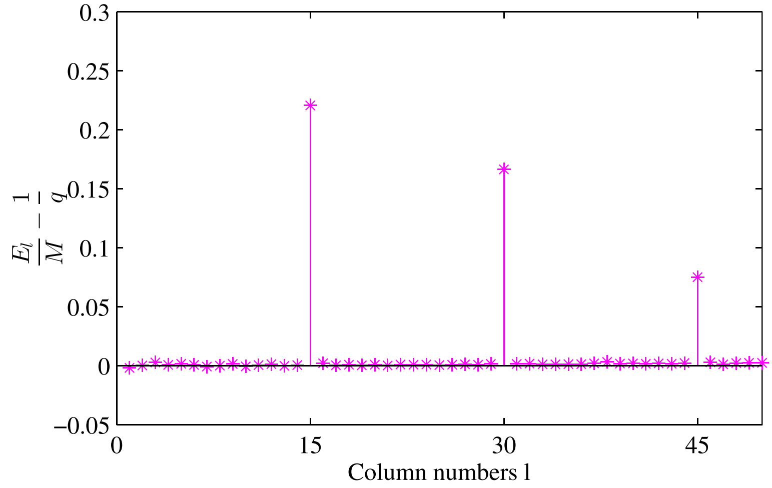

**Figure 6 Gap between the mean $E_l/M$ and $1/q$ of $RS(15, 11)$ over GF($2^4$) for $p_e = 0.01$.** The mean $E_l$ normalized by $M = 1,000$ is represented in the case of $p_e = 0.01$.

$l \neq \alpha \cdot n$, we can verify that the mean $E_l$ normalized by $M$ is stable at $1/q = 0.0625$. For $l = \alpha \cdot n$, the mean $E_l$ meets (23):

$$\frac{1}{q \cdot n} \cdot (q \cdot (n-k) + k) = 0.3125$$

So, the matrices of size $l = \alpha \cdot n$ have peaks for $\frac{E_l}{M} - \frac{1}{q} = 0.25 > 0$. In Figure 6, the gap $\frac{E_l}{M} - \frac{1}{q}$ is represented with respect to $l$ when $p_e = 0.01$ for one iteration of our algorithm. According to (25), the set $\mathcal{J}$ is shown in Table 1. Henceforth, using (26), the identified length of the code words is $\tilde{n} = 15$.

## Analysis and performances

The aim of our proposed algorithm is to blindly identify the length of non-binary code words in noisy environment. This purpose can be reached with an average complexity equal to $\mathcal{O}(M \cdot \mathrm{l}_{\max}^3 \cdot \mathrm{it}_{\max})$. Indeed, the proposed algorithm performs $((\mathrm{l}_{\max} - 1) \cdot \mathrm{it}_{\max})$ processes of Gaussian eliminations which have an average complexity equal to $\mathcal{O}(M \cdot l^2)$, where $l = 2 \cdots l_{\max}$. So, the average complexity is such that:

$$\mathcal{O}\left(M \cdot \sum_{\mathrm{it}=1}^{\mathrm{it}_{\max}} \sum_{l=2}^{\mathrm{l}_{\max}} l^2\right) = \mathcal{O}(M \cdot l_{\max}^3 \cdot \mathrm{it}_{\max}) \qquad (28)$$

**Table 1 Sizes of the matrices $\tilde{\mathrm{R}}_l$ for $\frac{E_l}{M} - \frac{1}{q} > 0$**

| $l \in \mathcal{J}$ | 15 | 30 | 45 |
|---|---|---|---|
| diff($\mathcal{J}$) | | 15 | 15 |

The sizes of matrices for which $\frac{E_l}{M} - \frac{1}{q} > 0$, $\mathcal{J}$, and the set diff($\mathcal{J}$) are given for $p_e = 0.01$ in the case of $RS(15, 11)$ over GF($2^4$).

In order to analyze the performances of our blind identification method, the probability of correct detection of the code word length $n$ is chosen as a performance criterion. In the simulations, our method is applied to the non-binary LDPC codes which became candidate for future communication systems. For each simulation, 2,000 Monte Carlo trials are run where the data symbols are randomly chosen at each trial. In this part, we focus on:

- the gain of the iteration process on the detection probability of $n$
- the performance comparison in the case of different channels
- the impact of increasing the Galois field dimension $q$ on the detection probabilities of $n$
- the impact of increasing the code word length $n$ on the detection probabilities for a given $q$

### *Gain of the iterative process*

In our simulations, we consider a LDPC ($n = 6$, $k = 3$) over GF(4). Figure 7 shows the probability of detecting $n$ according to $p_e$ for one, three, five, and ten iterations. We can see that the gain between the first and the tenth iteration is significantly important. Indeed, for $p_e = 0.07$, with one iteration, the detection probability is equal to 0.76 and it becomes equal to 0.99 after 10 iterations. We can deduce that the iterative process improves significantly the detection performances of the blind identification method based on the mean calculation.

