Identification des relations de parité d'un code de canal à partir de données souples

Aurélien Bonvard, Sébastien Houcke

IMT Atlantique - Département Mathematical and Electrical Engineering, Lab-STICC - SI3 - UMR 6285, Brest, France
aurelien.bonvard@imt-atlantique.fr, sebastien.houcke@imt-atlantique.fr

Résumé – Dans un contexte d'interception de signaux et d'identification aveugle de codes correcteurs d'erreurs, la recherche d'éléments du code dual est une étape nécessaire à l'identification du code. Nous proposons une approche tirant profit des données souples. Notre méthode exploite la distribution particulière des distances euclidiennes entre des vecteurs de fiabilités des bits impliqués dans une relation de parité de poids w. Enfin, nous comparons les performances de notre méthode avec une méthode existante.

Abstract – In a context of signal interception and blind identification of error-correcting codes, the search for elements of the dual code is a necessary step. To this end, we propose an approach using soft data. Our method is based on highlighting a particular distribution of distances in a w-dimensional space in the presence of a w-weighted parity check. It solves a common sub-problem of dual code reconstruction. Finally, we compare the performance of our approach with an existing method.

1 Introduction

Dans la plupart des systèmes de communication, l'utilisation de codes correcteurs est désormais d'usage. En effet, ces codes permettent de rendre les transmissions plus robustes aux aléas liés aux média utilisés. Á cette fin, l'émetteur et le récepteur se mettent d'accord sur le type de codage et échangent en suivant cette règle. Dans un contexte non-coopératif, il est courant que le récepteur (parfois intercepteur) n'ait pas de connaissances a priori sur la nature du code utilisé.

Dans ces circonstances, il est nécessaire de déterminer les paramètres du codeur de canal en aveugle. Une fois la taille du mot de code identifié (ce que nous supposerons dans cet article), il est nécessaire d'identifier les relations de parité du code ce qui revient à reconstruire son code dual. Une synthèse des approches existantes a été proposée dans [1]. Ces approches reposent sur des considérations combinatoires tels que [2] [3], [4] et [5]. Deux grandes approches existent. La première consiste à construire une matrice dont chaque ligne est un mot de code intercepté. La diagonalisation de cette matrice permet d'identifier le dual du code. Cette méthode est efficace mais devient vite prohibitive lorsque la taille du code augmente. Les performances chutent si les relations de parité sont de poids fort ou si niveau de bruit est élevé. La deuxième approche consiste à chercher et tester les parités sur l'ensemble des mots interceptés. Ces approches, basées sur l'algorithme de Dumer [5] [6] sont aussi vites prohibitives (même s'il existe des algorithmes permettant de réduire la complexité) lorsque la taille du mot de code augmente et ne peuvent être mise en place que pour la recherche de parités de poids faibles. Pour palier à ces problèmes [7] propose de réaliser un pivot de gauss partiel (i.e. première approche) et ensuite de chercher de manière exhaustive des relations de parité de poids faible (i.e. deuxième approche) sur la partie de la matrice non diagonalisée. En combinant les résultats, ils montrent qu'ils peuvent ainsi identifier des relations de parités de poids plus élevé et sur des codes plus longs. Il est donc tout a fait pertinent de s'intéresser à l'identification de relation de parité de poids faible.

Dans cet article, nous proposons une nouvelle méthode permettant d'identifier des relations de parité de poids faible mais qui contrairement à la méthode de Dumer exploite la connaissance de la fiabilité des bits interceptés. On montre ainsi que notre méthode permet d'identifier plus de relations pour un même niveau de SNR. De plus, elle ne nécessite que peu de mots de code interceptés pour fonctionner.

Pour l'identification des parités d'un code, nous supposons connaître la taille du code, notée n. Nous tirons aléatoirement un vecteur h de poids w : h est un vecteur de taille n ayant w composantes non nulles et égale à 1. Nous testons si ce vecteur h est une relation de parité en comparant la valeur d'un critère à un seuil. Le critère de décision est basé sur une mesure de distance Euclidienne entre les vecteurs des fiabilités des bits impliqués dans la parité testée.

