Régions atteignables pour la régression linéaire sur données compressées avec information adjacente

Jiahui WEI[1,2] Elsa DUPRAZ[2] Philippe MARY[1]

[1]IMT Atlantique, CNRS UMR 6285, Lab-STICC, Brest, France

[2]Univ. Rennes, INSA, IETR, CNRS, Rennes, France

Résumé – Nous nous plaçons dans le cadre d'une communication dont le but est d'effectuer une tâche d'apprentissage sur les données transmises. Nous étudions le compromis fondamental qu'il peut exister entre le taux de compression des données et l'erreur faite sur la tâche d'apprentissage sur ces données compressées. La tâche d'apprentissage est une régression linéaire, avec information adjacente au décodeur. Nous étudions dans un premier temps la région de débit-perte en régime asymptotique, c'est-à-dire lorsque la taille de la source tend vers l'infini, puis nous étendons les résultats au régime de longueur de blocs finie. Nous montrons qu'il n'y a pas de compromis entre la compression et l'estimation des paramètres de la régression en régime asymptotique, alors que la conclusion inverse s'applique dans le cas à taille finie.

Abstract – In the framework of goal-oriented communications, this paper investigates the fundamental achievable rate-loss function of a learning task performed on compressed data with side-information at the decoder. Considering that the learning task is a linear regression, we first investigate the rate-loss region in the asymptotic regime, *i.e.*, when the length of the source tends to infinity. The results are further extended to the finite blocklength regime. Our findings reveal that there is no tradeoff between data compression and the learning task in the asymptotic regime, while the opposite conclusion holds for the non-asymptotic case.

1 Introduction

Les communications orientées tâche suscitent de plus en plus d'intérêt et se rencontrent dans divers contextes, comme l'apprentissage distribué dans les réseaux de capteurs. Dans ce contexte, les capteurs transmettent leurs données à un point de collecte qui réalise l'apprentissage. Une question fondamentale est de savoir si le système de codage pour la tâche d'apprentissage doit être conçu de la même manière que pour une communication classique, où l'objectif principal est la reconstruction des données.

Plusieurs travaux théoriques ont étudié cette question dans le cas de deux sources corrélées X et Y, où X est la source à coder et Y est disponible comme information adjacente au décodeur. D'après [3], le débit nécessaire à l'estimation d'un paramètre θ associé à la distribution de probabilité conjointe P_{XY} est inférieur au débit de Slepian-Wolf pour le codage sans perte des deux sources. Le problème de test d'hypothèses sur la distribution P_{XY} a également été largement étudié [4, 8]. Enfin, [6] a établi des bornes génériques sur l'erreur de généralisation d'une tâche d'apprentissage sur X et Y, et a démontré son application a divers problèmes d'apprentissage distribué. Il existe cependant un écart important entre les bornes supérieures et inférieures proposées dans [6].

Une autre question importante est celle de l'existence ou non d'un compromis en terme de débit de codage entre la reconstruction des données et l'apprentissage. Ce problème a été modélisé dans [1] en utilisant la théorie débit-distorsion avec une contrainte supplémentaire sur la perception visuelle, représentée par une mesure de divergence entre deux distributions. Les auteurs ont montré que la région débit-distorsion-perception atteignable est réduite par rapport à la limite inférieure de Shannon, mettant ainsi en évidence un compromis entre les deux critères. Ce compromis a également été mis en évidence dans [4] pour le test d'hypothèse, et dans [10] pour l'identification de données bruitées dans une base de données.

Dans cet article, nous étudions le problème de la régression linéaire entre la source X et l'information adjacente Y, qui n'a pas été traité dans les travaux précédents. Nous souhaitons déterminer le débit de codage source minimum nécessaire à la régression linéaire, sous une contrainte sur l'erreur de généralisation. Nous fournissons la région débit-erreur de généralisation, à la fois dans les régimes asymptotique (Section 3) et non-asymptotique (Section 4), en nous appuyant sur des outils standard de la théorie de l'information [2] et sur des outils développés pour le problème de compression à longueur finie [5, 11, 9]. Ces travaux ne considèrent que le problème de compression pour la reconstruction des données, et une de nos contributions est aussi de les appliquer pour la première fois aux problèmes de codage pour l'apprentissage.