### *Performance comparison in the case of different channels*

Let us illustrate the detection obtained by the proposed method for a LDPC ($n = 16, k = 8$) over GF(8) when an AWGN channel (the first channel) and a multipath Rayleigh channel associated to an AWGN channel (the

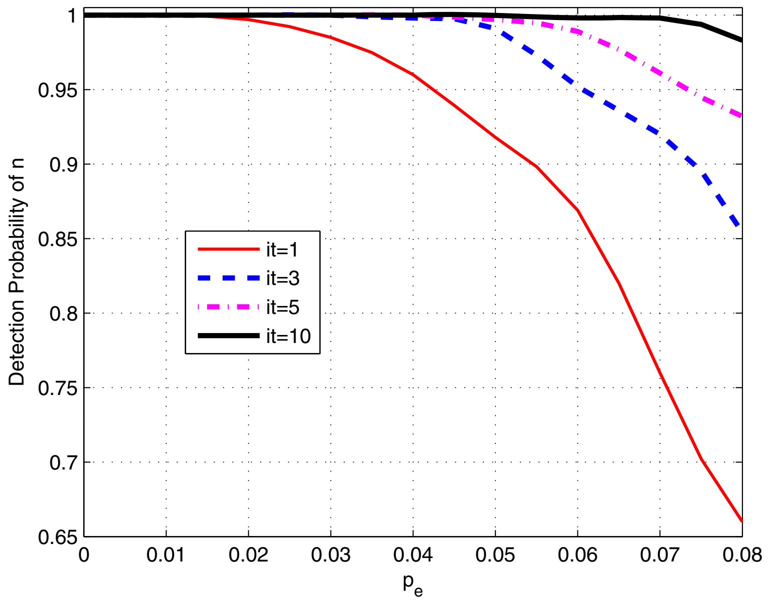

**Figure 7 The detection probability of the method based on the mean calculation for LDPC ($n = 6, k = 3$) in GF(4).** For LDPC ($n = 6, k = 3$) in GF(4), the probability of detecting $n$ is depicted compared with the error probability $p_e$ for one, three, five, and ten iterations.

second channel) are considered. In order to compensate and reduce the inter-symbol interference (ISI) caused by the multipath propagation, a linear mean square error (MSE) equalizer of length 20 was used.

We evaluate the performances of our method when the QAM or PAM modulation of order 8 (8-QAM and 8-PAM) is used to transmit the symbols coded by LDPC ($n = 16, k = 8$) over GF(8). In Figures 8 and 9, a comparison of performances of our blind identification method using 8-PAM or 8-QAM modulations in the case of an AWGN channel and a multipath channel with path number $L_{\text{path}} = 4$ and $\text{it}_{\max} = 1$ is presented. In Figure 8, a comparison of the detection performances of our method in the case of AWGN channel is depicted. We can see that the proposed method for 8-QAM modulation gives better performances than for 8-PAM modulation when SNR<

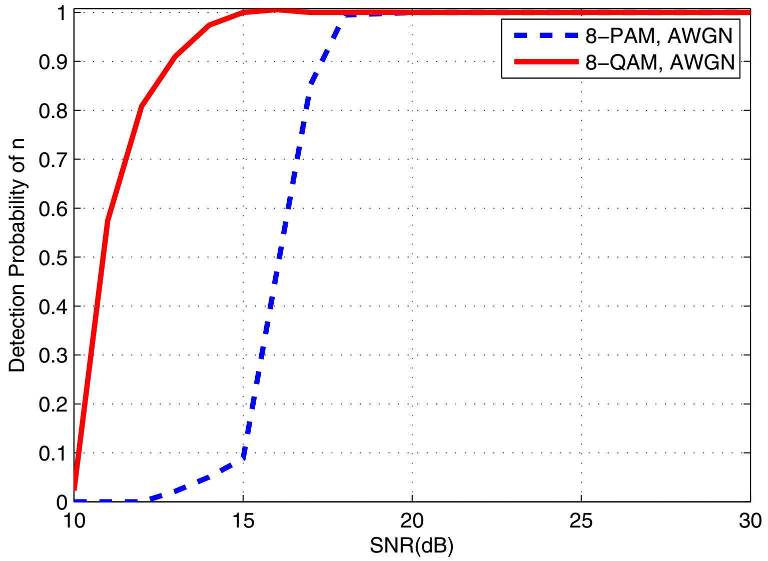

**Figure 8 The detection probability of the method based on the mean calculation for LDPC ($n = 16, k = 8$) in the case of AWGN channel.** For LDPC ($n = 16, k = 8$) over GF(8), the detection probabilities of the method based on the mean calculation in the case of AWGN channel are depicted when a 8-PAM and 8-QAM modulations are used.

