Codage de source sous contrainte de sécurité avec information adjacente aux récepteurs

Joffrey VILLARD*, Pablo PIANTANIDA

SUPELEC, Dpt. Télécommunications, Gif-sur-Yvette, France.
joffrey.villard@supelec.fr, pablo.piantanida@supelec.fr

Résumé – Cet article traite du codage de source sous contrainte de sécurité avec information adjacente aux récepteurs. Un nœud (Alice) souhaite compresser et transmettre une source à un autre nœud (Bob), qui possède une information adjacente, corrélée à la source. La communication entre Alice et Bob est réalisée à travers un lien sans erreur à débit limité. Un nœud espion (Ève), qui possède également une information adjacente, corrélée à la source, écoute (parfaitement) ce lien. Dans ce contexte, Alice souhaite à la fois permettre à Bob d'estimer la source avec une faible distorsion, et révéler le minimum d'information à Ève. Une caractérisation complète de la région débit-distorsion-incertitude est fournie dans le cas d'informations adjacentes arbitrairement corrélées. Quelques cas particuliers et un exemple dans le cas de sources binaires sont également étudiés. Il est montré que les différences statistiques entre les informations adjacentes ainsi que la distorsion tolérée peuvent être pleinement utilisées pour accroître la sécurité.

Abstract – This paper investigates the problem of lossy source coding under security constraint in the presence of side information at the receivers. This scenario consists of an encoder (referred to as Alice) that wishes to compress and send a source to a legitimate receiver (Bob), who has also access to a correlated source that can be used as side information. An error-free rate-limited link is assumed between Alice and Bob. An eavesdropper (Eve) perfectly observes the information bits sent by Alice to Bob and has also access to a correlated source. In this context, Alice wishes to simultaneously satisfy the desired requirements on: (i) the distortion level at Bob and (ii) the equivocation rate at Eve. A complete characterization of the rate-distortion-equivocation region for the case of arbitrarily correlated side informations is derived. Several special cases of interest and an application example to secure lossy source coding of binary sources are also considered. It is shown that the statistical differences between the side informations and the presence of non-zero distortion at the legitimate receiver can be useful in terms of secrecy.

Notations – Pour toute suite $(x_i)_{i\in\mathbb{N}^*}$, on note x^n la collection $(x_1, x_2, \ldots, x_n)$. Le cardinal d'un ensemble fini est noté $\|\cdot\|$. L'entropie est noté $H(\cdot)$ et l'information mutuelle $I(\cdot;\cdot)$. Soient X, Y et Z trois variables aléatoires de loi p. Si $p(x|y,z) = p(x|y)$ pour tous x, y, z, alors elles forment une chaîne de Markov, notée $X \multimap Y \multimap Z$. Pour tout $x \in \mathbb{R}$, $[x]_+$ signifie $\max(0; x)$.

1 Introduction

Cet article traite du codage de source sécurisé avec information adjacente aux récepteurs. Un nœud (Alice) souhaite compresser et transmettre son observation A à un autre nœud (Bob), qui possède sa propre observation B, corrélée à A. La communication entre Alice et Bob est réalisée à travers un lien sans erreur à débit limité. Un nœud espion (Ève) écoute le lien et possède également une information adjacente E. Dans ce contexte, Alice souhaite à la fois permettre à Bob d'estimer A avec une faible distorsion et révéler le minimum d'information à Ève (voir Figure 1). Notre but est d'étudier la région débit-distorsion-incertitude correspondante.