Malgré son apparente simplicité, la régression linéaire est un outil d'apprentissage supervisé très utilisé dans divers domaines tels que l'économie ou la biologie. Des travaux antérieurs, *e.g.*, [1, 4, 10], suggèrent qu'il y a toujours un compromis entre la distorsion et l'apprentissage, mais nous montrons que ce n'est pas le cas pour la régression linéaire en régime asymptotique. De plus, nous améliorons la borne supérieure débit-erreur de généralisation introduite dans [6], qui était très lâche pour la régression linéaire.

2 Énoncé du problème

Dans cet article, X représente une variable aléatoire et x est une réalisation. De plus, $\boldsymbol{X} = (X_1, ..., X_n)$ et $\boldsymbol{x} = (x_1, ..., x_n)$ représentent un vecteur aléatoire et sa réalisation, respectivement, de longueur n. D'autre part, pour une fonction $f : \mathcal{X} \to \mathcal{Y}$, $|f|$ est la cardinalité de l'ensemble $\mathcal{Y}$. Enfin,

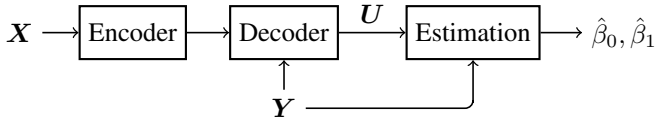

FIGURE 1 : Compression pour la régression linéaire

$\mathbb{E}[X]$, $\mathbb{V}[X]$ et $\mathbb{C}\text{ov}(X, Y)$ sont le premier moment de X, la variance de X, et la covariance des variables aléatoires X et Y, respectivement, et $\log(\cdot)$ est le logarithme en base 2.

2.1 Définition des source

Soient X et Y des variables aléatoires distribuées conjointement suivant P_{XY}, et Y est une information adjacente disponible au décodeur. Nous supposons que Y suit une distribution gaussienne $Y \sim \mathcal{N}(0, \sigma_Y^2)$, et X est définie à partir du modèle linéaire suivant :

$$X = \beta_0 + \beta_1 Y + N, \tag{1}$$

où $N \sim \mathcal{N}(0, \sigma^2)$ est un bruit Gaussien, et $\beta_0, \beta_1 \in \mathbb{R}$ sont des paramètres constants. Par conséquent, $X \sim \mathcal{N}(\beta_0, \sigma_X^2)$, et $\sigma_X^2 = \beta_1^2 \sigma_Y^2 + \sigma^2$. Nous définissons $\mathcal{S} = \{\sigma_Y^2, \sigma^2, \beta_0, \beta_1\}$ comme l'ensemble des paramètres qui définissent entièrement la distribution de probabilité Gaussienne conjointe P_{XY}, et nous supposons que tous les paramètres contenus dans $\mathcal{S}$ sont inconnus.

2.2 Régression linéaire

Au lieu de reconstruire la source X au décodeur, notre objectif est d'effectuer une régression linéaire, c'est-à-dire à estimer β_0 et β_1 à partir des séquences sources $\boldsymbol{X}$ et $\boldsymbol{Y}$ de longueur n, comme illustré sur la figure 1. En utilisant la notation introduite dans [6], nous formalisons le problème de la manière suivante.