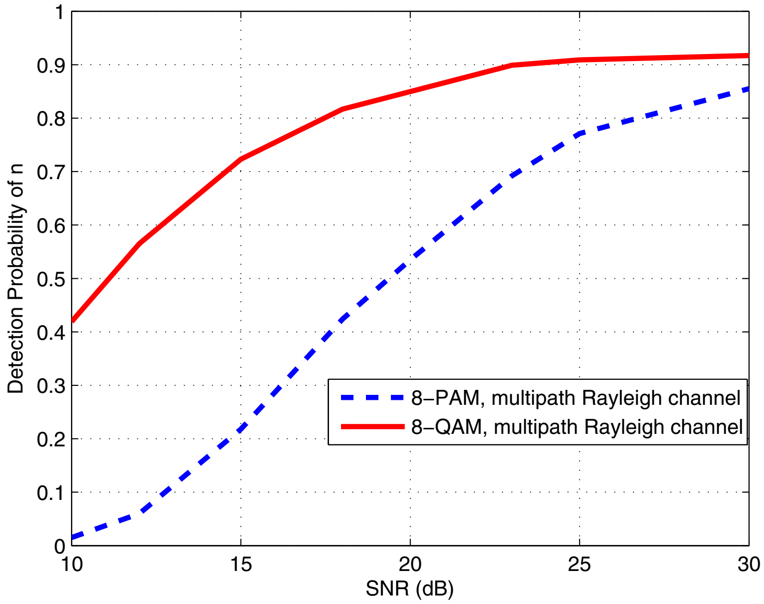

**Figure 9 The detection probability of the method based on the mean calculation in the case of multipath Rayleigh channel.** For LDPC ($n = 16, k = 8$) over GF(8), the detection probabilities of our method are depicted in the case of multipath Rayleigh channel with $L_{\text{path}} = 4$ when 8-PAM and 8-QAM modulations are used.

18 dB. The gain between both is equal to 5 dB. However, for SNR > 18 dB, the performances are similar and the detection probability is equal to 1. To obtain the detection probabilities presented in Figure 9, the modulated symbols by 8-PAM or 8-QAM modulations are transmitted in a quasi-static Rayleigh fading multipath channel with path number $L_{\text{path}} = 4$, then the received symbols are treated by the linear MSE equalizer of length 20. We can observe that, in the case of 8-QAM, our proposed method provides better performances than for 8-PAM. A gain equal to 5 dB is exhibited. We have chosen to evaluate our proposed methods in the worst case of 8-PAM modulation because our aim was to show that our method has the best performances even in the case of the PAM modulation.

In the following, the performance study of the impact of $n$ and $q$ on the proposed method is presented.

### *Impact of increasing q*

Let us consider a LDPC ($n = 6$, $k = 3$), constructed in the Galois field GF($q$), where $q = 4, 8, 16$. The matrices $\tilde{\mathbf{R}}_l$ are reshaped from $L = 30,000$ received symbols with $l = 2, \cdots, 30$ and $M = 1,000$. For each value of $q$, the method based on the mean calculation is applied to blindly identify the code word length of LDPC ($n = 6$, $k = 3$) over GF($q$) when $\text{it}_{\max} = 1$. Figure 10 depicts the probability of detecting the correct $n$ by our blind identification method according to the error probability $p_e$ in the cases of GF(4), GF(8), and GF(16). This figure shows that the curve behavior is nearly similar for all $q = 4$, 8, 16. We can deduce that the method based on the mean calculation is slightly sensitive to the increase of the Galois field dimension $q$.

### *Impact of increasing n*

To evaluate the detection performances of our blind identification method, the impact of increasing the code word length should be studied. In our simulations, we consider two LDPC codes over GF(8), a LDPC ($n = 6$, $k = 3$) and a LDPC ($n = 16$, $k = 8$). The matrices $\tilde{\mathbf{R}}_l$ are reshaped from $L = 64,000$ received symbols with $l = 2, \cdots, 64$ and $M = 1,000$. For each code, the method based on the mean calculation is applied to blindly identify the code word length $n$ when $\text{it}_{\max} = 1$. Figure 11 shows the detection probabilities of $n$ by the method based on the mean calculation. We can note that the increase of the code word length leads to lower detection performances with our proposed method. Indeed, for $p_e = 0.01$, the detection probability of the method of the mean calculation is constant and equal to 1 in the case of the two codes. For $p_e = 0.02$, the detection probability decreases from 0.99 to 0.94.

