Codage de Sources avec Information Adjacente et Connaissance Imparfaite de la Corrélation : le problème des cadrans

Elsa DUPRAZ[1], Aline ROUMY[2], Michel KIEFFER[1]

[1]L2S - CNRS - SUPELEC - Univ Paris-Sud, 3 rue Joliot Curie, 91192 Gif-sur-Yvette, France

[2]INRIA, Campus de Beaulieu, 35042 Rennes, France

elsa.dupraz@lss.supelec.fr, aline.roumy@inria.fr, michel.kieffer@lss.supelec.fr

Thème – 2.1 - Théorie de l'information, 2.6 - Synchronisation, estimation, 4.1 - Détection et estimation statistiques

Problème traité – Cet article aborde le problème du codage d'une source X avec information adjacente Y disponible uniquement au décodeur dans le cas où $P(Y|X)$ est mal connue au décodeur. Ce type de problème apparait lors de la compression de données prélevées par des capteurs en réseau. Nous considérons des sources q-aires et supposons qu'il existe Z tel que $Y = X + Z$ et $i \in GF(q)$, inconnu, tel que $P(Z = i) = 1 - (q-1)p$ et $P(Z = j) = p$, $j \neq i$, avec $p \in \mathbb{P} \subset [0, 1/q]$ également inconnu.

Originalité – Pour un paramètre p fixé, l'entropie de Z est indépendante de i. Cependant, l'incertitude sur p et la méconnaissance de i rend inefficaces les schémas classiques de codage avec information adjacente au décodeur. Cet article propose un schéma de codage reposant sur des codes LDPC non binaires. Le décodeur exploite un algorithme EM pour estimer conjointement p, i et les symboles émis par la source. Un estimateur permettant d'initialiser efficacement l'algorithme EM est également proposé.

Résultats – Nous montrons qu'un estimateur au sens du maximum de vraisemblance naïf ne permet pas d'obtenir i et p correctement. L'efficacité du schéma proposé est comparée avec celle de techniques classiques, montrant sa supériorité tant en termes d'erreurs d'estimation de X qu'en termes de temps de calcul.

1 Introduction

Dans les schémas de codage d'une source X avec information adjacente Y disponible uniquement au décodeur, la distribution conditionnelle $P(Y|X)$ est, en général, supposée parfaitement connue. Ceci a permis de concevoir des solutions de codage reposant sur des codes de canal [6], comme les codes LDPC [3, 4]. Cependant, la source, n'ayant pas accès à Y peut avoir des difficultés à déterminer $P(Y|X)$. Ainsi, dans un réseau de capteur, X représente la mesure d'un capteur, qui doit faire parvenir une version compressée de X en un point de collecte, et Y représente les mesures prélevées par les autres capteurs dont le point de collecte peut se servir pour retourver X : $P(Y|X)$ peut déprendre du temps, de la configuration des capteurs, *etc*.

Dans cet article, nous considèrons le codage de sources q-aires dans le cas où $P(Y|X)$ est mal connue au décodeur. Plus spécifiquement, nous supposons qu'il existe Z tel que $Y = X + Z$ et $i \in GF(q)$, inconnu, tel que $P(Z = i) = 1 - (q-1)p$ et $P(Z = j) = p$, $j \neq i$, avec $p \in \mathbb{P} \subset [0, 1/q]$ également inconnu. Ce type de modèle permet de tenir compte d'une incertitude sur le niveau de corrélation entre X et Y, par l'intermédiaire de $p \in \mathbb{P}$, mais également sur le type de corrélation. Dans le cas binaire, avec $\mathbb{P} = [0, 0.1]$, le fait de ne pas connaitre i se traduit par une méconnaissance de la probabilité de transition $p = P(Y = 1|X = 0) = P(Y = 0|X = 1)$, qui peut être dans $[0, 0.1]$ ou dans $[0.9, 1]$. Malgré cette incertitude, les cas $p = 0.1$ et $p = 0.9$, par exemple, donnent la même entropie, et donc théoriquement le même débit de codage. De plus, on peut démontrer facilement à partir de [5], qu'un code LDPC a la même performance pour une probabilité de transition p ou $1-p$ et les algorithmes de décodage proposés dans [3] sont tout à fait capables de prendre en compte le cas $p > 0.5$. En revanche, quand on ne sait pas *a priori* si la probabilité de transition est donnée par p ou $1 - p$, les décodeurs classiques sont incapables de reconstruire les informations provenant de la source X.

Cet article propose un schéma de codage reposant sur des codes LDPC non binaires, directement inspiré de [2]. Le décodeur exploite un algorithme EM pour estimer conjointement p, i et les symboles émis par la source. Nous montrons qu'un estimateur au sens du maximum de vraisemblance naïf ne permet pas de fournir une estimée de i et p satisfaisante pour initialiser l'algorithme EM correctement. L'efficacité du schéma proposé est comparée avec des techniques classiques, montrant sa supériorité tant en

termes d'erreurs d'estimation de X qu'en termes de temps de calcul.

