Classification Distribuée de Données avec la Méthode "Information Bottleneck"

Abdellatif ZAIDI[1,2]

[1] Université Paris-Est, Champs-sur-Marne, 77454, France

[2] France Research Center, Huawei Technologies, Boulogne-Billancourt, 92100, France

abdellatif.zaidi@univ-mlv.fr

Résumé – Dans ce article, nous étudions le problème de compression distribuée de sources avec information adjacente dans le cas où la mesure de la distortion est de type *logarithmic-loss*. Cette fonction de pénalité, par ailleurs très utilisé dans la théorie d'apprentissage et prédiction ainsi que certains travaux de la littérature de traitement d'images, permet d'établir des liens importants avec des domaines à-priori différents, telque celui de classification de données. Nous donnons une charactérization complète de la région taux de compression/distortion du modèle de compression distribuée de données étudié et nous appliquons les résultats obtenus au problème de classification distribuée de données via la méthode d'*Information Bottleneck*.

Abstract – In this work, we study the problem of multiterminal source coding with side information under logarithmic loss distortion measure. The logarithmic loss function, which is also widely used in the theory of learning and prediction as well as the image processing literature establishes important connections with problems that are seemingly different, such as that of data clustering. We establish a single-letter characterization of the rate-distortion region of the studied distributed source coding model and then apply the results to the problem of distributed clustering of data via distributed Information Bottleneck.

1 Introduction

Nous considérons le problème représenté par Figure 1. Dans ce modèle, deux source discrètes sans mémoire qui sont arbitrairement corrélées doivent être compréssées de façon séparée, et reconstruites de façon conjointe au niveau du décodeur. En plus, le décodeur a accès à une information adjacente (SI) qui, elle aussi, est arbitrairement corrélée aux sources. Pour des sources avec des alphabets et mesures de distortions généraux, le probème est encore ouvert depuis un peu plus d'une trentaine d'années. Une solution du problème, en termes de caractérisation de la région taux de compression/distortion est connue seulement dans quelques cas particuliers, parmi lesquels le cas où les sources sont indépendantes conditionnellement à l'information disponible au décodeur, ¸à-d, $X_1 \multimap Y \multimap X_2$ est une chaine de Markov [1, Theorem 6]. En réalité, pour des sources et mesures de distortions générales, la solution du problème est à nos jours inconnue même dans le cas où le décodeur n'observe aucune information adjacente, ç-à-d., $Y = \emptyset$. Dans ce dernier cas, ¸à-d, quand $Y = \emptyset$, quelques cas importants ont déjà été résolus, parmi lesquels le cas où une des deux sources n'est pas à reconstruite [2], et le cas où une des deux sources est encodé de faparfaitte et envoyée au décodeur [3]. Dans le cas gaussien sans mémoire, les cas de codage distribuée avec mesure quadratique de distortion et celui du 'Chief Executive Officer problem' (CEO) avec mesure quadratique de distortion ont, eux aussi, déjà été résolus.

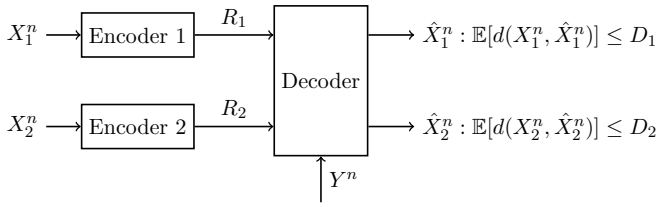

FIGURE 1 – Codage distribué de sources avec information adjacente au décodeur.