Le scénario étudié dans cet article implique un grand nombre de problématiques de la théorie de l'information. En termes de codage de source, Slepian et Wolf [1], et Wyner et Ziv [2] ont introduit le problème de codage de source avec information adjacente au décodeur, problème ayant reçu par la suite une attention considérable, conduisant à des progrès remarquables sur les plans théorique et pratique. D'autre part, les communications sécurisées sur canaux bruités ont été l'objet de nombreux travaux ces dernières années, depuis le lien sur écoute de Wyner [3], adoptant le point de vue défini par Shannon [4] qui propose de mesurer le niveau de sécurité via l'incertitude résiduelle à Ève à propos du message. Peu de travaux portent sur le codage de source avec contrainte de sécurité. Lorsqu'une information adjacente est disponible aux récepteurs, le cas où Bob doit parfaitement reconstruire la source A a récemment été étudié [5–7]. Cependant, le cas général d'une estimation avec perte lorsque les informations adjacentes sont corrélées de manière arbitraire n'a pas été résolu. C'est le problème que nous abordons ici.

*Les travaux de J. Villard sont soutenus par la DGA.

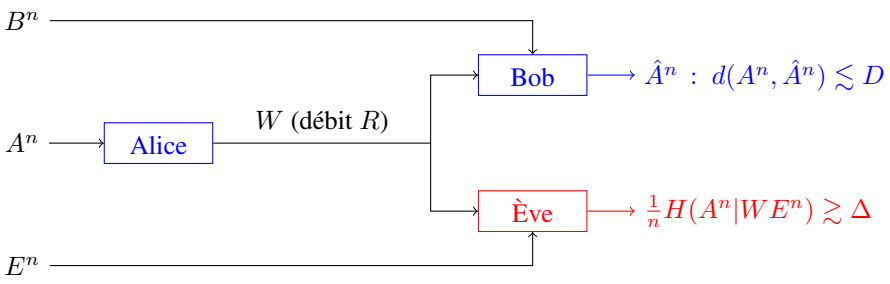

FIGURE 1 – Codage de source sécurisé avec information adjacente aux récepteurs.

Dans cet article, nous fournissons une caractérisation complète de la région débit-distorsion-incertitude dans le cas d'informations adjacentes arbitrairement corrélées. Quelques cas particuliers sont également étudiés. Un exemple est fourni dans le cas d'une source binaire avec information adjacente à Bob (resp. Ève) produite à partir de la source par un canal binaire à effacement (resp. un canal binaire symétrique). L'intérêt de ce modèle réside dans le fait que Bob (ou Ève) n'est pas toujours un décodeur moins bruité.

2 Définitions et résultat général

2.1 Définitions

Dans cette section, nous donnons une définition plus rigoureuse du contexte représenté à la Figure 1. Soient $\mathcal{A}$, $\mathcal{B}$ et $\mathcal{E}$ trois ensembles finis. Alice, Bob et Ève observent les suites de variables aléatoires $(A_i)_{i\in\mathbb{N}^*}$, $(B_i)_{i\in\mathbb{N}^*}$ et $(E_i)_{i\in\mathbb{N}^*}$ respectivement, à valeurs dans $\mathcal{A}$, $\mathcal{B}$ et $\mathcal{E}$, resp. Pour tout $i \in \mathbb{N}^*$, les variables aléatoires A_i, B_i et E_i sont distribuées selon la probabilité jointe $p(a, b, e)$ sur $\mathcal{A} \times \mathcal{B} \times \mathcal{E}$. Elles sont de plus indépendantes en fonction du temps i. Soit $d : \mathcal{A} \times \mathcal{A} \to [0\,; d_{max}]$ une mesure de distorsion, telle que $0 \leq d_{max} < \infty$. On note également d la distorsion moyenne sur toutes les composantes : pour tous $a^n, b^n \in \mathcal{A}^n$, $d(a^n, b^n) = \frac{1}{n} \sum_{i=1}^n d(a_i, b_i)$.

