Nouvelle méthode robuste d'identification aveugle de la taille des mots de code pour une transmission entachée d'erreurs avec généralisation aux codes correcteurs d'erreurs non-binaires

Yasamine ZRELLI, Roland GAUTIER, Mélanie MARAZIN, Emanuel RADOI

Université Européenne de Bretagne, Université de Brest; CNRS, UMR 6285 Lab-STICC
6 avenue Victor Le Gorgeu, 29200 Brest, France
yasamine.zrelli@univ-brest.fr, roland.gautier@univ-brest.fr
melanie.marazin@univ-brest.fr, emanuel.radoi@univ-brest.fr

Résumé – Dans un contexte non-coopératif, le récepteur doit être capable d'identifier en aveugle les paramètres des codes correcteurs d'erreurs avec la seule connaissance des données reçues. L'objet de cet article est de proposer une approche robuste permettant d'identifier en aveugle la taille des mots de code pour des codes non-binaires lorsque les données reçues sont entachées d'erreurs. Dans ces travaux, une étude théorique des approches existantes qui traitent du problème de l'identification des codes binaires est proposée. Cette étude nous a permis de mettre en avant des points faibles de ces méthodes. Nous proposons également une nouvelle méthode conçue autour d'un critère basé sur la variance du nombre de zéros présents dans les colonnes de matrices construites à partir des données reçues et transformées en des matrices triangulaires à l'aide de l'algorithme de pivot de Gauss. La méthode proposée est généralisée au cas des codes non-binaires. En comparant cette approche par rapport aux approches existantes, les résultats montrent que la méthode basée sur le critère de la variance est plus robuste et ne nécessite pas l'estimation de la probabilité d'erreur du canal pour identifier la taille des mots de code.

Abstract – In a non-cooperative context, the receiver must be able to blindly identify the parameters of error correcting codes from the only knowledge of the received data. The aim of this article is to propose a robust approach to blindly identify the codewords size parameter of non-binary codes in a noisy transmission environment. On the one hand, we study theoretically current approaches which deal with the problem of blind identification in the case of binary errors correcting codes in order to analyze their weak points. On the other hand, we propose a method developed by using the variance of the number of zeros in columns of matrices reshaped from received data and transformed into triangular matrices by the Gauss-Jordan elimination through pivoting algorithm. Our proposed method is generalized to non-binary errors correcting codes. The results obtained by this proposed approach, compared to the existing approaches show that our method based on the variance criterion is more robust and does not require the estimation of the error probability of the channel to identify the size of the codewords.

1 Introduction

L'introduction d'un système de codage performant à l'émission comme les codes correcteurs d'erreurs est indispensable pour combattre l'effet des perturbations introduites par le canal de transmission. Afin d'être en mesure d'effectuer l'opération de décodage, le récepteur a besoin de connaître les paramètres de codage utilisés à l'émission. Les technologies actuelles utilisées sont basées sur l'entente préalable entre l'émetteur et le récepteur sur le schéma de codage et ses paramètres. Ces technologies sont fonctionnelles, mais elles ne sont plus adaptées au développement des nouveaux schémas de codage plus performants et à la prolifération des nouvelles normes et standards de communication. Ainsi, les systèmes radio cognitifs fournissent une solution pertinente à ce problème à travers la conception de récepteurs intelligents qui sont capables d'identifier en aveugle les paramètres du système de codage avec la seule connaissance des données reçues.

Dans cet article, nous nous intéressons à l'identification aveugle de la taille des mots de codes en considérant une transmission bruitée. Les travaux sur la thématique de l'identification aveugle des paramètres du bloc de codage canal ont été limités pour la plupart jusqu'à présent à des codeurs binaires. Des techniques d'identification aveugle des codeurs en bloc ont été proposées dans [1, 2]. Dans [3], une méthode basée sur le critère du rang a été proposée pour des codes binaires et non-binaires. Nous avons démontré la pertinence de cette méthode pour des codes dans GF(2^m) sous l'hypothèse que les paramètres du corps de Galois (l'entier m et le polynôme primitif) utilisés à l'émission soient connus ou correctement identifiés par le récepteur. Dans le cas d'une transmission bruitée, les auteurs dans [4, 5] ont proposé une technique basée sur la recherche des colonnes dépendantes dans les matrices construites à partir des données reçues. Un des inconvénients de cette technique est qu'elle nécessite la connaissance de la probabilité d'erreur du canal de transmission.