Soit $\mathcal{F}$ l'ensemble des fonctions linéaires $f : \mathbb{R} \to \mathbb{R}$ de la forme $f(y) = \alpha_0 + \alpha_1 y$, où $\alpha_0, \alpha_1 \in \mathbb{R}$. L'apprentissage produit une suite de fonctions $\widehat{f}^{(n)} \in \mathcal{F}$, appelées prédicteurs, telles que $\widehat{f}^{(n)} : \boldsymbol{Z}^n \times \mathbb{R} \to \mathbb{R}$, à partir d'une séquence d'entraînement $\boldsymbol{Z} = (\boldsymbol{U}, \boldsymbol{Y}) \in \mathcal{Z}^n$, où $\boldsymbol{U}$ est une séquence aléatoire représentant une version compressée de $\boldsymbol{X}$ dont la définition sera détaillée en Section 3. Étant donné que la régression linéaire estime les coefficients α_0 et α_1 à partir de $\boldsymbol{Z}$, nous pouvons écrire :

$$\widehat{f}^{(n)}(\boldsymbol{Z}, y) = \alpha_0(\boldsymbol{Z}) + \alpha_1(\boldsymbol{Z}) y. \tag{2}$$

Considérons la fonction de perte quadratique $\ell : \mathbb{R}^2 \to \mathbb{R}$ définie par $\ell(x, \hat{x}) = (x - \hat{x})^2$. Pour une certaine fonction $f \in \mathcal{F}$, la perte moyenne est définie comme[1] :

$$J(f, \mathcal{S}) = \mathbb{E}\left[\ell(X, f(Y))\right]. \tag{3}$$

La perte attendue minimale pour un ensemble de fonctions $\mathcal{F}$ et pour $\mathcal{S}$ fixé est définie comme :

$$L^\star(\mathcal{F}, \mathcal{S}) = \inf_{f \in \mathcal{F}} J(f, \mathcal{S}). \tag{4}$$

[1] On peut également définir une perte sur une séquence. Cependant, étant donné que les échantillons des phases d'entraînement et d'inférence sont i.i.d, cela ne change pas l'analyse.

L'erreur de généralisation est définie comme l'espérance de la fonction de perte évaluée sur une nouvelle paire $(\tilde{X}, \tilde{Y})$ qui suit la distribution P_{XY} mais qui est indépendante de la séquence d'apprentissage $\boldsymbol{Z}$, c'est-à-dire :

$$L(\widehat{f}^{(n)}, \mathcal{S}) = \mathbb{E}_{\tilde{X}\tilde{Y}}\left[\ell\left(\tilde{X}, \widehat{f}^{(n)}(\boldsymbol{Z}, \tilde{Y})\right) | \boldsymbol{Z}\right]. \tag{5}$$

2.3 Schéma de codage pour la régression linéaire

Définition 1. *Un schéma d'apprentissage au taux de compression R est défini par une séquence $(e_n, d_n, R, \mathcal{L}_n)$ avec un encodeur $e_n : \mathcal{X}^n \to [\![1, M_n]\!]$, un décodeur $d_n : \mathcal{Y}^n \times [\![1, M_n]\!] \to \mathcal{U}^n$ et une fonction $\mathcal{L}_n : \mathcal{Y}^n \times \mathcal{U}^n \to \mathcal{F}$ tels que*

$$\limsup_{n \to \infty} \frac{\log M_n}{n} \leq R.$$

Définition 2. *Un code (n, M, l) pour la séquence $(e_n, d_n, R, \hat{f}^{(n)})$ est un code avec $|e_n| = M$ tel que*

$$\mathbb{E}_{\boldsymbol{Z}}\left[L(\widehat{f}^{(n)}, \mathcal{S})\right] \leq l, \;\; et \;\; \frac{\log M}{n} \leq R. \tag{6}$$

Définition 3. *Un code (n, M, l, ε) pour la séquence $(e_n, d_n, R, \hat{f}^{(n)})$ et $\varepsilon \in (0, 1)$ est un code avec $|e_n| = M$ tel que*

$$\mathbb{P}\left[L(\widehat{f}^{(n)}, \mathcal{S}) \geq l\right] \leq \varepsilon, \;\; et \;\; \frac{\log M}{n} \leq R. \tag{7}$$