Récémment, la région optimale de taux de compression distortion a été caractérisée pour un modèle de codage distribué de sources avec deux encodeurs et mesure de distortion qui est de type 'logarithmic loss' [4, Theorem 6]. Ce résultat est assez important car il s'applique à toutes les sources qui ont un alphabet fini. Cette mesure de distortion est telle que le decodeur produit une estimation 'softe' de la source, ç-à-d., une distribution de probabilité au lieu de valeur déterministe dans le cas classique. Plus spécifiquement, $\hat{x}_i$, $i = 1, 2$, est une distribution de probabilité qui est définie sur l'alphabet $\mathcal{X}_i$ de X_i et $d(x_i, \hat{x}_i)$ est l'entropie relative (¸à-d, divergence de Kullback-Leibler) entre la distribution empirique de l'évènement $\{X_i = x_i\}$ et l'estimé $\hat{x}_i$. En utilisant une propriété importante de cette mesure de distortion log-loss, que les auteirs démontrent et qui consiste en le fait que toute valeur de distortion qui est atteignable est au moins aussi large qu'une entropie conditionnelle,

ils caracterisent la région taux de compression/distortion du problème CEO [4, Theorem 3] and celle du problème de codage distribué dans le cas de deux encodeurs [4, Theorem 6]. Il est important de noter qu'alors que le résultat du CEO s'etend bien au cas de $m \geq 3$ encodeurs, à condition que les obervations aux encodeurs soient indépendantes conditionnellement à la source cachée, celui du codage distribué ne s'etend pas facilement au cas d'encodeurs multiples, avec $m \geq 3$. En réalité le codage Berger-Tung, qui est optimal dans le cas de deux encodeurs est strictement sous-optimal (exemple, problème de Korner-Marton de reconstitution de la somme de deux sources binaires).

Dans ce papier, nous étudions le problème représenté par Figure 1 dans le cas d'une mesure de distortion de type logarloss. Ce modèle généralise celui de [4] au cas où le décodeur possède, ou a accès, à une information adjacente Y^n qui est arbitrairement corrélée aux sources à compresser. Nous développons une caractérisation de la région taux de compression/distortion de ce problème dans le cas discret. A cet effet, nous montrons qu'une généralisation de la région atteignable de Gastpar [1, Theorem 2], obtenue en introduisant partage de temps (time sharing) est optimale. La preuve de la réciproque est inspirée par celle de [4, Theorem 12], que nous étendons au cas où le décodeur est informé. Aussi, cela requiert aussi une redéfinition appropriée des variables auxiliaires. Par ailleurs, en spécialisant notre résultat au cas où un seul encodeur communique avec le décodeur, ç-à-d., $R_1 = 0$, et le décodeur est intéressé en seulement la reconstruction de la source cachée, nous caractérisons le compromis optimal entre la complexité (complexity) et la précision (accuracy) de la description de X_2^n au décodeur. Cela constitue une généralisation de la méthode d'Information Bottleneck [5, 6] au cas avec SI au décodeur.

Finalement, nous notons que la mesure de distortion log-loss est très utile en pratique dans beaucoup de domaines, comme l'apprentissage statistique, l'éstimation, prédiction, traitement d'images et d'autres [5, 6]. Par exemple, la méthode d'Information Bottleneck a déjà étée appliquée à différents problèmes de classification de divers types de données, ainsi que pour la sécurité des données [7] et d'autres.

1.1 Notation

Les variables aléatoires et leurs réalisations sont représentées par des lettres majuscules (exp., X) et miniscules (exp., x), respectivement. Les lettres en calligraphiques dénotent les alphabets (exp., $\mathcal{X}$), et leurs cardinalités sont dénotés par $|\cdot|$ $(exp., |\mathcal{X}|)$. Nous abbprévions une séquence $(X_1, X_2, \ldots, X_n)$ par X^n, et pour $i \leq j$, nous dénotons $(X_i, \ldots, X_j)$ par X_i^j.