2 Modèle considéré

Les variables aléatoires sont en lettres majuscules et leurs réalisations en minuscules. Les vecteurs sont en gras. Les sources X et Y génèrent des suites de variables aléatoires $\{X_n\}_{n=1}^{+\infty}$ et $\{Y_n\}_{n=1}^{+\infty}$ indépendantes, identiquement distribuées (iid) à valeurs dans GF(q), où l'addition et la soustraction sont notée $\oplus$ et $\ominus$. Le modèle de corrélation est décrit par $Y = X \oplus Z$, où Z est une source qui génère des variables aléatoires iid dans GF(q). X est distribuée uniformément, tandis que la distribution de Z est décrite par un vecteur de paramètres $\boldsymbol{\theta} = [\theta_0, \ldots, \theta_{q-1}]$, avec $\theta_k = P(Z = k)$. La valeur prise par $\boldsymbol{\theta}$ n'est connue ni du codeur ni du décodeur et $\boldsymbol{\theta}$ peut varier d'une suite $\{Z_n\}_{n=1}^{+\infty}$ à une autre. Nous faisons l'hypothèse que $\boldsymbol{\theta}$ peut prendre q formes différentes : $\boldsymbol{\theta}^{(0)}(p) = [1-(q-1)p, \ldots, p]$, $\boldsymbol{\theta}^{(1)}(p) = [p, 1-(q-1)p, p, \ldots, p]$, *etc.*, avec $p \in \mathbb{P} \subset [0, 1/q]$, avec $\mathbb{P}$ connu. On note $\mathcal{P}_{\boldsymbol{\theta}}$ l'ensemble des vecteurs $\boldsymbol{\theta}$ possibles. Cet ensemble n'est pas connexe. L'index i de $\boldsymbol{\theta}^{(i)}(p)$ est appelé le *cadran* de $\boldsymbol{\theta}$, il permet de désigner la composante connexe de $\mathcal{P}_{\boldsymbol{\theta}}$ à laquelle $\boldsymbol{\theta}^{(i)}(p)$ appartient.

Pour ce modèle, le plus petit débit atteignable par un schéma de codage de source avec information adjacente au décodeur est donné par [1]

$$R = \sup_{\boldsymbol{\theta} \in \mathcal{P}_{\boldsymbol{\theta}}} H(X|Y, \boldsymbol{\theta}) \,. \tag{1}$$

le sup étant atteint pour q valeurs différentes de $\boldsymbol{\theta}$, de cadrant différent, mais de même p.

Dans [2], un schéma de codage est proposé pour le type de source considéré ci-dessus, mais dans le cas où $\mathcal{P}_{\boldsymbol{\theta}}$, n'est composé que d'une seule composante connexe (un seul cadran). Ce schéma ne peut être appliqué directement dans le cas où $\mathcal{P}_{\boldsymbol{\theta}}$ comporte plusieurs cadrans. En effet, le décodeur introduit dans [2] estime conjointement la source et le vecteur de paramètres $\boldsymbol{\theta}$, à l'aide d'un algorithme de type Expectation Maximization (EM). Ce dernier doit être correctement initialisé, ce qui a nécessité de proposer un estimateur initial de $\boldsymbol{\theta}$. Cependant, cet estimateur initial ne fonctionne pas lorsque $\mathcal{P}_{\boldsymbol{\theta}}$ comporte plusieurs cadrans, comme le montre le paragraphe suivant.

3 Codage, décodage et initialisation de l'algorithme EM

Le codeur proposé dans [2] réalise le codage du vecteur de source $\mathbf{x}$ de longueur n, à l'aide d'une matrice LDPC non-binaire H de taille $n \times m$, dimensionnée en utilisant (1). Le décodeur dispose du mot de code $\mathbf{u} = H^T\mathbf{x}$ et d'un vecteur d'information adjacente $\mathbf{y}$ de longueur n, avec lesquels il estime conjointement $\mathbf{x}$ et $\boldsymbol{\theta}$ à l'aide d'un algorithme EM. Cet algorithme itératif qui produit à l'itération j des estimées $\hat{\boldsymbol{\theta}}^{(j)}$ et $\hat{\mathbf{x}}^{(j)}$ de $\boldsymbol{\theta}$ et $\mathbf{x}$. Pour cela, l'algorithme EM a besoin de $P(X_k = i|y_k, \hat{\boldsymbol{\theta}}^{(j-1)})$, pour $i = 0 \ldots q-1$. Ces quantités sont fournies par un algorithme de décodage LDPC de type somme-produit initialisé avec $\hat{\boldsymbol{\theta}}^{(j-1)}$.