Définition 4. *Pour une perte l, une longueur de bloc n et une une probabilité d'excès ε fixées, la fonction débit-perte est définie par :*

$$R(n, l, \varepsilon) = \inf_R \{\exists \;\; (n, M, l, \varepsilon) \; code\} \tag{8}$$

Définition 5. *Un couple (R, δ) est réalisable si un code (n, M, l) existe tel que*

$$\limsup_{n \to \infty} \mathbb{E}_{\boldsymbol{Z}}\left[L(\hat{f}^{(n)}, \mathcal{S})\right] \leq L^*(\mathcal{F}, \mathcal{S}) + \delta. \tag{9}$$

L'espérance dans le terme de gauche est l'erreur de généralisation moyenne sur la distribution des séquences d'apprentissage. On note que les régions définies dans cette section correspondent à des régions d'erreur de généralisation en fonction du débit. Mais nous y référons comme des régions de débit-perte pour simplifier et par un léger abus de langage.

3 Borne asymptotique sur la fonction débit-perte

Dans [6], il est démontré que l'erreur de généralisation peut être bornée de la manière suivante :

$$\sigma \leq \limsup_{n \to \infty} \mathbb{E}\left[L(\hat{f}^{(n)}, P)^{\frac{1}{2}}\right] \leq \sigma(1 + 2^{-R+1}), \tag{10}$$

où P est la distribution de (X, Y) dans [6]. De plus, σ^2 est le bruit du modèle dans (1), et on peut montrer que $L^\star(\mathcal{F}, \mathcal{S}) = \sigma^2$. Notre prochain résultat fournit une région débit-perte qui améliore la borne supérieure dans (10) pour des sources gaussiennes.

Théorème 1. *Pour tout débit $R > 0$, la paire $(R, 0)$ est réalisable pour une régression linéaire sur des sources gaussiennes avec une perte quadratique.*

Démonstration. On définit U comme la sortie d'un canal de test $P_{U|X}$ qui satisfait la chaîne de Markov $U - X - Y$, et tel que :

$$U = \alpha(X + \Phi) \tag{11}$$

où $\alpha = \frac{\sigma_X^2 - D}{\sigma_X^2}$, $\Phi \sim \mathcal{N}\left(0, \sigma_\phi^2\right)$, $\sigma_\phi^2 = \frac{D\sigma_X^2}{\sigma_X^2 - D}$, et $D > 0$ est un paramètre. Étant donné que l'on suppose que les paramètres $\mathcal{S}$ de la distribution Gaussienne conjointe P_{XY} sont inconnus de l'encodeur et du décodeur, nous utilisons le schéma d'atteignabilité proposé dans [2], où σ_x^2 est estimé à l'encodeur directement puis transmis au décodeur. Ce schéma repose sur le principe de "binning" et "debinning". On génère 2^{nR_1} séquences $\mathbf{u}$, qui sont ensuite uniformément réparties dans 2^{nR} compartiments (les "bins"), avec $R_1 > R$. L'encodeur identifie une séquence $\mathbf{u}$ typique avec $\mathbf{x}$, et transmet l'indice du compartiment auquel $\mathbf{u}$ appartient. Au décodeur, à l'aide de l'information adjacente $\mathbf{y}$ et d'un test de typicalité, la séquence $\mathbf{u}$ est identifiée dans le compartiment. Draper montre dans [2] que la probabilité d'erreur de débinning peut être rendue aussi petite que souhaitée si la taille de séquence n est suffisamment grande. Si $D < \sigma_x^2$, les résultats de [2] montrent que le débit $R_b(D) = \frac{1}{2}\log\left(1 + \frac{\sigma^2}{\sigma_\phi^2}\right)$ est atteignable pour $\mathbb{E}_{XU}\left[d(X, U)\right] \leq D$.