Définition 1 *Dans ce contexte, un code (n, R) pour le codage de source est défini par*
- *Une fonction d'encodage à Alice $f : \mathcal{A}^n \to \{1, \ldots, 2^{nR}\}$,*
- *Une fonction de décodage à Bob*
$$g : \{1, \ldots, 2^{nR}\} \times \mathcal{B}^n \to \mathcal{A}^n.$$

Définition 2 *Un triplet $(R, D, \Delta) \in \mathbb{R}_+^3$ est dit* atteignable*, si pour tout $\epsilon > 0$ il existe un code $(n, R + \epsilon)$ (f, g) tel que :*

$$\mathbb{E}\big[d(A^n, g(f(A^n), B^n))\big] \leq D + \epsilon\,,$$
$$\frac{1}{n} H(A^n|f(A^n), E^n) \geq \Delta - \epsilon\,.$$

L'ensemble des triplets atteignables, appelé région débit-distorsion-incertitude*, est noté $\mathcal{R}^*$.*

2.2 Résultat

Le théorème suivant fournit une caractérisation de la région débit-distorsion-incertitude. La démonstration de l'atteignabilité repose sur un schéma de codage en deux couches superposées (*superposition coding*), utilisant les techniques de Wyner et Ziv [2] (*random binning*), et une évaluation de l'incertitude à Ève basée sur les propriétés des séquences typiques. La réciproque est démontrée en utilisant les techniques classiques et l'identité de Csiszàr et Körner [8]. Voir [9, 10] pour une démonstration détaillée.

Théorème 1 ([9,10]) *La région $\mathcal{R}^*$ est l'ensemble des triplets $(R, D, \Delta) \in \mathbb{R}_+^3$ tels qu'il existe des variables aléatoires U, V sur des ensembles finis $\mathcal{U}$, $\mathcal{V}$, resp., et une fonction $\hat{A} : \mathcal{V} \times \mathcal{B} \to \mathcal{A}$, telles que $U \multimap V \multimap A \multimap (B, E)$ forment une chaîne de Markov et*

$$R \geq I(V; A|B)\,, \tag{1}$$
$$D \geq \mathbb{E}\big[d(A, \hat{A}(V, B))\big]\,, \tag{2}$$
$$\Delta \leq H(A|VB) + I(A; B|U) - I(A; E|U)\,. \tag{3}$$

Les inégalités (1) et (2) sont classiques dans la théorie débit-distorsion [2]. La variable V est la description de la source A permettant d'estimer celle-ci sous la contrainte de distorsion.

L'inégalité (3) peut être interprétée de la façon suivante. Le premier terme $H(A|VB)$ correspond à l'incertitude à Bob. Alice exploite donc la distorsion tolérée à Bob pour augmenter l'incertitude à Ève. De plus, pour une variable V donnée, qui détermine le débit R et le niveau de distorsion D, la variable aléatoire U peut être choisie pour rendre Bob *meilleur* que Ève, c'est-à-dire maximiser $I(A; B|U) - I(A; E|U)$. Cette quantité représente le gain (ou la perte) en termes d'incertitude à Ève (par rapport à Bob).

D'autre part, cette inégalité s'écrit également

$$\Delta \leq H(A|UE) - I(V; A|UB)\,.$$

La variable U (une « sous-partie » de V) est donc considérée comme un *message commun*, c-à-d comme si Ève pouvait le décoder. La quantité d'information restante $I(V; A|UB)$ (transmise à Bob dans un deuxième temps pour estimer A à partir de B et U sous la contrainte de distortion) est directement soustraite de l'incertitude. Cela signifie qu'elle est traitée comme des bits *bruts*.

La proposition suivante permet de borner la taille des alphabets $\mathcal{U}$ et $\mathcal{V}$. La démonstration, basée sur le théorème de Fenchel-Eggleston-Carathéodory, est omise ici. Elle peut être trouvée dans [9].