La suite de cette étude suit la progression suivante : la section 2 introduit le contexte et les notations utilisées. La section 3 présente les principes de l'algorithme de re-

construction dépendant de la distribution des distances euclidiennes. Enfin, nous proposons des résultats permettant de comparer cette méthode à l'état de l'art.

2 Notations et contexte

2.1 Modèle de transmission

Dans cet article, nous notons $\mathcal{C}(n,k)$ un code binaire linéaire en bloc de longueur n et de rendement $\rho = \frac{k}{n}$ avec $k < n$. Notez que ce code peut être celui généré par la sous matrice issue du pivot de Gauss partiel (cf [7]).

La matrice Y comporte M lignes (i.e. M mots de code interceptés) et n colonnes. Soit $y(i,j)$ l'élément situé à la i-ième ligne et j-ième colonne. $y(i,j)$ représente la fiabilité du bit correspondant. $y(i,j)$ est positif si le bit décidé est un et est négatif si le bit décidé est zéro. Une manière simple de générer des fiabilités consiste à moduler les bits par une BPSK et d'ajouter un bruit blanc Gaussien, on obtient ainsi une valeur d'un échantillon qui est homogène à une fiabilité. C'est ce que nous considérons dans la suite de l'article.

$$y(i,j) = s(i,j) + w(i,j)$$

avec $(i,j) \in [\![1,M]\!] \times [\![1,n]\!]$. Les échantillons $s(i,j)$ sont le résultat d'une modulation de phase à deux états sur les valeurs issues du codage de canal : pour tout (i,j), $s(i,j)$ appartient à $\{\pm 1\}$. $w(i,j)$ représente un échantillon de bruit blanc gaussien centré et de variance σ_b^2. La méthode proposée n'est cependant pas restrictive à ce type de modulation, elle peut être mise en place dès que nous disposons d'une mesure de fiabilité sur chaque bit.

2.2 Construction de la sous-matrice Y_h

La recherche exhaustive de relations de taille n et de poids w revient à tester $\binom{n}{w}$ vecteurs h de poids w. Pour chaque vecteur h, on extrait de Y un sous-matrice Y_h en ne gardant que les w colonnes de Y indexées par le support de h. Nous verrons par la suite que pour décider si le vecteur h testé est bien une parité du code, il n'est pas nécessaire d'avoir dans la matrice Y_h un nombre de lignes important. Nous pouvons donc sélectionner et retenir uniquement les lignes les plus fiables. La fiabilité de chaque ligne est déterminée par celle de son bit le moins fiable.

Définition 1 *Soit deux mots de taille w, $\mathbf{m_1}$ et $\mathbf{m_2}$:*
$\mathbf{m_1}$ *est plus fiable que* $\mathbf{m_2}$ $(\mathbf{m_1} < \mathbf{m_2})$ *ssi* $\min_{j\in[\![1,w]\!]}(\{|m_1(j)|\}) < \min_{j\in[\![1,w]\!]}(\{|m_2(j)|\})$.

Nous gardons donc dans Y_h les L lignes $\{i_j\}_{j=1,\dots,L}$ les plus fiables (au sens de la définition 1). Cette procédure permet de virtuellement augmenter le SNR sur la partie des données que nous allons utiliser ensuite. Y_h est donc une matrice de taille $L \times w$. Nous proposons dans la section 3.2 une heuristique permettant de fixer L.