Ensuite, l'estimation par les moindres carrés de β_0 et β_1 à partir de $\boldsymbol{u}$ et $\boldsymbol{y}$ est [7, Chapitre 6] :

$$\hat{\beta}_1 = \frac{1}{\alpha}\frac{\sum_{i=1}^n u_i y_i - n\bar{u}\bar{y}}{\sum_{i=1}^n y_i^2 - n\bar{y}^2}, \quad \hat{\beta}_0 = \frac{1}{\alpha}(\bar{u} - \bar{y}\hat{\beta}_1), \tag{12}$$

où $\bar{y}$ et $\bar{u}$ sont les moyennes empiriques des vecteurs $\boldsymbol{y}$ et $\boldsymbol{u}$. Les estimateurs $\hat{\beta}_0$ et $\hat{\beta}_1$ sont non biaisés. Notons B_0 et B_1 les variables aléatoires représentant $\hat{\beta}_0$ et $\hat{\beta}_1$. L'erreur de généralisation, définie dans l'équation (5), s'écrit :

$$L(\widehat{f}^{(n)}, \mathcal{S}) = (B_0(\boldsymbol{Z}) - \beta_0)^2 + \mathbb{E}[\tilde{Y}^2](B_1(\boldsymbol{Z}) - \beta_1)^2 + \sigma^2. \tag{13}$$

En définissant $S_{yy} = \frac{1}{n}\sum_{i=1}^n (y_i - \bar{y})^2$, nous avons

$$\begin{aligned}\mathbb{V}\left[B_1|\boldsymbol{Y} = \boldsymbol{y}\right] &= \frac{\sigma^2 + \sigma_\phi^2}{nS_{yy}},\\ \mathbb{V}\left[B_0|\boldsymbol{Y} = \boldsymbol{y}\right] &= \frac{\sigma^2 + \sigma_\phi^2}{n}\left(1 + \frac{\bar{y}^2}{\alpha^2 S_{yy}}\right).\end{aligned} \tag{14}$$

L'erreur de généralisation moyenne s'écrit alors :

$$\begin{aligned}&\mathbb{E}_{\mathbf{Z}}[L(\widehat{f}^{(n)}, \mathcal{S})]\\ &= \sigma^2 + \mathbb{E}_{\boldsymbol{Y}}\left[\mathbb{V}\left[B_0|\boldsymbol{Y} = \boldsymbol{y}\right]\right] + \sigma_Y^2\mathbb{E}_{\boldsymbol{Y}}\left[\mathbb{V}\left[B_1|\boldsymbol{Y} = \boldsymbol{y}\right]\right]\\ &= \sigma^2 + \frac{\sigma^2 + \sigma_\phi^2}{n}\left(1 + \frac{1}{\alpha^2}\mathbb{E}\left[\frac{\bar{y}^2}{S_{yy}}\right] + \sigma_Y^2\mathbb{E}\left[\frac{1}{S_{yy}}\right]\right).\end{aligned} \tag{15}$$

Avec $\frac{nS_{yy}}{\sigma_Y^2} \sim \chi^2(n-1)$, $\frac{n\bar{y}^2}{\sigma_Y^2} \sim \chi^2(1)$, où $\chi^2(n)$ est une distribution du Chi-2 avec n degrés de liberté [7, Chapitre 5], et $\frac{(n-1)\bar{y}^2}{S_{yy}} \sim \mathcal{F}_{\text{sne}}(1, n-1)$, où $\mathcal{F}_{\text{sne}}$ est la distribution de Fisher-Snedecor [7, Chapitre 5]. Par les propriétés des distributions du Chi-2 et de Fisher-Snedecor, on obtient [7, Chapitre 5] :

$$\mathbb{E}\left[\frac{1}{S_{yy}}\right] = \frac{n}{(n-3)\sigma_Y^2}, \quad \mathbb{E}\left[\frac{\bar{y}^2}{S_{yy}}\right] = \frac{1}{n-3}, \tag{16}$$

ce qui donne

$$\mathbb{E}_{\mathbf{Z}}[L(\widehat{f}^{(n)}, \mathcal{S})] = \sigma^2 + \frac{(\sigma^2 + \sigma_\phi^2)(1 + 2n\alpha^2 - 3\alpha^2)}{n(n-3)\alpha^2}. \tag{17}$$