Proposition 1 *Dans la caractérisation de la région débit-distorsion-incertitude $\mathcal{R}^*$ donnée par le Théorème 1, il suffit de considérer les ensembles $\mathcal{U}$ et $\mathcal{V}$ vérifiant $\|\mathcal{U}\| \leq \|\mathcal{A}\| + 2$ et $\|\mathcal{V}\| \leq (\|\mathcal{A}\| + 2)(\|\mathcal{A}\| + 1)$.*

3 Cas particuliers

3.1 Codage de source sans perte

Le cas du codage de source sans perte correspond à un niveau de distorsion nul à Bob ($D = 0$). Dans ce cas, le corollaire suivant (qui peut également être trouvé dans [5, 6]), est une conséquence directe du Théorème 1 (avec $V = A$) :

Corollaire 1 *Un triplet $(R, D = 0, \Delta)$ est atteignable si et seulement si il existe une variable aléatoire U sur un ensemble fini $\mathcal{U}$, telle que $U \multimap A \multimap (B, E)$ forment une chaîne de Markov et*

$$\begin{aligned} R &\geq H(A|B) \,, \\ \Delta &\leq I(A;B|U) - I(A;E|U) \,. \end{aligned}$$

3.2 L'information adjacente de Bob est moins bruitée que celle de Ève

L'information adjacente B est dite *moins bruitée* que E si $I(U;B) \geq I(U;E)$ pour toute variable aléatoire U telle que $U \multimap A \multimap (B, E)$ forment une chaîne de Markov. On note cette relation $B \succeq_A E$. Notons que cette condition est strictement plus faible que la condition *stochastiquement dégradée*. Le corollaire suivant indique que, dans ce cas, le codage de Wyner-Ziv [2] atteint les performances optimales en termes de sécurité (la variable auxiliaire U du Théorème 1 est choisie constante).

Corollaire 2 *Si $B \succeq_A E$, alors la région $\mathcal{R}^*$ est l'ensemble des triplets (R, D, Δ) tels qu'il existe une variable aléatoire V sur un ensemble fini $\mathcal{V}$, et une fonction $\hat{A} : \mathcal{V} \times \mathcal{B} \to \mathcal{A}$ telles que $V \multimap A \multimap (B, E)$ forment une chaîne de Markov et*

$$\begin{aligned} R &\geq I(V;A|B) \,, \\ D &\geq \mathbb{E}\big[d(A, \hat{A}(V,B))\big] \,, \\ \Delta &\leq H(A|VB) + I(A;B) - I(A;E) \,. \end{aligned}$$

4 Caractérisation alternative

La proposition suivante est une conséquence du Théorème 1.

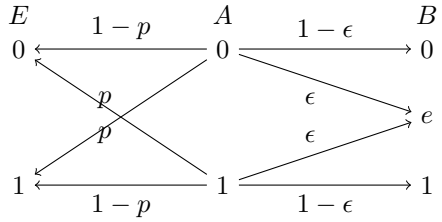

FIGURE 2 – Modèle de source et d'informations adjacentes.

Proposition 2 *La région $\mathcal{R}^*$ est l'ensemble des triplets $(R, D, \Delta) \in \mathbb{R}_+^3$ tels qu'il existe des variables alétoires U, V sur des ensembles finis $\mathcal{U}$, $\mathcal{V}$, resp., et une fonction $\hat{A} : \mathcal{V} \times \mathcal{B} \to \mathcal{A}$, telles que $U \multimap V \multimap A \multimap (B, E)$ forment une chaîne de Markov et*

$$R \geq \big[I(U;B) - I(U;E)\big]_+ + I(V;A|B) \,, \tag{4}$$

$$D \geq \mathbb{E}\big[d(A, \hat{A}(V,B))\big] \,, \tag{5}$$

$$\Delta \leq H(A|VB) + I(A;B|U) - I(A;E|U) \,. \tag{6}$$

Démonstration. Les inégalités (4)–(6) décrivent une région plus petite que (1)–(3). L'atteignabilité de la Proposition 2 est donc une conséquence de celle du Théorème 1.