Une estimée $\hat{\boldsymbol{\theta}}^{(0)}$ raisonnable pour initialiser l'algorithme EM est obtenue par le décodeur à l'aide de $\mathbf{s} = H^T\mathbf{x} - H^T\mathbf{y} = H^T\mathbf{z}$. On fait ensuite l'hypothèse [1] que les variables aléatoires intervenant dans le calcul des composantes s_j de $\mathbf{s}$ sont distinctes et indépendantes de celles intervenant dans le calcul de s_k, $j \neq k$.

A partir de cette hypothèse, on effectue une estimation au sens du Maximum de Vraisemblance (MV) de $\boldsymbol{\theta}$ à partir de $\mathbf{s}$. La log-vraisemblance $L_i(p)$ en fonction du cadran i et de p est

$$L_i(p) = \sum_{m=1}^{M} \log \mathcal{F}_{u_m}^{-1} \left(\prod_{j=1}^{d_c} \mathcal{F}(W[H_{j,m}]\boldsymbol{\theta}^{(i)}(p)) \right) , \tag{2}$$

où $\mathcal{F}$ et $\mathcal{F}^{-1}$ sont les transformées de Fourier directe et inverse associées au calcul de la probabilité d'une somme de variables aléatoires à valeurs dans GF(q) ; $W[a]$ est une matrice de dimension $q \times q$ telle que $W[a]_{k,n} = \delta(a \otimes n \ominus k), 0 \leq k, n \leq q-1$.

Une estimée $\hat{\boldsymbol{\theta}}$ de $\boldsymbol{\theta}$ est obtenue en maximiser $L_i(p)$ par rapport à i et à p. Pour cela, on peut produire une estimée $\hat{p}_i = \arg\max_p L_i(p)$ par cadran i et conserver le couple $(i, \hat{p}_i)$ maximisant $L_i(\hat{p}_i)$. Cependant, nous avons vérifié expérimentalement que cette méthode ne produit des estimées satisfaisantes de i et de p que lorsque $i = 0$ est le vrai cadran.

Ce résultat peut être interprété de la manière suivante. Lorsque le cadran est $i = 0$, $\mathbf{z}$ contient une majorité de composantes nulles et il en est de même pour $\mathbf{s}$. En fait, la probabilité qu'une composante de $\mathbf{s}$ soit nulle varie fortement avec la probabilité $1-(1-q)*p$ qu'une composante de $\mathbf{z}$ soit nulle, voir la figure 1. Lorsque le cadran est $i \neq 0$, $\mathbf{z}$ contient beaucoup d'éléments non nuls. Ces éléments sont combinés aléatoirement via la matrice H pour donner des composantes de $\mathbf{s}$ dont les valeurs sont distribuées plus ou moins uniformément sur GF(q), avec une distribution dépendant très peu de p, voir la figure 1.

1. Cette hypothèse est fausse, puisqu'une variable X_n peut intervenir dans plusieurs sommes, mais comme chaque X_n intervient seulement dans un petit nombre de sommes, elle semble raisonnable pour obtenir une initialisation grossière de $\boldsymbol{\theta}$.

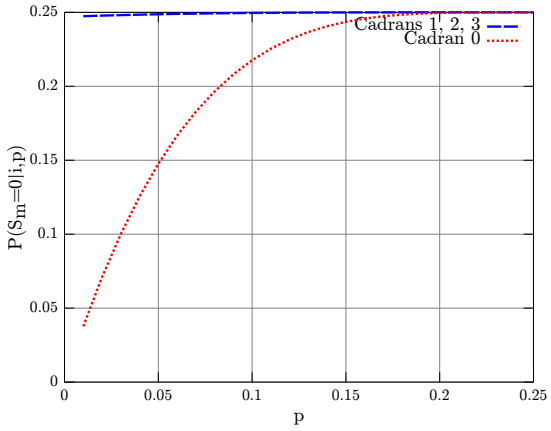

FIGURE 1 – $P(S_m = 0|i, p)$ en fonction du cadran i et de p, pour $q = 4$; les courbes pour $i = 1, 2, 3$ sont superposées

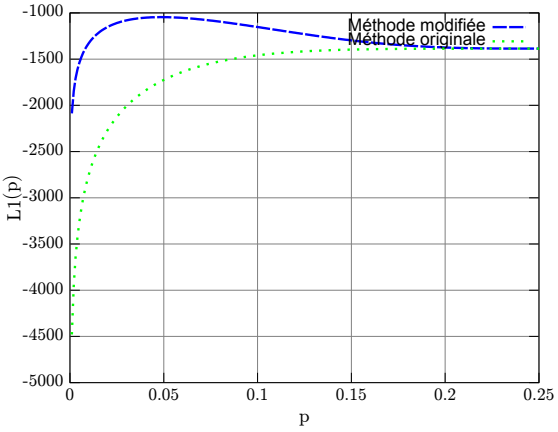

FIGURE 2 – $L_1(p)$ et $\tilde{L}_1(p)$ lorsque le vrai cadran est $i = 1$, calculées à l'aide d'un vecteur $\mathbf{s}$ de 1000, pour $p = 0.05$; $L_1(p)$ ne permet pas d'estimer p