Finalement, $\lim_{n\to\infty}\mathbb{E}_{\mathbf{Z}}[L(\widehat{f}^{(n)}, \mathcal{S})] = \sigma^2$. Par conséquent, lorsque $n \to \infty$, $L(\hat{f}^{(n)}, \mathcal{S}) \to L^*(\mathcal{F}, \mathcal{S}) = \sigma^2$. □

Ce résultat montre que l'erreur de généralisation minimale $L^\star(\mathcal{F}, \mathcal{S}) = \sigma^2$ peut être atteinte même avec un taux R très faible, tant que la longueur de la séquence d'entraînement est suffisamment grande. Ce résultat montre aussi que la borne de [6], voir (10), est lâche pour le problème de régression linéaire. Enfin, le schéma d'atteignabilité de [2], que nous utilisons dans la preuve précédente, permet aussi d'atteindre la région débit-distortion optimale. La combinaison de cette remarque et du Théorème 1 nous permet d'énoncer le résultat suivant.

Corollaire 1. *Pour des sources gaussiennes conjointes, il n'y a pas de compromis en termes de débit de codage entre la distorsion et l'erreur de généralisation de la régression.*

4 Borne non asymptotique sur la fonction débit-perte

Dans le problème classique de débit-distorsion avec information adjacente en taille finie, la probabilité d'excès joue un rôle important, car tous les mots de code ne peuvent pas satisfaire la contrainte de distorsion. Ce problème a récemment été étudié en utilisant la notion de dispersion [5, 11]. Dans ce qui suit, nous nous appuyons sur les outils théoriques introduits dans [11], pour proposer une borne atteignable non asymptotique pour la région débit-erreur de généralisation.

Considérons les trois ensembles suivants, similaires à ceux définis dans [11],

$$\mathcal{T}_{\text{p}}(\gamma_{\text{p}}) := \left\{(u, y) : \log\frac{P_{Y|U}(y|u)}{P_Y(y)} \geq \gamma_{\text{p}}\right\}, \tag{18}$$

$$\mathcal{T}_{\text{c}}(\gamma_{\text{c}}) := \left\{(u, x) : \log\frac{P_{X|U}(x|u)}{P_X(x)} \leq \gamma_{\text{c}}\right\}, \tag{19}$$

$$\mathcal{T}_e(l) := \left\{(\tilde{x}, \tilde{y}, u, y) : \ell(\tilde{x}, \hat{f}^{(n)}(\mathbf{z}, \tilde{y})) \leq l\right\}, \tag{20}$$

où γ_p, γ_c sont des seuils prédéfinis, et l est le seuil d'erreur de généralisation. Les deux premiers ensembles sont définis dans [11] pour le problème de reconstruction de données avec information adjacente, tandis que le troisième est spécifique à notre problème de régression linéaire. En conséquence, définissons le vecteur densité d'information-perte comme suit :

$$\mathbf{i}(U, X, Y, \tilde{X}, \tilde{Y}) := \begin{bmatrix} -\log\dfrac{P_{Y|U}(Y|U)}{P_Y(Y)} \\ \log\dfrac{P_{X|U}(X|U)}{P_X(X)} \\ \ell(\tilde{X}, \hat{f}^{(n)}(\mathbf{Z}, \tilde{Y})) \end{bmatrix}. \tag{21}$$

En prenant l'espérance de ce vecteur par rapport à la distribution $P_{UXY\tilde{X}\tilde{Y}}$, on obtient :

$$\mathbf{J}(\mathbf{i}) := \mathbb{E}[\mathbf{i}(U, X, Y, \tilde{X}, \tilde{Y})] \tag{22}$$

$$= \begin{bmatrix} -I(U;Y) \\ I(U;X) \\ \mathbb{E}_{\boldsymbol{Z}\tilde{X}\tilde{Y}}\left[\ell(\tilde{X}, \hat{f}^{(n)}(\mathbf{Z}, \tilde{Y})\right] \end{bmatrix}. \tag{23}$$