3 Reconstruction par distribution des distances euclidiennes

3.1 Description de la méthode

Á partir de Y_h, nous calculons toutes les distances deux à deux des lignes de Y_h. Nous allons étudier la distribution de ces distances euclidiennes. Notons $x^{(h)} = (x_1^{(h)}, x_2^{(h)}, \cdots, x_r^{(h)})$ les $r = L(L-1)/2$ distances mesurées à partir de la matrice Y_h. Ces distances sont ordonnées de telle sorte que : $x_1^{(h)} \leqslant x_2^{(h)} \leqslant \cdots \leqslant x_r^{(h)}$.

Deux cas de figure sont possibles :

- Lorsque le vecteur testé n'appartient pas au code dual de $\mathcal{C}(n,k)$, la distribution des distances correspond alors à celle des distances pour des mots i.i.d. de longueur w. Notons $X^{(iid)} = \{X_i^{(iid)}\}_{i\in\{1,\dots,r\}}$ le vecteur aléatoire des distances ordonnées obtenues en calculant les distances euclidiennes deux à deux de mots BPSK i.i.d. bruités (de variance σ_b^2) et de longueur w. Le vecteur est ordonné de sorte que : $X_1^{(iid)} \leqslant X_2^{(iid)} \leqslant \cdots \leqslant X_r^{(iid)}$. La figure 1 représente $x_i^{(h)}$ et $\mathbb{E}\left(X_i^{(iid)}\right)$ en fonction de i. On peut constater que les deux courbes suivent le même comportement.

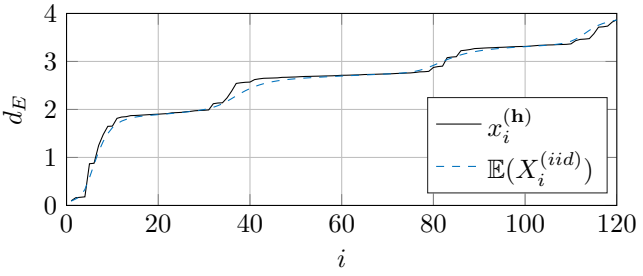

FIGURE 1 – Valeur des distances moyennes ordonnées dans le cas de mots iid

- Lorsque le vecteur testé appartient effectivement au code dual de $\mathcal{C}(n,k)$, la distribution des distances deux à deux des mots formés correspond à celle des distances pour un code de parité de longueur w. Notons $X^{(par)} = \{X_i^{(par)}\}_{i\in\{1,\dots,r\}}$le vecteur aléatoire des distances ordonnées obtenues en calculant les distances euclidiennes deux à deux de mots BPSK bruités de taille w tirés uniformément parmi les mots de longueur w vérifiant la condition de parité (la somme des bits est égale à zéro). La figure 2 représente $\mathbb{E}\left(X_i^{(par)}\right)$ et $x_i^{(h)}$ en fonction de i. On peut constater que les deux courbes suivent le même comportement. De plus si nous comparons les figures 1 et 2, nous constatons que les distances dans le cas où h correspond à une parité et dans le cas contraire sont clairement différentes.

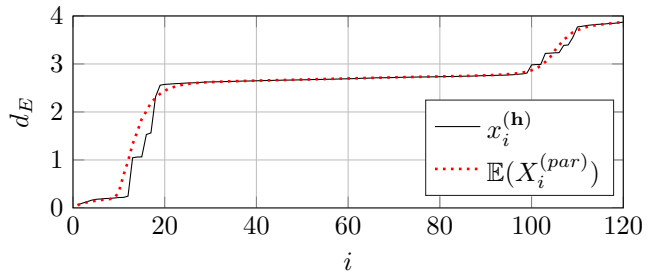

FIGURE 2 – Valeur des distances moyennes ordonnées dans le cas d'une relation de parité

La détection de potentielles relations de parité h appartenant au code se fait donc par la comparaison de la distribution observée des distances à celles dans les cas d'une trame i.i.d. (i.e. $\mathbf{X}^{(iid)}$) et d'une trame issue d'un code de parité (i.e. $\mathbf{X}^{(par)}$). Ces deux quantités peuvent être calculées au préalables et stockés dans des tables, notons que pour les générer il est nécessaire d'estimer la variance σ_b^2.