In order to show that our method works in the case of codes of a reasonable code word length, we computed the detection probability of the Reed-Solomon code RS ($n = 31, k = 25$) over GF(32) which corresponds to an equivalent code over GF(2) of length $m \cdot n = 5 \cdot 31 = 155$. For an error probability $p_e = 0.01$ and 1,000 trials

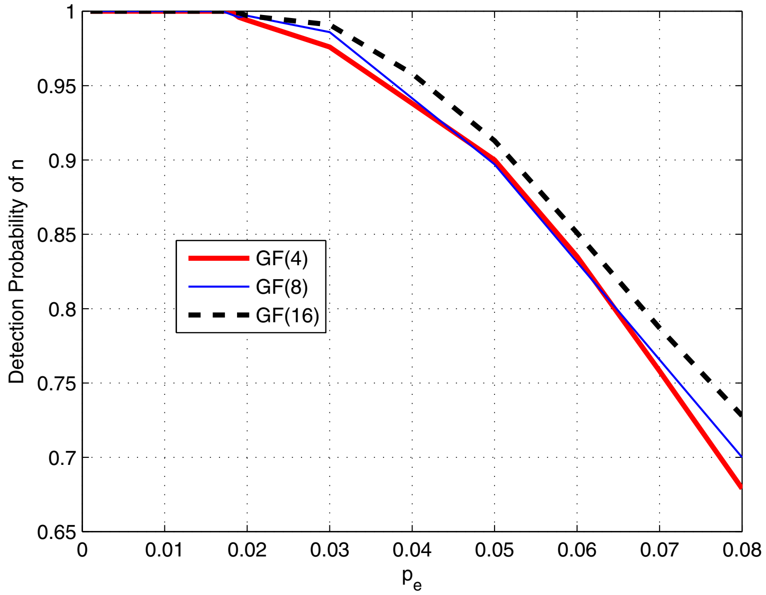

**Figure 10 Impact of the Galois field dimension *q* on the detection probability of *n* by the proposed method considering $it_{max} = 1$.** For LDPC ($n = 6, k = 3$), the probability of detecting the correct *n* by the method based on the arithmetic mean computation is depicted according to the error probability $p_e$ in the cases of GF(4), GF(8), and GF(16).

of Monte Carlo, we obtained a detection probability of 0.87 for $it_{max} = 50$. This probability can be improved by increasing the number of iteration of our algorithm. For $it_{max} = 100$, we obtained a detection probability of 0.95.

## Conclusions

In this paper, we have introduced an algorithm devoted to the blind identification of the code word length for a non-binary code in a noisy transmission environment. Using this algorithm, the code word length can be identified by

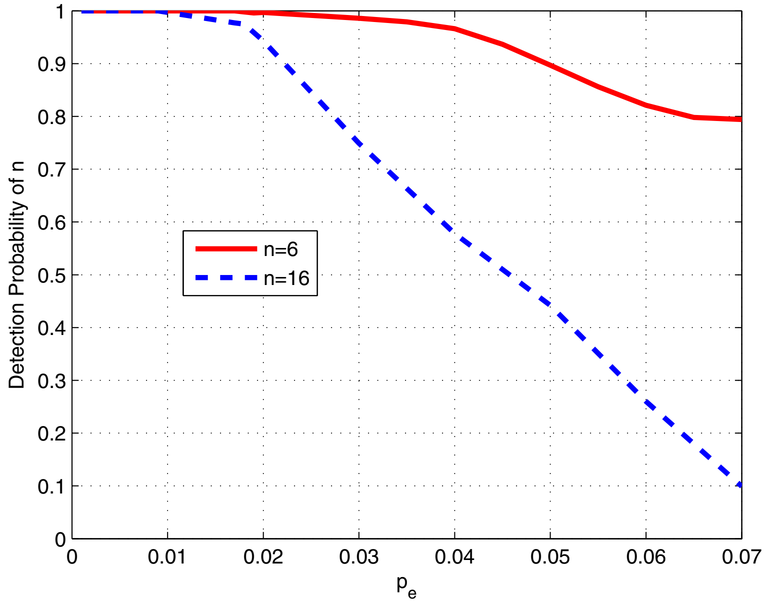

**Figure 11 Impact of increasing *n* on the detection probability for LDPC codes.** Of sizes $n = 6$ and $n = 16$ by using the proposed method considering $it_{max} = 1$. For LDPC ($n = 6, k = 3$) and LDPC ($n = 16, k = 8$) over GF(8), the probability of detecting the correct *n* by the method based on the arithmetic mean calculation is depicted according to the error probability $p_e$.

calculating the arithmetic mean of the number of zeros that occur in the columns of the matrix obtained by the Gauss elimination. We have shown that the proposed algorithm is robust because it does not require the estimation of error probability, is insensitive to the high order of Galois field, and has the best detection performances for the most of modulation types. Furthermore, this method provides better performances of detection when an iterative process is considered in order to increase the probability to obtain non-erroneous pivots during the Gauss elimination.