2 Modèle

Considérons une source discrète sans mémoire (X_1, X_2, Y) d'alphabet fini $\mathcal{X}_1 \times \mathcal{X}_2 \times \mathcal{Y}$ et de distribution jointe (PMF) $P_{X_1,X_2,Y}(x_1, x_2, y)$. Soit $\{(X_{1,i}, X_{2,i}, Y_i)\}_{i=1}^n$ une séquence de n independent and identiquement distributés (i.i.d.) triplets de variables aléatoires de PMF jointe $P_{X_1,X_2,Y}(x_1, x_2, y)$, ç-à-d, $(X_1^n, X_2^n, Y^n) \sim \prod_{i=1}^n P_{X_1,X_2,Y}(x_{1,i}, x_{2,i}, y_i)$. Considérons maitenant le problème représenté par Figure 1. Dans ce modèle, Encodeur 1 observe la source X_1^n and utilise R_1 bits par symbole afin de la décrire au décodeur. De façon similaire, Encodeur 2 observe la source X_2^n et utilise R_2 bits par symbole afin de la décrire au décodeur. Le décodeur observe une séquence d'information adjacente Y^n qui est statistiquement dependente des sources.

De façon analogue à [4], l'alphabet de reproduction $\hat{\mathcal{X}}_i$, $i = 1, 2$, est pris égal à l'ensemble des PMF qui sont définies sur l'alphabet de la source $\mathcal{X}_i$. Par conséquence, pour un vecteur $\hat{X}_i^n \in \hat{\mathcal{X}}_i^n$, la notation $\hat{X}_{i,j}(x_i)$ dénote la j^{eme}-coordonnée de $\hat{X}_i^n$, $1 \leq j \leq n$, qui est une distribution de probabilité sur $\mathcal{X}_i$, évaluée en $x_i \in \mathcal{X}_i$. Nous considérons la mesure de distortion log-loss définie par

$$d(x_i, \hat{x}_i) = \log\left(\frac{1}{\hat{x}_i(x_i)}\right). \tag{1}$$

De façon équivalente, $d(x_i, \hat{x}_i)$ est l'entropie relative (ç-à-d., divergence de Kullback-Leibler) entre la distribution empirique de l'évenement $\{X_i = x_i\}$ et l'estimé $\hat{x}_i$. La distortion entre les séquences est donnée par

$$d(x_i^n, \hat{x}_i^n) = \frac{1}{n}\sum_{i=1}^n d(x_i, \hat{x}_i) \quad \text{for} \quad i = 1, 2. \tag{2}$$

Définition 1. *Un code (de longuer n) pour le modèle de la Figure 1 consiste en deux fonstions d'encodage*

$$\phi_i^{(n)} \ : \ \mathcal{X}_i^n \to \{1, \ldots, M_i^{(n)}\} \quad \textit{pour} \ i = 1, 2 \tag{3}$$

et fonctions de décodage

$$\psi_i^{(n)} : \{1, \ldots, M_1^{(n)}\} \times \{1, \ldots, M_2^{(n)}\} \times \mathcal{Y}^n \to \hat{\mathcal{X}}_i^n \quad \textit{pour} i = 1, 2. \tag{4}$$

Définition 2. *Un vecteur de taux de compression/distortions (R_1, R_2, D_1, D_2) est atteignable pour le modèle de la Figure 1 s'il existe n, deux fonctions d'encodage $\phi_1^{(n)}$ et $\phi_2^{(n)}$ et fonctions de décodage $\psi_1^{(n)}$ et $\psi_2^{(n)}$ telles que*

$$R_i \geq \frac{1}{n}\log M_i^{(n)} \quad \textit{pour} \ i = 1, 2 \tag{5}$$

$$D_i \geq \mathbb{E}[d(X_i^n, \hat{X}_i^n)] \quad \textit{pour} \ i = 1, 2 \tag{6}$$

où

$$\hat{X}_i^n = \psi_i^{(n)}\left(\phi_1^{(n)}(X_1^n), \phi_2^{(n)}(X_2^n), Y^n\right) \quad \textit{pour} \, i = 1, 2. \tag{7}$$

La région taux de compressions/distortions $\mathcal{RD}^\star$ du modèle de la Figure 1 est définie par l'enveloppe convexe de l'ensemble des quadriplets (R_1, R_2, D_1, D_2) qui sont atteignables.