3.2 Estimation du vecteur des distances deux à deux ordonnées

La valeur de r (le nombre de distances calculées deux à deux) dépend du poids des relations recherchées. En effet, les distances calculées sont celles entre des mots de taille w. Il y a donc 2^w mots différents sur w bits. L'utilisation de 2^w mots suffit à l'obtention d'une mesure représentative de la distribution des distances dans le cas du code de parité ou de la trame i.i.d. Ainsi en fixant à 2^w le nombre de mots traités, on obtient $r = \frac{2^w(2^w-1)}{2}$ distances calculées.

Pour obtenir une estimation fiable du vecteur de distance ordonnée, nous calculons ce vecteur sur plusieurs blocs de 2^w mots issus de la matrice Y_h. Nous notons N_T, le nombre de blocs retenu pour l'identification ($N_T \in [\![1, \lfloor \frac{M}{2^w} \rfloor]\!]$). Pour chacun des blocs, les $r = \frac{2^w(2^w-1)}{2}$ distances euclidiennes sont calculées et ordonnées. On note $\mathbf{x}_j^{(h)}$, le vecteur de distances ordonnées issu du j-ième bloc de Y_h : $\mathbf{x}_j^{(h)} = (x_{j,1}^{(h)}, x_{j,2}^{(h)}, \cdots, x_{j,r}^{(h)})$, avec $j \in [\![1, N_T]\!]$. Dans un soucis de clarté, $x_{j,2}^{(h)}$ est, par exemple, la deuxième valeur la plus faible du vecteur de distance issu du j-ième bloc lors du test du vecteur h. Les distances moyennes ordonnées $\hat{x}^{(h)}$ pour le vecteur testé h sont calculées à partir de tous ces blocs :

$$\begin{aligned}\hat{\mathbf{x}}^{(h)} &= \frac{1}{N_T}\sum_{j=1}^{N_T}(x_{j,1}^{(h)}, x_{j,2}^{(h)}, \cdots, x_{j,r}^{(h)}) \\ &= \left(\frac{1}{N_T}\sum_{j=1}^{N_T} x_{j,1}^{(h)}, \frac{1}{N_T}\sum_{j=1}^{N_T} x_{j,2}^{(h)}, \cdots, \frac{1}{N_T}\sum_{j=1}^{N_T} x_{j,r}^{(h)}\right) \\ &= (\hat{x}_1^{(h)}, \hat{x}_2^{(h)}, \cdots, \hat{x}_r^{(h)}) \qquad (1)\end{aligned}$$

Le nombre L de lignes à conserver dans Y_h est donc de $L = N_T \times 2^w$.

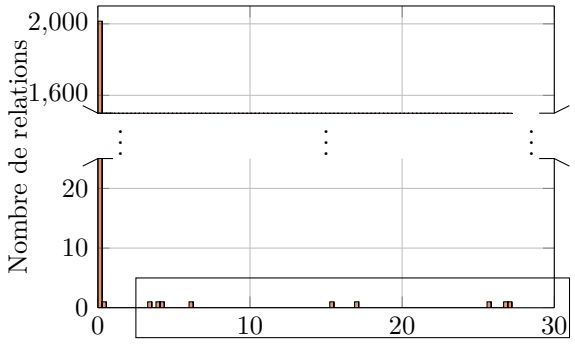

FIGURE 3 – Histogramme des $\phi(h)$ pour$\sigma_b = 5$ dB

Les variables aléatoires distribuées comme les distances d'une trame i.i.d. (et pour des mots issus d'un code de parité respectivement) sont notées $X^{(iid)}$ (respectivement $X^{(par)}$). Ainsi, lorsque h est une relation de parité, $\hat{x}^{(h)}$ suit la même loi de probabilité que $X^{(par)}$. Á l'inverse, lorsque h n'est pas une relation de parité, $\hat{x}^{(h)}$ suit la même loi de probabilité que $X^{(iid)}$. Deux calculs de distance quadratique permettent de quantifier l'adéquation des distances mesurées $\hat{\mathbf{x}}^{(h)}$ à l'une ou l'autre des distributions :