Dans cet article, nous proposons une technique d'identification aveugle, robuste et généralisée aux codes correcteurs d'erreurs non-binaires qui ne nécessite pas en entrée la connais-

sance de la probabilité d'erreur du canal. Le principe de cette technique est basée sur l'utilisation du critère de la variance.

2 Principe de l'identification aveugle pour les codes binaires et non-binaires

Dans le cadre de cette étude, nous considérons un canal non-binaire symétrique de probabilité d'erreur p_e. En effet, l'ensemble de modulateur d'ordre $q = 2^m$, canal de transmission et démodulateur est équivalent à un canal non-binaire symétrique lors d'une décision dure [6]. Pour un canal de transmission, il suffit de déterminer la probabilité d'erreur d'un symbole à la sortie du démodulateur afin d'appliquer notre algorithme d'identification aveugle. Dans ce cas, la probabilité d'erreur dépend du type de modulation utilisée et des caractéristiques physiques du canal de transmission.

Les données reçues $r_i, \forall i \in \{1, ..., L\}$, sont réorganisées pour construire des matrices $\mathbf{R}_l$, de taille $(M \times l)$, avec $M > l$. Le nombre de colonnes l varie entre 2 et l_{max} et le nombre de lignes M dépendant de l est déterminé par $M = \lfloor L/l \rfloor$.

Dans le cas d'une transmission non-bruitée, les matrices $\mathbf{R}_l$ présentent des déficiences de rang pour $l = \alpha \cdot n, \forall \alpha \in \mathbb{N}$, avec n la taille des mots de code. En pratique, le rang d'une matrice est calculé en déterminant le nombre de colonnes qui sont linéairement dépendantes, ce qui permet d'en déduire le nombre de celles qui sont indépendantes. Dans le contexte d'une transmission bruitée, cette dépendance est perturbée par la présence des erreurs dans certains symboles. Par conséquent, toutes les matrices $\mathbf{R}_l$ sont de rang plein. Dans un tel contexte, les auteurs dans [4, 5] ont proposé deux algorithmes pour identifier les paramètres de codes convolutifs binaires et les paramètres des entrelaceurs binaires. L'idée de ces deux algorithmes consiste à chercher les colonnes qui sont "presque dépendantes" dans les matrices $\mathbf{R}_l$ en utilisant l'algorithme de Gauss dans GF(2). Dans le cas des codes non-binaires, il suffit de trianguliariser ces matrices à l'aide de l'algorithme d'élimination de Gauss adapté au corps fini GF($q = 2^m$) tout en faisant des permutations sur les colonnes. La transformation appliquée à $\mathbf{R}_l$ est une application linéaire définie par :

$$\mathbf{R}_l \cdot \mathbf{A}_l = \mathbf{T}_l \tag{1}$$

où $\mathbf{A}_l$ est une matrice de taille $(l \times l)$ qui représente les permutations et les combinaisons de colonnes. La matrice triangulaire obtenue, notée $\mathbf{T}_l$, est de taille $(M \times l)$. La méthode d'identification de la taille des mots de code binaires présentée dans [5] est basée sur la recherche des colonnes dépendantes en étudiant le nombre de 1 (ou 0) dans les colonnes des matrices binaires $\mathbf{T}_l$.

Dans cet article, nous proposons une nouvelle méthode basée sur le calcul de la variance du nombre de 0 dans les colonnes de $\mathbf{T}_l$.