3 Principaux Résultats

3.1 Rate-Distortion Region

Le principal résultat de ce papier est une caractérisation de la région de taux de compression/distortions $\mathcal{RD}^\star$ du modèle de Figure 1 sous une mesure de distortion qui est de type log-loss.

Théorème 1. *La région taux de compression/distortions $\mathcal{RD}^\star$ du modèle de Figure 1 sous mesure de distortion de tye log-loss est donnée par l'ensemble des quadriplets $(R_1, R_2, D_1, D_2$ qui vérifient*

$$R_1 \geq I(U_1; X_1|Y, U_2, Q) \tag{8}$$
$$R_2 \geq I(U_2; X_2|Y, U_1, Q) \tag{9}$$
$$R_1 + R_2 \geq I(U_1, U_2; X_1, X_2|Y, Q) \tag{10}$$
$$D_1 \geq H(X_1|Y, U_1, U_2, Q) \tag{11}$$
$$D_2 \geq H(X_2|Y, U_1, U_2, Q) \tag{12}$$

pour une PMF jointe de la forme

$$P_{X_1,X_2,Y}(x_1, x_2, y)P_Q(q) \times P_{U_1|X_1,Q}(u_1|x_1, q)P_{U_2|X_2,Q}(u_2|x_2, q). \tag{13}$$

Remarque 1. *Les variables auxiliaires du Théorème 1 sont tels que $U_1 \multimap (X_1, Q) \multimap (Y, X_2, U_2)$ et $U_2 \multimap (X_2, Q) \multimap (Y, X_1, U_1)$ forment des chaines de Markov.*

Remarque 2. *Le Théoreme 1 généralise celui de [4, Theorem 6] au cas où le décodeur observe une information adjacente Y^n qui est statistiquement dependente des sources observées par les encodeurs.*

Remarque 3. *Dans le cas où les sources X_1 et X_2 sont indépendantes conditionallement à Y, ç-à-d., $X_1 \multimap Y \multimap X_2$ est une chaine de Markov, on peut démontrer facilement que le résultat du Théorème 1 se réduit à l'ensemble des quadriplets (R_1, R_2, D_1, D_2) qui satisfont*

$$R_1 \geq I(U_1; X_1) - I(U_1; Y) \tag{14}$$
$$R_2 \geq I(U_2; X_2) - I(U_2; Y) \tag{15}$$

et

$$D_1 \geq H(X_1|Y, U_1) \tag{16}$$
$$D_2 \geq H(X_2|Y, U_2) \tag{17}$$

pour une mesure de la forme

$$P_{X_1,X_2,Y}(x_1, x_2, y)P_{U_1|X_1}(u_1|x_1)P_{U_2|X_2}(u_2|x_2). \tag{18}$$

Ce résultat peut aussi être obtenu à partir de [1, Theorem 6] avec les fonctions de reproduction choisies telles que

$$f_i(U_i, Y) := Pr[X_i = x_i|U_i, Y] \quad pour\ i = 1, 2. \tag{19}$$

Pour un tel choix, on a

$$\mathbb{E}[d(X_i, f_i(U_i, Y)] = H(X_i|U_i, Y) \quad pour\ i = 1, 2. \tag{20}$$

4 Application à la Classification des Données

Les résultats de la section précédente peuvent être appliqués au problème de classification (ou clusterin) des données de façon distribuée. Par exemple, considérons le cas où la source $< x_1^n$ est cachée ou non-observée. C'est à dire, seul le deuxième encodeur communique avec le décodeur qui souhaite obtenir un estimé de la source non-observée X_1^n. ¡cela veut dire que $U_1 = \emptyset$ et $D_2 = \infty$. Dans ce cas, avec les substitutions $R := R_2$, $U_2 := U$ et $D_1 = D$, le résultat du Théoreme 1 se simplifie et devient l'ensemble des paires (R, D) qui satisfont