$$\mathcal{A}_h^{(par)} = \sum_{i=1}^{r}(\hat{x}_i^{(h)} - \mathbb{E}(X_i^{(par)}))^2 \qquad (2)$$

et

$$\mathcal{A}_h^{(iid)} = \sum_{i=1}^{r}(\hat{x}_i^{(h)} - \mathbb{E}(X_i^{(iid)}))^2 \qquad (3)$$

3.3 Critère de détection

Finalement, notre critère $\phi(h)$ pour décider si un vecteur h est effectivement une relation de parité repose sur la valeur que prend le rapport :

$$\phi(h) = \mathcal{A}_h^{(iid)} / \mathcal{A}_h^{(par)}.$$

Plus sa valeur est élevée, plus il est probable que la relation testée appartienne effectivement au code dual. En effet, cela correspond au cas où $\mathcal{A}_h^{(iid)}$ et $\mathcal{A}_h^{(par)}$ ont respectivement une valeur très grande et très faible conjointement.

Pour $n = 20$ et $w = 4$, la Figure 3 est un histogramme donnant un exemple de distribution empirique de notre critère $\phi(h)$ à 5 dB. Pour la plupart des relations, le critère a une valeur comprise entre 0 et 1 : ces relations n'appartiennent pas au code dual. Cependant, 9 relations voient leur valeur de critère s'écarter significativement du mode de cette distribution empirique. Ces valeurs correspondent aux vecteurs h telles que $h \in \mathcal{C}^T$. La sélection de ces relations passe donc par le choix d'un seuil sur la valeur minimale du critère de décision. Par exemple, le choix d'un seuil à 2 permet de détecter ces 9 relations de parité.

TABLE 1 – Valeur du seuil optimal pour différents E_b/N_0

E_b/N_0	-0.5	0	0.5	1	1.5	2
Seuil	4.5	3.9	2.7	2.7	1.1	0.7

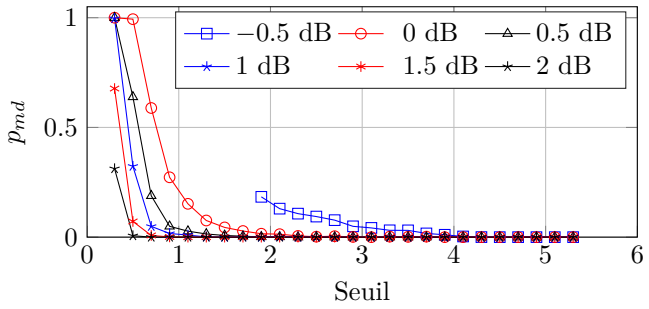

FIGURE 4 – Probabilité de mauvaise détection en fonction du seuil sur la valeur de $\phi(h)$

4 Performances de la méthode

Afin de mesurer l'intérêt de l'utilisation des informations souples pour la reconstruction d'un code correcteur d'erreur, nous comparons des résultats d'une méthode par collision sur les informations fermes (comme dans [3]) à la nôtre. Les simulations ont été effectuées sur un code en bloc de taille $n = 20$ et de dimension $k = 10$ pour 500 mots interceptés. La recherche se limite aux relations de poids $w = 4$, elles sont au nombre de 10.