Initialisation	P_e	Temps (s)
Aléatoire	7.2×10^{-3}	47.0
Méthode de [2]	7.5×10^{-1}	20.0
Méthode proposée	$< 10^{-5}$	9.1

TABLE 1 – Comparaison des initialisations possibles

Pour résoudre cette difficulté, nous proposons, pour chaque cadran i, de calculer $\tilde{\mathbf{s}}^{(i)} = \mathbf{s} \ominus H^T\mathbf{i} = H^T(\mathbf{z} \ominus \mathbf{i})$ où $\mathbf{i}$ est une vecteur de taille n contenant uniquement des i. Lorsque le vrai cadrant est i, $\mathbf{z}$ contient une majorité de i et $\mathbf{z} - \mathbf{i}$ contient une majorité de 0. Avec cette modification, la probabilité $P(\tilde{S}_m^{(i)} = 0|p, i)$ que la $m-$ème composante de $\mathbf{s}$ soit nulle exprimée en fonction de p pour le cadran i aura la même allure que $P(S_m = 0|p, i = 0)$, voir la Figure 1. Ainsi, p et i peuvent alors être estimés efficacement, en calculant $\hat{p}_i = \arg\max \tilde{L}_i(p)$, la log-vraisemblance (2) calculée pour $\tilde{\mathbf{s}}^{(i)}$, puis on conserve le couple $(i, \hat{p}_i)$ maximisant $\tilde{L}_i(\hat{p}_i)$.

La figure 2 montre sur un exemple où $i = 1$ que, contrairement à $L_1(p)$, $\tilde{L}_1(p)$ permet d'estimer efficacement p.

4 Simulations

Considèrons le modèle décrit au paragraphe 2, avec $q = 4$ et $\mathbb{P} = [0, 0.08]$. Quelle que soit la méthode de décodage considérée, 100 vecteurs de 1000 symboles sont générés, 20 itérations de décodeur LDPC sont réalisées et 3 itérations de l'algorithme EM. Pour chaque vecteur, le cadran i et et le paramètre p sont générés aléatoirement et uniformément. Un code LDPC de distributions de degrés $\lambda(x) = 0.413x + 0.375x^2 + 0.012x^4$ et $\rho(x) = x$ est utilisé, ce qui donne un débit de $R = 1.6$ bits/symbole.

Pour un premier jeu d'expériences, l'algorithme EM est initialisé aléatoirement. On suppose que le vrai cadran est le $i = 0$, on tire un p aléatoirement et uniformément dans $\mathbb{P}$ et on applique l'algorithme EM. S'il ne converge pas, on teste le cadran $i = 1$ et on effectue la même opération jusqu'à convergence. Dans le second jeu d'expériences, l'algorithme EM est initialisé à l'aide de l'argument du maximum de $L_i(p)$. Dans le troisième jeu d'expériences l'algorithme EM est initialisé avec l'argument du maximum de $\tilde{L}_i(p)$.

Le tableau 1 décrit la probabilité d'erreur de décodage de $\mathbf{x}$ ainsi que le temps nécessaire au décodage pour les trois jeux d'expériences, montrant l'importance d'une initialisation correcte de l'algorithme EM et les performances de la méthode que nous avons proposée.

Références

[1] I. Csiszar. Linear codes for sources and source networks : Error exponents, universal coding. *IEEE Transactions on Information Theory*, 28(4) :585–592, 1982.

[2] E. Dupraz, A. Roumy, and M. Kieffer. Practical coding scheme for universal source coding with side information at the decoder. In *Accepted to the Data Compression Conference, Snowbird*, March 2013.

[3] A. Liveris, Z. Xiong, and C. Georghiades. Compression of binary sources with side information at the decoder using LDPC codes. *IEEE Communications Letters*, 6 :440–442, 2002.

[4] T. Matsuta, T. Uyematsu, and R. Matsumoto. Universal Slepian-Wolf source codes using Low-Density Parity-Check matrices. In *IEEE International Symposium on Information Theory, Proceedings.*, pages 186–190, june 2010.

[5] T.J. Richardson, M.A. Shokrollahi, and R.L. Urbanke. Design of capacity-approaching irregular Low-Density Parity-Check codes. *IEEE Trans. on Inf. Th.*, 47(2) :619–637, 2001.

[6] Z. Xiong, A.D. Liveris, and S. Cheng. Distributed source coding for sensor networks. *IEEE Signal Processing Magazine*, 21(5) :80–94, Sep 2004.