Le membre de droite de (3) et (6) s'écrit

$$H(A|VB) + I(A;B) - I(A;E) - \Big[I(U;B) - I(U;E)\Big] \,.$$

Maximiser ce terme par rapport à U revient donc à minimiser $I(U;B) - I(U;E)$. Dans le pire des cas, $U = \emptyset$ annule cette différence. Le choix optimal U^* impliquera nécessairement $I(U^*;B) - I(U^*;E) \leq 0$, rendant (1) et (4) identiques. □

La Proposition 2 et la démonstration ci-dessus indiquent que le choix optimal de Uest une variable aléatoire U^* qui peut être décodée par Ève. Cependant, puisqu'il minimise la quantité $I(U;B) - I(U;E)$ par rapport U, ce *message commun* devrait donner peu d'information à Ève. Ce résultat est en accord avec ceux sur les canaux de diffusion avec messages confidentiels [8], pour lesquels le message commun peut être également donné à l'espion sans réduire la région atteignable.

5 Exemple : source binaire avec informations adjacentes CBE et CBS

On considère le modèle de source représenté à la Figure 2 où la source est binaire et les informations adjacentes à Bob et Ève sont les sorties respectives d'un canal binaire à effacement (CBE) de probabilité d'effacement $\epsilon \in [0, 1/2]$ et d'un canal binaire symétrique (CBS) de probabilité d'erreur $p \in [0, 1/2]$, d'entrée A. Ce modèle présente un intérêt car aucun décodeur (ni Bob, ni Ève) est un décodeur moins bruité de l'autre pour toutes valeurs de (p, ϵ). En fait, en fonction des paramètres

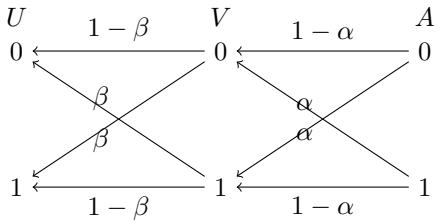

FIGURE 3 – Variables auxiliaires binaires.

(p, ϵ), on peut montrer [11] que les informations adjacentes vérifient les propriétés [1] résumées à la Figure 4. Le Corollaire 2 donne donc la région débit-distorsion-incertitude lorsque $\epsilon \leq 4p(1-p)$.

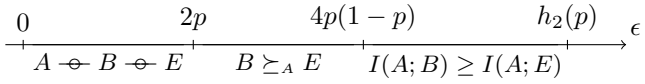

FIGURE 4 – Propriétés des informations adjacentes.

Supposons à partir de maintenant que la fonction de distorsion d est la distance de Hamming et que la source est uniformément distribuée ($\Pr\{A=0\} = \Pr\{A=1\} = 1/2$). En utilisant des technique classiques, on peut montrer qu'il suffit dans ce cas de considérer des variables auxiliaires U et V obtenues en sortie de deux CBS en cascade avec A en entrée (voir Figure 3).

Nous pouvons alors calculer numériquement quelques triplets atteignables pour $p = 0.1$ et $\varepsilon = h_2(p) \approx 0.469$. Dans le cas d'une compression sans perte (colonnes 1 et 2 de la Table 1), $V = A$, et la variable U permet d'atteindre une incertitude non nulle à Ève. Si le débit est limité à 80% du débit requis pour une reconstruction parfaite (colonne 3), une distorsion de 1.5% est présente à Bob et on peut atteindre une incertitude à Ève de 0.133 bits. Cela signifie qu'un faible niveau de distorsion à Bob peut être pleinement exploité par Alice pour obtenir des gains significatifs en termes de sécurité.

Notons qu'un niveau de distorsion atteignable D fournit ici une borne supérieure du taux d'erreur binaire (TEB) à Bob (pour l'estimation de A) :

$$\mathbb{E}\big[d(A^n, g(f(A^n), C^n))\big] = \frac{1}{n}\sum_{i=1}^{n} \Pr\left\{\hat{A}_i \neq A_i\right\} .$$