Dans le cas de notre méthode, il est nécessaire de choisir un seuil sur la valeur de notre critère $\phi(h)$. Nous n'avons pas d'expression théorique du seuil, cependant nous pouvons le fixer expérimentalement. En effet, la figure 4 représente la probabilité de mauvaise détection d'une parité en fonction de la valeur du seuil et cela pour des rapports signal-à-bruit allant de -0.5 à 2 dB, cette simulation permet de déterminer le seuil minimum pour lequel la probabilité de mauvaise détection est minimale. La Table 1 recense les seuils optimaux pour différents niveaux de bruit au regard des résultats de cette simulation. Le choix du seuil nécessite donc la connaissance de σ_b^2 qui est estimée directement à partir des fiabilités :

$$\hat{\sigma_b^2} = \frac{1}{Mn} \sum_{k=1,\dots,n \;\; l=1,\dots,M} \left(y(l,k) - \text{sign}\left(y(l,k)\right)\right)^2$$

La Figure 5 compare les performances des deux méthodes en terme de robustesse au bruit. Dans ces simulations, le nombre de blocs N_T de 2^4 mots interceptés est de 10.

5 Conclusion

En l'état, la méthode proposée est plus robuste au bruit que la méthode par collision. Une amélioration immédiate de la méthode consisterait à trouver une expression analytique pour le seuil sur la valeur de notre critère : $\phi(h)$. Par

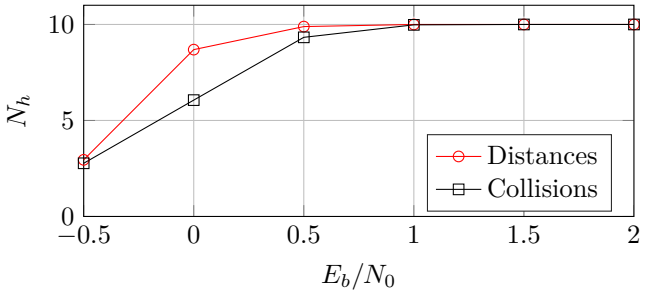

FIGURE 5 – Nombre N_h de relations identifiées en fonction du SNR

ailleurs, l'objectif est d'identifier les relations h pour lesquelles $\mathcal{A}_h^{(iid)}$ est grand et $\mathcal{A}_h^{(par)}$ est petit conjointement : il est donc envisageable de poser des seuils spécifiques à $\mathcal{A}_h^{(iid)}$ et à $\mathcal{A}_h^{(par)}$.

Enfin, dans certaines méthodes de reconstruction de codes longs (e.g. [7]), l'application d'un pivot partiel sur les données reçues permet de réduire la complexité. En effet, cette manipulation crée une sous-matrice sur laquelle des relations de poids et de tailles faibles sont identifiées. La méthode présentée dans cet article est adaptée à ce type de problème. C'est pourquoi les résultats exposés portent sur un code court. Une fois que des relations ont été identifiées pour cette sous-matrice, il est aisé d'utiliser la forme réduite de la matrice des données reçues pour obtenir les relations de parité du code long.

Références

[1] A. Tixier. *Reconnaissance de codes correcteurs.* Theses, Université Pierre et Marie Curie, Octobre 2015.

[2] G. Sicot, S. Houcke, and J. Barbier. Blind Detection of interleaver parameters. *ELSEVIER Signal Processing*, 20 :450–462, November 2008.

[3] M. Cluzeau and M. Finiasz. Recovering a Code's Length and Synchronization from a Noisy Intercepted Bitstream. ISIT, 2009.

[4] Jacques Stern. A method for finding codewords of small weight. In Gérard Cohen and Jacques Wolfmann, editors, *Coding Theory and Applications*, pages 106–113, Berlin, Heidelberg, 1989. Springer Berlin Heidelberg.

[5] Ilya Dumer. On minimum distance decoding of linear codes. In *5th Joint Soviet-Swedish Int. Workshop on Inform. Theory*, pages 50–52, Moscow, Russia, 1991.

[6] M. Bellard and J.-P. Tillich. Detecting and reconstructing an unknown convolutional code by counting collisions. pages 2967–2971. ISIT, 2014.

[7] K. Carrier and J.-P. Tillich. Identifying an unknown code by partial Gaussian elimination. WCC, 2017.