Our future work will focus on identifying the remainder of the non-binary code parameters as well as a parity check matrix, permitting to implement a generic decoder in a noisy environment. Furthermore, a method based on using soft information that allows us to improve the performances of the blind identification algorithms will be published soon [31].

## Appendix

### Proof of Equation 8

We define $\mathcal{H}_0$ and $\mathcal{H}_1$ by:

$$\mathcal{H}_0 \text{ if } \tilde{\mathbf{a}}_i^{(l)} \in \mathcal{C}^\perp \text{ and } \mathcal{H}_1 \text{ if } \tilde{\mathbf{a}}_i^{(l)} \notin \mathcal{C}^\perp \tag{29}$$

The two behaviors of $B_l(i)$ are limited in (7). The aim of this appendix is to demonstrate (8). In order to determine the probabilities of $P_{\text{fa}}$ and $P_{\text{nd}}$, we should study the behaviors of the variable $B_l(i)$ according to the hypotheses $\mathcal{H}_0$ and $\mathcal{H}_1$:

**Under the hypothesis $\mathcal{H}_0$:** the variable $B_l(i)$ follows a binomial distribution $\mathcal{B}(M, P_i)$. So, the probability that $B_l(i)$ is greater than $\frac{M}{q} \cdot \eta$ is as follows:

$$Pr\left[B_l(i) > \frac{M}{q} \cdot \eta \mid \mathcal{H}_0\right] = \sum_{j=\lfloor \frac{M}{q} \cdot \eta \rfloor + 1}^{M} \binom{M}{j} \cdot P_i^j \cdot (1-P_i)^{M-j} \tag{30}$$

**Under the hypothesis $\mathcal{H}_1$:** the variable $B_l(i)$ follows a binomial distribution $\mathcal{B}(M, 1/q)$. So, the probability that $B_l(i)$ is less than or equal to $\frac{M}{q} \cdot \eta$ is as follows:

$$Pr\left[B_l(i) \leq \frac{M}{q} \cdot \eta \mid \mathcal{H}_1\right] = \sum_{j=0}^{\lfloor \frac{M}{q} \cdot \eta \rfloor} \binom{M}{j} \cdot \frac{(q-1)^{M-j}}{q^M} \tag{31}$$

Using these two probabilities, we will calculate the false alarm probability $P_{\text{fa}}$, the probability of not detecting a theoretical dependent column $P_{\text{nd}}$ and the probability of detection $P_{\text{det}}$.

**Calculation of the false alarm probability $P_{\text{fa}}$:** this probability corresponds to decide that a column $\tilde{\mathbf{a}}_i^{(l)}$ belongs to a dual code $\mathcal{C}^\perp$ even thought in reality it does not belong. This probability can be determined by:

$$\begin{aligned} P_{\text{fa}} &= Pr\left[B_l(i) > \frac{M}{q} \cdot \eta \mid \mathcal{H}_1\right] \\ &= \sum_{j=\lfloor \frac{M}{q} \cdot \eta \rfloor + 1}^{M} \binom{M}{j} \cdot \frac{(q-1)^{M-j}}{q^M} \end{aligned} \tag{32}$$

**Calculation of the probability of not detecting a theoretical dependent column $P_{\text{nd}}$:** this probability corresponds to decide that a column $\tilde{\mathbf{a}}_i^{(l)}$ does not belong to $\mathcal{C}^\perp$ even thought in reality it belongs. This probability can be determined by:

$$\begin{aligned} P_{\text{nd}} &= Pr\left[B_l(i) \leq \frac{M}{q} \cdot \eta \mid \mathcal{H}_0\right] \\ &= \sum_{j=0}^{\lfloor \frac{M}{q} \cdot \eta \rfloor} \binom{M}{j} \cdot P_i^j \cdot (1-P_i)^{M-j} \end{aligned} \tag{33}$$

**Calculation of the probability of detection $P_{\text{det}}$:** this probability is defined by:

$$\begin{aligned} P_{\text{det}} = 1 - P_{\text{nd}} &= Pr\left[B_l(i) > \frac{M}{q} \cdot \eta \mid \mathcal{H}_0\right] \\ &= \sum_{j=\lfloor \frac{M}{q} \cdot \eta \rfloor + 1}^{M} \binom{M}{j} \cdot P_i^j \cdot (1-P_i)^{M-j} \end{aligned} \tag{34}$$