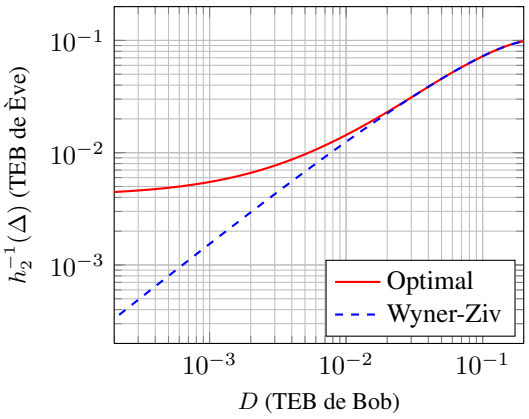

FIGURE 5 – Bornes sur les TEB de Bob et Ève ($p = 0.1$, $\epsilon = h_2(p) \approx 0.469$).

Dans le même temps, un niveau d'incertitude atteignable Δ constitue une borne inférieure du TEB à Ève :

$$\frac{1}{n}H(A^n|JE^n) \leq h_2\left(\frac{1}{n}\sum_{i=1}^{n} \Pr\left\{\breve{A}_i \neq A_i\right\}\right) .$$

La Figure 5 représente la borne inférieure du TEB à Ève $h_2^{-1}(\Delta)$ en fonction du TEB à Bob D pour $p = 0.1$ et $\varepsilon = h_2(p) \approx 0.469$. En particulier, le TEB à Bob peut tendre vers zéro alors que celui à Ève reste supérieur à $4 \cdot 10^{-3}$.

Cette relation entre incertitude et distorsion est cependant spécifique au cas binaire avec distance de Hamming (une autre existe aussi dans le cas gaussien avec erreur quadratique). Dans le cas général, le Théorème 1 ne permet pas de garantir une borne inférieure pour la distorsion de Ève. Un tel problème avec une contrainte de distorsion sur l'espion aurait certainement du sens, mais sa démonstration, en particulier la réciproque, apparaît plus difficile à obtenir.

1. $h_2(x) = -x\log_2(x) - (1-x)\log_2(1-x)$ est l'entropie binaire.

TABLE 1 – Quelques triplets atteignables ($p = 0.1$, $\epsilon = h_2(p) \approx 0.469$).

	Codage proposé	Slepian-Wolf	Codage proposé	Wyner-Ziv
Débit R	0.469	0.469	0.375	0.375
Distorsion D	0	0	0.015	0.015
Incertitude Δ	0.039	0	0.133	0.126

Références

[1] D. Slepian and J. Wolf. Noiseless coding of correlated information sources. *IEEE Trans. Inf. Theory*, 19(4) :471–480, 1973.

[2] A.D. Wyner and J. Ziv. The rate-distortion function for source coding with side information at the decoder. *IEEE Trans. Inf. Theory*, 22(1) :1–10, 1976.

[3] A.D. Wyner. The wire-tap channel. *BSTJ*, 54(8) :1355–1387, 1975.

[4] C.E. Shannon. Communication theory of secrecy systems. *BSTJ*, 28 :656–715, 1949.

[5] V. Prabhakaran and K. Ramchandran. On secure distributed source coding. In *Proc. ITW*, pages 442–447, 2007.

[6] D. Gunduz, E. Erkip, and H.V. Poor. Secure lossless compression with side information. In *Proc. ITW*, 2008.

[7] R. Tandon, S. Ulukus, and K. Ramchandran. Secure source coding with a helper. In *Proc. Allerton*, 2009.

[8] I. Csiszar and J. Korner. Broadcast channels with confidential messages. *IEEE Trans. Inf. Theory*, 24(3) :339–348, 1978.

[9] J. Villard and P. Piantanida. Secure lossy source coding with side information at the decoders. In *Proc. Allerton*, 2010.

[10] J. Villard and P. Piantanida. Secure multiterminal source coding with side information at the eavesdropper. *arXiv cs.IT*, 1105.1658 :1–65, 2011.

[11] C. Nair. Capacity regions of two new classes of 2-receiver broadcast channels. In *Proc. ISIT*, pages 1839–1843, 2009.