$$R \geq I(U; X_2|Y) \tag{21a}$$
$$D \geq H(X_1|Y, U) \tag{21b}$$

pour U choisi tel que $U \multimap X_2 \multimap (X_1, Y)$ est une chaine de Markov. Alternativement, en faiant la substitution $\tau = H(X_1|Y) - D$, le compromis exprimé par (21) peut être ré-ecrit de façon équivalente de la façon suivante

$$R(\tau) = \min_{p(u|x_2)\,:\, I(U;X_1|Y) \geq \tau} I(U; X_2|Y). \tag{22}$$

L'expression (22) étend la fonction *Information Bottleneck* de [8] au cas où le décodeur observe SI Y^n qui est corrélée avec la ource cachée X_1^n qui est à estimer. Plus précisément, en utilisant la terminologie de [5], la fonction $R(\tau)$ décrit le compomis optimal entre complexité and précision de la description U dans ce cas. Le concept d'Information Bottleneck, qui a été introduit pour la première fois par Tishby et al. [5], a été prouvé utile pour diverses applications d'apprentissage statistique, comem la classification [9], feature selection [10] and others. Thus, the result of Theorem 1, and the tradeoff (22), can be applied to similar problems, especially in the context of supervised learning.

5 Preuve du Théorème 1

5.1 Attaignabilité

Pour la preuve de l'atteignabilité du Théorème 1, nous utilisons une généralisation du résultat de [1, Theorem 2], qui donne une région atteignable pour le modèle de Figure 1 dans le cas d'une mesure générale de distortion. La généralisation comprend l'introduction du partage de temps ('time sharing').

Proposition 1. *(Gastpar Inner Bound [1, Theorem 2] with time-sharing) Un quadriplet (R_1, R_2, D_1, D_2) est atteignable si*

$$R_1 \geq I(U_1; X_1|Y, U_2, Q) \tag{23}$$
$$R_2 \geq I(U_2; X_2|Y, U_1, Q) \tag{24}$$
$$R_1 + R_2 \geq I(U_1, U_2; X_1, X_2|Y, Q) \tag{25}$$
$$D_1 \geq \mathbb{E}[d(X_1, f_1(U_1, U_2, Y, Q))] \tag{26}$$
$$D_2 \geq \mathbb{E}[d(X_2, f_2(U_1, U_2, Y, Q))] \tag{27}$$

pour une mesure de la forme

$$P_{X_1,X_2,Y}(x_1, x_2, y)P_Q(q) \times P_{U_1|X_1,Q}(u_1|x_1, q)P_{U_2|X_2,Q}(u_2|x_2, q), \tag{28}$$

et fonctions de reproduction

$$f_i : \mathcal{U}_1 \times \mathcal{U}_2 \times \mathcal{Y} \times Q \longrightarrow \hat{X}_i \quad pour\ i = 1, 2. \tag{29}$$

La preuve d'atteignabilité s'obtient facilement appliquant le résultat de Proposition 1 au cas d'une mesure de distortion de

type log-loss. PLus spécifiquement, les fonctions de reproductions sont choisies de la façon suivante.

$$f_i(U_1, U_2, Y, Q) := \Pr[X_i = x_i | U_1, U_2, Y, Q] \quad \text{pour } i = 1, 2. \tag{30}$$

Il est facile de voir qu'avec un tel choix, on a

$$\mathbb{E}[d(X_i, f_i(U_1, U_2, Y, Q)] = H(X_i|U_1, U_2, Y, Q) \quad \text{pour} i = 1, 2. \tag{31}$$

5.2 Preuve Réciproque

Premièrement, nous énonçons le lemme suivant, qui est une extension assez facile à obtenir de [4, Lemma 1] au cas où le décodeur also observes une information adjacente Y^n qui est arbitrairement corrélée aux sources à compresser.