Using (32) and (34), the optimal threshold can be determined by:

$$\begin{aligned} \eta_{\text{opt}} &= \arg\min_{\eta} (P_{\text{wd}}) = \arg\min_{\eta} (P_{\text{nd}} + P_{\text{fa}}) \\ &= \arg\min_{\eta} (1 + P_{\text{fa}} - P_{\text{det}}) \\ &= \arg\min_{\eta} \left(1 + \sum_{j=\left\lfloor \frac{M}{q} \cdot \eta \right\rfloor + 1}^{M} \binom{M}{j} \right. \\ &\qquad \left. \times \left[q^{-M} \cdot (q-1)^{M-j} - P_i^j \cdot (1-P_i)^{M-j}\right]\right) \end{aligned} \tag{35}$$

### Proof of the equation (16)

The probability $P_i$ is initially expressed by (15). We denote $P_1(s) = Pr[X = s]$ and $P_2(s) = Pr$

$\left[\sum_{k=1}^{s} a_i^{(l)}(k) \cdot e_k^{(l)}(j) = 0\right]$ such that these two probabilities are independent. Henceforth, (15) becomes:

$$P_i = P_1(0) + \sum_{s=1}^{N_i(l)} P_1(s) \cdot P_2(s) \qquad (36)$$

Assuming that the errors are independent from each other and uniformly distributed in $\mathrm{GF}(q)\backslash\{0\}$, the variable $X$ follows a binomial distribution with parameters $N_i(l)$ and $p_e$. Thereby, the probability $P_1(s)$ is determined by:

$$P_1(s) = \binom{N_i(l)}{s} \cdot (p_e)^s \cdot (1-p_e)^{N_i(l)-s} \qquad (37)$$

The probability $P_2(s)$ is the probability of having $\sum_{k=1}^{s} a_i^{(l)}(k) \cdot e_k^{(l)}(j) = 0$ where $e_k^{(l)}(j) \in \mathrm{GF}(q)\backslash\{0\}$.

We demonstrate by the mathematical induction that the probability $P_2(s)$ can be expressed by:

$$P_2(s) = \frac{1 - P_2(s-1)}{q-1} \qquad (38)$$

We have $P_2(0) = 1$ because there are no erroneous positions. In the case of a single erroneous position, we have $P_2(1) = 0$. However, considering the example of $\mathrm{GF}(2^2)$, the probability $P_2(s = 2)$ can be obtained by the matrix $\mathbf{M}$ whose the indexes of rows and columns correspond to non-zero elements of this field. The coefficients of this matrix correspond to the sum over $\mathrm{GF}(2^2)$ of the indexes of a row and a column.

$$\mathbf{M} = \begin{pmatrix} 0 & 3 & 2 \\ 3 & 0 & 1 \\ 2 & 1 & 0 \end{pmatrix} \qquad (39)$$

If we have $\left(a_i^{(l)}(1), e_1^{(l)}(j)\right) \in \left(GF(2^2)^*\right)^2$, and $\left(a_i^{(l)}(2), e_2^{(l)}(j)\right) \in \left(GF(2^2)^*\right)^2$, the probability of having $a_i^{(l)}(1) \cdot e_1^{(l)}(j) + a_i^{(l)}(2) \cdot e_2^{(l)}(j) = 0$ will be $P_2(2) = 3/9 = 1/3$. The computed probability verifies (38).

We assume that (38) is verified for $s$, and we demonstrate it for $s+1$. If we have $\sum_{k=1}^{s+1} a_i^{(l)}(k) \cdot e_k^{(l)}(j) = 0$, we will have $e_{s+1}^{(l)} = -\frac{1}{a_i^{(l)}(s+1)} \cdot \sum_{k=1}^{s} a_i^{(l)}(k) \cdot e_k^{(l)}(j)$ that belongs to $\mathrm{GF}(q)^*$ with a probability equal to $1/(q-1)$. Therefore, the probability $P_2(s+1)$ is determined by:

$$\begin{aligned} P_2(s+1) &= Pr\left(e_{s+1}^{(l)} \in \mathrm{GF}(q)^*, \sum_{k=1}^{s} a_i^{(l)}(k) \cdot e_k^{(l)}(j) \neq 0\right) \\ &= Pr\left(e_{s+1}^{(l)} \in \mathrm{GF}(q)^*\right) \cdot Pr\left(\sum_{k=1}^{s} a_i^{(l)}(k) \cdot e_k^{(l)}(j) \neq 0\right) \\ &= \frac{\overline{P_2(s)}}{q-1} \\ &= \frac{1 - P_2(s)}{q-1} \end{aligned} \qquad (40)$$