3 Méthode classique d'identification pour les codes binaires

Notons $N_i(l)$ une variable contenant le nombre de 0 de la i-ème colonne de la matrice $\mathbf{T}_l$. $N_i(l)$ est étudiée en fonction de l :

- Pour $l \neq \alpha \cdot n$ ou $l < n_a$ avec n_a la taille de la première matrice de rang déficient : les matrices $\mathbf{R}_l$ seront de rang plein. Dans ce cas, la variable $N_i(l)$ suivra la loi binomiale $\mathcal{B}(M, 1/2)$.
- Pour $l = \alpha \cdot n$ et $l \geqslant n_a$: la variable $N_l(i)$ aura deux comportements en fonction de i :
  - Si i est l'indice d'une colonne dépendante : $N_l(i)$ suivra la loi binomiale $\mathcal{B}(M, P_i)$, avec P_i la probabilité qu'un élément de la i-ème colonne de $\mathbf{T}_l$ soit nul.
  - Si i n'est pas l'indice d'une colonne dépendante : $N_l(i)$ suivra la loi binomiale $\mathcal{B}(M, 1/2)$.

Notons $Q(l)$ un ensemble contenant les colonnes dépendantes de la matrice $\mathbf{R}_l$ et γ le seuil de décision. L'appartenance de la i-ème colonne de $\mathbf{R}_l$, notée $\mathbf{r}_i^{(l)}$, à $Q(l)$ peut être déterminée au regard des deux comportements de $N_l(i)$ comme suit :

$$\begin{cases} \text{Si } N_l(i) > \frac{M}{2} \cdot \gamma \text{ alors } \mathbf{r}_i^{(l)} \in Q(l) \\ \text{Si } N_l(i) \leq \frac{M}{2} \cdot \gamma \text{ alors } \mathbf{r}_i^{(l)} \notin Q(l) \end{cases} \tag{2}$$

Les auteurs dans [5] ont montré que le seuil γ dépend de la probabilité d'erreur du canal p_e. Après avoir déterminé ce seuil, la taille des mots de code est identifiée par :

$$n = \text{mode}(\text{diff}(\mathcal{I})) \tag{3}$$

avec $\mathcal{I}$ l'ensemble défini par :

$$\mathcal{I} = \{l = 2, \cdots, l_{max} | card(Q(l)) \neq 0\} \tag{4}$$

Le vecteur diff($\mathcal{I}$) contient la différence entre deux éléments consécutifs du vecteur $\mathcal{I}$. La fonction mode permet d'obtenir la valeur de la plus fréquente occurrence dans le vecteur diff($\mathcal{I}$).

La méthode présentée dans cette section peut être appliquée uniquement aux codes binaires et elle n'est pas robuste puisqu'elle a besoin d'une estimation fiable de la probabilité d'erreur p_e.

4 Nouvelle méthode robuste généralisée (codes binaires et non-binaires)

Dans cette section, nous présentons une méthode d'identification plus robuste et généralisée aux codes correcteurs d'erreurs non-binaires qui traitent des symboles appartenant au corps de Galois GF(2^m). Cette méthode consiste à calculer la variance du nombre de 0 dans les colonnes des matrices non-binaires $\mathbf{T}_l$ ($N_l(i)$). Notons V_l la variable représentant la variance de $N_l(i)$. Cette variable est définie par :

$$V_l = \frac{\sum_{i=1}^{l} (N_l(i) - E_l)^2}{l} \tag{5}$$

où E_l est la moyenne arithmétique de $N_l(i)$. La variance V_l a deux comportements différents en fonction de $l, \forall \alpha \in \mathbb{N}$:

– Si $l \neq \alpha \cdot n$ ou $l < n_a$, $N_l(i)$ suit la loi normale de paramètres $\mu_1 = M/q$ et $\sigma_1^2 = M \cdot (q-1)/q^2$, notée $\mathcal{N}(\mu_1, \sigma_1^2)$, pour toutes les colonnes i de $\mathbf{T}_l$. Alors la variance V_l suivra :

$$V_l \rightarrow \frac{\sigma_1^2}{l} \cdot \mathcal{X}_{l-1}^2 \tag{6}$$

où $\mathcal{X}_{l-1}^2$ est la loi du chi-2 de paramètre $l-1$. Dans ce cas, la moyenne de V_l normalisée par M^2 sera proche de $(q-1)^2/q^4$.