La somme des deux premières composantes donne le taux de codage de Wyner-Ziv, $R_b(D)$, donné dans la démonstration du Théorème 1. La matrice de covariance de ce vecteur est

$$\mathbf{V} = \mathbb{C}\text{ov}(\mathbf{i}(U, X, Y, \tilde{X}, \tilde{Y})). \tag{24}$$

Soit k un entier positif et $\mathbf{V} \in \mathcal{R}^{k\times k}$ une matrice positive semi-définie. Étant donné un vecteur aléatoire gaussien $\mathbf{B} \sim \mathcal{N}(0, \mathbf{V})$, le terme de dispersion est défini par rapport à la matrice de covariance comme [9] :

$$\mathscr{S}(\mathbf{V}, \varepsilon) := \{\mathbf{b} \in \mathbb{R}^k : \mathbb{P}[\mathbf{B} \leq \mathbf{b}] \geq 1 - \varepsilon\}. \tag{25}$$

En remplaçant la mesure de distorsion dans [11] par l'erreur de généralisation et en adaptant certaines étapes de l'analyse, nous obtenons le résultat suivant.

Théorème 2. *Pour des constantes arbitraires $\gamma_p, \gamma_c, l \geq 0$, et un entier positif N, il existe un code (n, M, l, ε) satisfaisant*

$$\begin{aligned}\varepsilon \leq & P_{UXY\hat{X}}[(u, y) \in \mathcal{T}_{\mathrm{p}}(\gamma_{\mathrm{p}})^c \cup (u, x) \in \mathcal{T}_{\mathrm{c}}(\gamma_{\mathrm{c}})^c \\ & \cup (\tilde{x}, \tilde{y}, u, y) \in \mathcal{T}_e(l)^c] \\ & + \frac{N}{2^{\gamma_{\mathrm{p}}}|\mathcal{M}|} + \frac{1}{2}\sqrt{\frac{2^{\gamma_{\mathrm{c}}}}{N}}.\end{aligned} \tag{26}$$

Démonstration. La preuve suit les mêmes étapes que dans [11]. □

En choisissant $\gamma_p = \log \frac{N}{|\mathcal{M}_n|} + \log n$ et $\gamma_c = \log N - \log n$, et en appliquant le théorème 2 avec le théorème de Berry-Esséen, et en suivant les mêmes étapes que dans [11], on peut montrer que pour tout $0 < \varepsilon < 1$ et n assez grand, la fonction débit-perte satisfait

$$R_{\mathrm{b}}(n, \varepsilon, l) \leq \inf \left\{ \mathbf{M} \left(\mathbf{J} + \frac{\mathscr{S}(\mathbf{V}, \varepsilon)}{\sqrt{n}} + \frac{2 \log n}{n} \mathbf{1}_3 \right) \right\} \tag{27}$$

avec $\mathbf{M} = [1 \quad 1 \quad 0]$ et $\mathbf{1}_3 = [1 \quad 1 \quad 1]^T$.

5 Résultats et discussion

Dans cette section, nous considérons $\beta_0 = 2$, $\beta_1 = 5$, $\sigma^2 = 4$, $\sigma_Y^2 = 16$, et $\sigma_X^2 = \sigma^2 + \beta_1^2\sigma_Y^2 = 404$. Nous utilisons l'expression (27) pour tracer la région débit-perte à taille finie, où le terme de dispersion $\mathscr{S}(\mathbf{V}, \varepsilon)$ est estimé en générant des échantillons à partir de la distribution conjointe connue P_{UXY}. La densité d'information $i(x; u|y)$ est ensuite estimée à partir de ces échantillons, et la dispersion est estimée à partir de (25).