Lemme 1. *Soit $Z = (\phi_1^{(n)}(X_1^n), \phi_2^{(n)}(X_2^n))$. Pour le modèle de Figure 1 sous une mesure de distortion de type log-loss, on a $n\mathbb{E}[d(X_i^n, \hat{X}_i^n)] \geq H(X_i^n|Z, Y^n)$ pour $i = 1, 2$.*

L'ingrédient principal de la preuve de la réciproque de Theorem 1 est le lemma suivant.

Lemme 2. *Si un quadriplet $(R_1, R_2, \tilde{D}_1, D_2)$ est atteignable alors il existe une PFM jointe de la forme*

$$P_{X_1,X_2,Y}(x_1, x_2, y)P_Q(q) \times P_{U_1|X_1,Q}(u_1|x_1, q)P_{U_2|X_2,Q}(u_2|x_2, q) \tag{32}$$

, et il existe $D_1 \leq \tilde{D}_1$ tels que

$$D_1 \geq H(X_1|Y, U_1, U_2, Q) \tag{33a}$$
$$D_2 \geq D_1 + H(X_2|Y, U_1, U_2, Q) - H(X_1|Y, U_1, U_2, Q), \tag{33b}$$

et

$$R_1 \geq H(X_1|Y, U_2, Q) - D_1 \tag{34a}$$
$$R_2 \geq I(U_2; X_2|Y, X_1, Q) + H(X_1|Y, U_1, Q) - D_1 \tag{34b}$$
$$R_1 + R_2 \geq I(U_2; X_2|Y, X_1, Q) + H(X_1|Y) - D_1. \tag{34c}$$

Le reste de la preuve de Théorème 1 s'obtient en appliquant le rsultat du lemma suivant.

Lemme 3. *Soit un quadriple (R_1, R_2, D_1, D_2). S'il existe une PMF de la forme* (32) *telle que* (33) *and* (34) *soit satisfaite, alors le quadriple (R_1, R_2, D_1, D_2) appartient à la region décrite par Théoreme 1.*

Références

[1] M. Gastpar, "The wyner-ziv problem with multiple sources," *IEEE Trans. on Info. Theory*, vol. IT-50, pp. 2762–2768, Nov. 2004.

[2] T. Berger and R. W. Yeung, "Multiterminal source coding with one distortion criterion," *IEEE Trans. on Info. Theory*, vol. IT-35, pp. 228–236, 1989.

[3] A. D. Wyner and J. Ziv, "The rate-distortion function for source coding with side information at the decoder," *IEEE Trans. Inf. Theory*, vol. 22, pp. 1–10, Jan. 1976.

[4] T.-A. Courtade and T. Weissman, "Multiterminal source coding under logarithmic loss," *IEEE Trans. on Info. Theory*, vol. 60, pp. 740–761, Jan. 2014.

[5] N. Tishby, F. Pereira, and W. Bialek, "The information bottleneck method," in *Proc. 37th Annu. Allerton Conf. Commun., Control, Comput.*, 1999, pp. 368–377.

[6] P. Harremoes and N. Tishby, "The information bottleneck revisited or how to choose a good distortion theory," in *Proc. IEEE Int. Symp. Information Theory*, Jun. 2007, pp. 566–570.

[7] K. Kittichokechai, Y.-K. Chia, T.-J. Oechtering, M. Skoglund, and T. Weissman, "Secure source coding with a public helper," *IEEE Trans. on Info. Theory*, vol. 62, pp. 3930–3949, Jul. 2016.

[8] P. Gilad-Bachraf, A. Navot, and N. Tishby, "An information theoretic tradeoff between complexity and accuracy," in *Proc. COLT*, 2003, pp. 595–609.

[9] N. Slonim and N. Tishby, "The power of word clusters for text classification," in *Proc. 23rd Eur. Colloq. Inf. Retr. Res.*, 2001, pp. 1–12.

[10] Y. Baram, R. El-Yaniv, and K. Luz, "Online choice of active learning algorithms," *Journal of Machine Learning Research*, vol. 5, pp. 255–291, 2004.