– Si $l = \alpha \cdot n$ et $l \geq n_a$:
  – Si la i-ème colonne est dépendante, $N_l(i)$ suivra la loi normale $\mathcal{N}(\mu_0, \sigma_0^2)$ avec $\mu_0 = M \cdot P_i$ et $\sigma_0^2 = M \cdot P_i \cdot (1-P_i)$.
  – Si la i-ème colonne n'est pas dépendante, $N_l(i)$ suivra la loi normale $\mathcal{N}(\mu_1, \sigma_1^2)$.

  Ainsi, la variable V_l suivra la loi :

$$V_l \rightarrow \frac{\sigma_0^2}{l} \cdot \mathcal{X}_{Q(l)-1}^2 + \frac{\sigma_1^2}{l} \cdot \mathcal{X}_{l-Q(l)-1}^2 \tag{7}$$

On peut remarquer deux comportements de V_l en fonction de l qui peuvent être distingués par :

$$\begin{cases} \text{Si } l \neq \alpha \cdot n \text{ ou } l < n_a \text{ alors } \frac{V_l}{M^2} \leq \frac{(q-1)^2}{q^4} \\ \text{Si } l = \alpha \cdot n \text{ et } l \geq n_a \text{ alors } \frac{V_l}{M^2} > \frac{(q-1)^2}{q^4} \end{cases} \tag{8}$$

Nous noterons $\mathcal{J}$ un ensemble défini par :

$$\mathcal{J} = \left\{ l = 2, \cdots, l_{max} | \frac{V_l}{M^2} > \frac{(q-1)^2}{q^4} \right\} \tag{9}$$

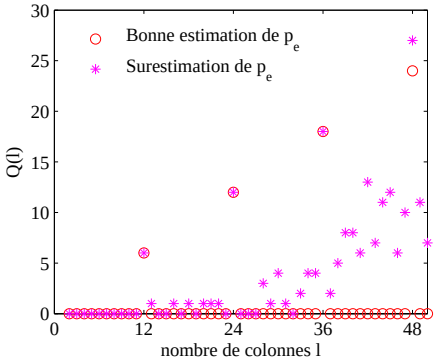
De ce fait, la taille des mots de code n sera identifiée en utilisant la méthode de la variance par : $n = \text{mode}(\text{diff}(\mathcal{J}))$.

5 Étude comparative des méthodes présentées

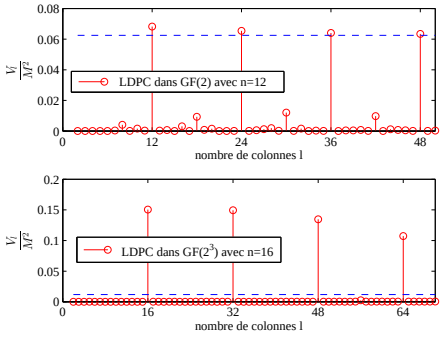
L'objectif des méthodes présentées dans cet article est d'identifier en aveugle la taille des mots de code n. La nouvelle méthode proposée permet de traiter aussi bien les codes binaires que non-binaires et ne nécessite pas d'estimer la probabilité d'erreur p_e. Nous illustrons la différence entre les deux méthodes lors d'une surestimation de p_e à 0.1 au lieu de 0.01. Pour cela, nous considérons un code LDPC binaire de taille de mots de code $n = 12$. La figure 1(a) représente le nombre de colonnes dépendantes $Q(l)$ en fonction de l dans les cas d'une bonne estimation et d'une surestimation de p_e. Dans le cas d'une bonne estimation de p_e, l'ensemble $\mathcal{I} = \{12, 24, 36, 48\}$ nous permet de déterminer la bonne taille $n = 12$. Par contre, il est clair que lors d'une surestimation de p_e, il est impossible d'identifier la bonne valeur de n par la méthode classique présentée dans la section 3 puisque la taille de mots de code identifiée est $n = \text{mode}(\text{diff}(\mathcal{I})) = 1$, avec :