La figure 2 montre les limites des régions atteignables pour différentes valeurs de n et ε. Plus n augmente, quel que soit ε, et plus la région atteignable se rapproche de la limite asymptotique tracée en noir. De même, pour une longueur de bloc fixe n, lorsque la contrainte sur la probabilité d'excès devient moins stricte, la région atteignable s'agrandit. Cela implique que, pour une erreur de généralisation et une longueur de bloc n fixées, une compression plus importante est possible si on admet une probabilité d'excéder l'erreur cible plus importante.

Cependant, ces résultats ne nous disent rien sur ce qui se passe pour une paire débit-perte en dehors de la région, car il s'agit d'une région atteignable et la caractérisation de la borne extérieure reste un problème ouvert. D'autre part, l'analyse de la région débit-perte pour des problèmes d'apprentissage plus complexes est en cours d'étude[2].

[2]Cette recherche est financée par le laboratoire d'excellence CominLabs ANR-10-LABX-07-01

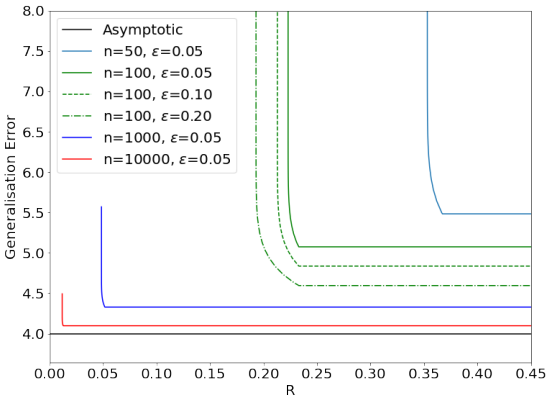

FIGURE 2 : Région débit-erreur de généralisation en fonction de la longueur de bloc n et la probabilité d'excès ε.

Références

[1] Y. BLAU et T. MICHAELI : Rethinking lossy compression : The rate-distortion-perception tradeoff. *In International Conference on Machine Learning*, pages 675–685. PMLR, 2019.

[2] Stark C DRAPER : Universal incremental slepian-wolf coding. *In Proc. 42nd Allerton Conf. on Communication, Control and Computing*, pages 1332–1341, 2004.

[3] M. EL GAMAL et L. LAI : Are slepian-wolf rates necessary for distributed parameter estimation ? *In 2015 53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 1249–1255. IEEE, 2015.

[4] G. KATZ, P. PIANTANIDA et M. DEBBAH : Distributed Binary Detection With Lossy Data Compression. *IEEE Trans. Inf. Theory*, 63(8):5207–5227, 2017.

[5] V. KOSTINA et S. VERDU : Fixed-length lossy compression in the finite blocklength regime. *IEEE Trans. Inf. Theory*, 58(6):3309–3338, 2012.

[6] M. RAGINSKY : Learning from compressed observations. *In 2007 IEEE Information Theory Workshop*, pages 420–425, 2007.

[7] A. C. RENCHER et G. B. SCHAALJE : *Linear models in statistics*. John Wiley & Sons, 2008.

[8] S. SALEHKALAIBAR, M. WIGGER et L. WANG : Hypothesis testing over the two-hop relay network. *IEEE Trans. on Inf. Theory*, 65(7):4411–4433, 2019.

[9] Vincent Y. F. TAN et Oliver KOSUT : On the dispersions of three network information theory problems. *IEEE Trans. Inf. Theory*, 60(2):881–903, 2014.

[10] E. TUNCEL et D. GÜNDÜZ : Identification and lossy reconstruction in noisy databases. *IEEE Trans. Inf. Theory*, 60(2):822–831, 2014.

[11] S. WATANABE, S. KUZUOKA et V. YF TAN : Nonasymptotic and second-order achievability bounds for coding with side-information. *IEEE Trans. Inf. Theory*, 61(4):1574–1605, 2015.