In order to simplify the expression of $P_2(s)$, a change of variable is done by considering $\varphi(s) = (q-1)^{s-1} \cdot P_2(s)$. When $P_2(s)$ is replaced by $\varphi(s)$, the expression (38) becomes:

$$\varphi(s) + \varphi(s-1) = (q-1)^{s-2} \qquad (41)$$

Denoting $\rho(s) = (-1)^s \cdot \varphi(s)$, the expression (41) can be written as:

$$\rho(s) = \rho(1) + \sum_{i=0}^{s-1} (1-q)^i \qquad (42)$$

but, the sum $\sum_{i=0}^{s-1}(1-q)^i$ is a geometric sequence of common ratio $1-q$. So, it can be written as:

$$\sum_{i=0}^{s-1} (1-q)^i = \frac{1 - (-1)^{s-1} \cdot (q-1)^{s-1}}{q} \qquad (43)$$

The computation of $\rho(1)$ gives $\rho(1) = 0$. Therefore, using (43) and (41), the simplified expression of $P_2(s)$ is written as:

$$P_2(s) = \frac{(-1)^s + (q-1)^{s-1}}{q \cdot (q-1)^{s-1}} \qquad (44)$$

Using (37) and (44), the overall probability $P_i$ is given by:

$$\begin{aligned} P_i = \tfrac{1}{q} \cdot \Bigg( & \textstyle\sum_{j=0}^{N_i(l)} \binom{N_i(l)}{j} \cdot p_e^j \cdot (1-p_e)^{N_i(l)-j} + (q-1) \\ & \times \textstyle\sum_{j=0}^{N_i(l)} \binom{N_i(l)}{j} \cdot \left(\frac{-p_e}{q-1}\right)^j \cdot (1-p_e)^{N_i(l)-j} \Bigg) \end{aligned} \qquad (45)$$

In order to simplify this equation, the Newton's binomial formula can be applied:

$$(Z+Y)^{N_i(l)} = \sum_{j=0}^{N_i(l)} \binom{N_i(l)}{j} \cdot Z^j \cdot Y^{N_i(l)-j}$$

Thus, the probability of having an element of the $i$-th column of $\tilde{\mathbf{T}}_l$ equal to 0 is determined by:

$$P_i = \frac{1 + (q-1) \cdot \left(1 - \frac{p_e \cdot q}{q-1}\right)^{N_i(l)}}{q} \qquad (46)$$

**Competing interests**

The authors declare that they have no competing interests.

**Author details**

[1]Université Européenne de Bretagne, 5 Boulevard Laënnec, 35000 Rennes, France. [2]Université de Brest; CNRS, UMR 6285 Lab-STICC, 6 avenue Victor Le Gorgeu, 29238 Brest, France. [3]Université de Brest; CNRS UMR 6205, Laboratoire de Mathématiques Bretagne Atlantique, 6 avenue Victor Le Gorgeu, 29238 Brest, France.

Received: 12 February 2014 Accepted: 11 February 2015
Published online: 01 March 2015