$$\mathcal{I} = \{10, 13, 14, 15, 16, 18, 20, 21, 23, 24, \cdots\}$$

Dans la figure 1(b), la variance V_l normalisée par M^2 et la droite du seuil $(q-1)^2/q^4$ sont représentées en fonction de



(a) Nombre de colonnes dépendantes $Q(l)$



(b) Variance $\frac{V_l}{M^2}$

FIGURE 1 – Impact de la surestimation de p_e sur $Q(l)$ et $\frac{V_l}{M^2}$

l dans le cas d'un code LDPC dans GF(2), avec $n = 12$ et dans le cas d'un code LDPC dans GF(2^3), avec $n = 16$. Dans les deux cas, n peut être identifiée correctement. Cette seconde méthode ne nécessitant pas l'estimation de p_e est plus robuste que la première.

On peut aussi remarquer que le seuil qui permet de distinguer les deux comportements de V_l/M^2 n'est pas optimal puisqu'il dépend du cardinal du corps q. Pour cette raison, nous choisissons un seuil optimal qui peut être déterminé par :

$$\max\left(V_l/M^2\right)/2$$

Ce seuil sera considéré pour l'analyse des performances de deux méthodes d'identification.

Nous considérons le critère de la probabilité de détection de la bonne taille n pour comparer les performances des deux méthodes d'identification. Un code LDPC de paramètres $n = 6$ et $k = 3$, travaillant dans le corps GF(q), avec $q \in \{2, 4, 8, 16\}$, est pris en compte. Les matrices R_l sont construites à partir des données reçues bruitées de taille $L = 30000$. Le nombre de lignes M de ces matrices est fixé à 1000 et le nombre de colonnes l varie entre 2 et 30. Afin de déterminer les probabilités de détection de n avec les deux méthodes présentées dans cet article, 1000 itérations de Monte-Carlo sont réalisées. Dans la figure 2, les courbes de ces probabilités sont représentées en fonction de p_e avec plusieurs dimensions du corps de Galois

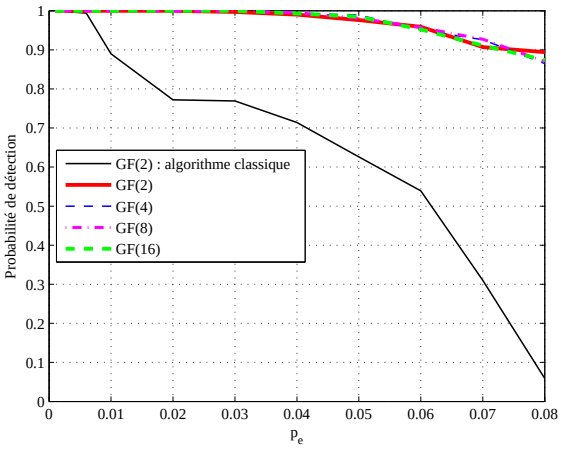


FIGURE 2 – Probabilité de détection de n pour un code LDPC ($n = 6$, $k = 3$) dans GF(q), avec $q \in \{2, 4, 8, 16\}$