**References**

1. MC Davey, D MacKay, Low-density parity-check codes over GF($q$). IEEE Commun. Lett. **2**, 165–167 (1998)
2. JA Briffa, HG Schaathun, in *5th International Symposium on Turbo Codes and Related Topics*. Non-binary turbo codes and applications (IEEE Lausanne, 2008)
3. D Declercq, M Fossorier, Decoding algorithms for nonbinary LDPC codes over GF($q$). IEEE Trans. Commun. **55**(4), 633–643 (2007)
4. L Barnault, D Declercq, in *Proceedings ITW*. Fast decoding algorithm for LDPC over GF($2^q$) (IEEE, Paris, France, 2003), pp. 70–73
5. A Voicila, D Declercq, F Verdier, M Fossorier, P Urard, Low-complexity decoding for non-binary LDPC codes in high order fields. IEEE Trans. Commun. **58**(5), 1365–1375 (2010)
6. Yang Yu, W Chen, Design of low complexity non-binary LDPC codes with an approximated performance-complexity tradeoff. IEEE Commun. Lett. **16**(4), 514–517 (2012)
7. T Xia, HC Wu, Identification of nonbinary LDPC codes using average LLR of syndrome a posteriori probability. IEEE Commun. Lett. **17**(7), 1301–1304 (2013)
8. J Stern, A method for finding code words of small weight. Coding Theory Appl. **388**, 106–113 (1989)
9. A Canteaut, F Chabaud, A new algorithm for finding minimum-weight words in a linear code: application to McElieces cryptosystem and to narrow-sense BCH codes of length 511. IEEE Trans. Inf. Theory. **44**, 367–378 (1998)
10. A Valembois, Detection and recognition of a binary linear code. Discrete Appl. Math. **111**(1-2), 199–218 (2001)
11. M Cluzeau, in *2006 IEEE International Symposium on Information Theory*. Block code reconstruction using iterative decoding techniques (IEEE, Seattle, WA, 2006), pp. 2269–2273
12. M Cluzeau, M Finiasz, in *IEEE International Symposium on Information Theory 2009*. Recovering a code's length and synchronization from a noisy intercepted bitstream (IEEE, Seoul, 2009), pp. 2737–2741
13. M Côte, N Sendrier, in *IEEE International Symposium on Information Theory (ISIT)*. Reconstruction of convolutional codes from noisy observation (IEEE, Seoul, 2009), pp. 546–550
14. G Burel, R Gautier, in *IASTED International Conference on Communications, Internet and Information Technology*. Blind estimation of encoder and interleaver characteristics in a non cooperative context (ACTA Press, Scottsdale, AZ, USA, 2003)
15. Y Zrelli, R Gautier, M Marazin, E Rannou, E Radoi, Focus on theoretical properties of blind convolutional codes identification methods based on rank criterion. MTA Review. **XXII**(4), 213–234 (2012)
16. Y Zrelli, M Marazin, R Gautier, E Rannou, in *Proceedings of the International Conference on Computer Communication Networks*. Blind identification of convolutional encoder parameters over GF($2^m$) in the noiseless case (IEEE, Maui, Hawaii, 2011)
17. G Sicot, S Houcke, in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 3. Blind detection of interleaver parameters (IEEE, Philadelphia, Pennsylvania, 2005), pp. 829–832
18. G Sicot, S Houcke, J Barbier, Blind detection of interleaver parameters. Signal Process. **89**(4), 450–462 (2009)
19. M Marazin, R Gautier, G Burel, Blind recovery of $k/n$ rate convolutional encoders in a noisy environment. EURASIP J. Wireless Commun. Netw. **2011**(168), 1–9 (2011)
20. Z Jing, H Zhiping, L Chunwu, S Shaojing, Z Yimeng, Information-dispersion-entropy-based blind recognition of binary BCH codes in soft decision situations. Entropy. **15**(5), 1705–1725 (2013)
21. A Hocquenghem, Codes correcteurs d'erreurs. Chiffres. **2**, 147–156 (1959)
22. RC Bose, DK Ray-Chaudhuri, On a class of error correcting binary group codes. Inf. Control. **3**(3), 68–79 (1960)
23. I Reed, G Solomon, Polynomial codes over certain finite fields. J. Soc. Ind. Appl. Math. **8**(2), 300–304 (1960)
24. M Baldi, M Bianchi, F Chiaraluce, R Garello, N Maturo, IA Sanchez, S Cioni, in *2013 IEEE Military Communications Conference*. Advanced coding schemes against jamming in telecommand links (IEEE, San Diego, CA, 2013), pp. 1220–1226
25. C Junbin, W Lin, L Yong, in *International Conference on Communications, Circuits and Systems*. Performance comparison between non-binary LDPC codes and Reed-Solomon codes over noise bursts channels, vol. 1 (IEEE, Hong Kong, China, 2005), pp. 1–4
26. B Zhou, L Zhang, J Kang, Q Huang, YY Tai, S Lin, M Xu, in *Information Theory and Applications Workshop*. Non-binary LDPC codes vs. Reed-Solomon codes (IEEE, San Diego, CA, 2008), pp. 175–184
27. version 8.8.0 Release 8 GT, LTE; Evolved Universal Terrestrial Radio Access (E-UTRA); Multiplexing and Channel Coding. The 3rd Generation Partnership Project 2, Technical Specification Group Radio Access Network (2010). http://www.3gpp.org, (2013)
28. M Marazin, R Gautier, G Burel, in *IEEE GLOBECOM Workshops*. Dual code method for blind identification of convolutional encoder for cognitive radio receiver design (IEEE, Honolulu, HI, 2009)
29. EM Moro, *Algebraic geometry modeling in information theory*. (World Scientific, Singapore, 2012)
30. E Anderson, Z Bai, C Bischof, S Blackford, J Demmel, J Dongarra, JD Croz, A Greenbaum, S Hammarling, A McKenney, D Sorensen, *LAPACK user's guide*. (SIAM, Philadelphia, 1999)
31. Y Zrelli, *Identification aveugle de codes correcteurs d'erreurs basés sur des grands corps de Galois et recherche d'algorithmes de type décision souple pour les codes convolutifs*. (PhD thesis, Université de Brest, France, 2013)