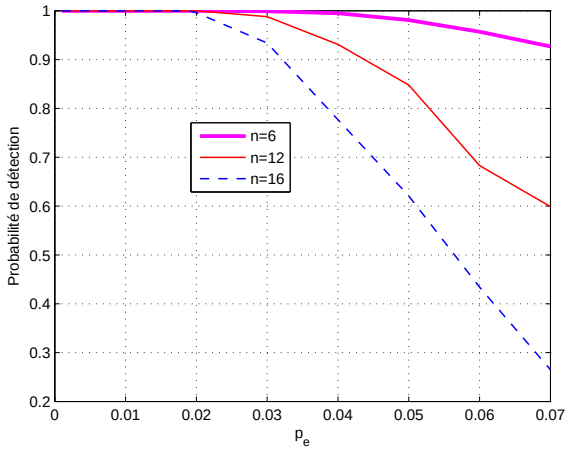


FIGURE 3 – Probabilité de détection de n pour des codes LDPC de taille $n = 6, 12, 16$ dans GF(8)

pour la méthode de la variance. Dans le cas du corps GF(2), nous constatons une amélioration importante de la probabilité de détection de n avec la méthode de la variance. En effet, pour $p_e = 0.07$, notre méthode apporte un gain de 66% par rapport à la méthode classique. Analysons maintenant l'effet de l'augmentation de la dimension du corps q sur les performances de détection de la méthode de la variance pour un code LDPC de taille $n = 6$. Dans la figure 2, on peut remarquer que les performances pour différentes valeur de q sont quasiment similaires et que la probabilité de détection est proche de 1 pour $p_e \leq 0.05$. Ainsi, la dimension du corps GF(q) n'a pas d'influence importante sur les performances de notre méthode pour le code LDPC de paramètre $n = 6$.

Comparons maintenant les performances obtenues dans la figure 3 lorsqu'on augmente la taille du code LDPC dans GF(8). Nous constatons que les performances sont moins bonnes lorsque la taille n croît. Notons que pour des tailles de n grandes, il est nécessaire d'augmenter la taille des données reçues $L > 4 \cdot n \cdot M$, ce qui provoque une augmentation de la complexité de l'algorithme de triangularisation.

6 Conclusion

Nous avons proposé dans cet article une nouvelle méthode d'identification aveugle de la taille des mots de code. Cette méthode est généralisée pour les codes correcteurs d'erreurs non-binaires. Elle est basée sur le critère de la variance pour détecter les matrices R_l dont le nombre de colonnes l est multiple de n. En comparant notre méthode avec la méthode existante d'identification pour les codes binaires, nous montrons que notre méthode a l'avantage de ne pas nécessiter l'estimation de la probabilité d'erreur du canal et qu'elle offre d'excellentes performances, même en augmentant la dimension du corps de Galois q. Cependant, pour des tailles de code plus grandes, on a besoin de plus de données afin d'obtenir d'excellentes probabilités de détection. Le problème de la complexité de l'algorithme proposé fait l'objet de nos travaux futurs.

Références

[1] M. Cluzeau and M. Finiasz, *Recovering a Code's Length and Synchronization from a Noisy Intercepted Bitstream*, IEEE International Symposium on Information Theory, 2009.

[2] G. Burel and R. Gautier, *Blind estimation of encoder and interleaver characteristics in a non cooperative context*, IASTED International Conference on Communications, Internet and Information Technology, Scottsdale, AZ, USA, 2003.

[3] Y. Zrelli, M. Marazin, R. Gautier et E. Rannou. *Blind identification of convolutional encoder parameters over* GF*(2^m) in the noiseless case*. Proceedings of the International Conference on Computer Communication Networks, Maiu, Hawaii, 2011.

[4] G. Sicot and S. Houcke, *Blind detection of interleaver parameters*, IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), vol. 3, pp. 829-832, Philadelphia, Pennsylvania, 2005.

[5] M. Marazin, R. Gautier and G. Burel, *Blind recovery of k/n rate convolutional encoders in a noisy environment*, EURASIP Journal on Wireless Communications and Networking 2011, 168 (2011) 1-9.

[6] W. Ryan and S. Lin, *Channel Codes : Classical and Modern*, Cambridge University Press, 2009.