N° d'ordre :

UNIVERSITÉ PARIS-SUD

FACULTÉ DES SCIENCES D'ORSAY

ÉCOLE DOCTORALE 142
MATHÉMATIQUES DE LA RÉGION PARIS-SUD

Laboratoire
LM-Orsay et LSTA

THÈSE SUR TRAVAUX

Spécialité : Mathématiques

Quantification vectorielle en grande dimension : vitesses de convergence et sélection de variables.

par

Clément LEVRARD

Soutenue le 30 Septembre 2014 devant la Commission d'examen :

M.	TAMÀS LINDER	(Rapporteur)
M.	PHILIPPE BERTHET	(Rapporteur)
M.	GÉRARD BIAU	(Directeur de thèse)
M.	PASCAL MASSART	(Directeur de thèse)
M.	FRÉDÉRIC CHAZAL	(Examinateur)
M.	STÉPHANE BOUCHERON	(Président du jury)

Table des matières

Chapitre 1

État de l'art et résumé des travaux

Sommaire

1.1 Introduction à la quantification

Soit P une distribution de probabilité sur un espace euclidien de dimension finie, assimilé à $\mathbb{R}^d$. Un quantificateur Q de taille k, ou k-quantificateur est une fonction de $\mathbb{R}^d$ à valeurs dans un sous-ensemble fini de taille k de $\mathbb{R}^d$. Une telle fonction partitionne l'espace $\mathbb{R}^d$ en k zones, et associe un représentant à chaque zone.

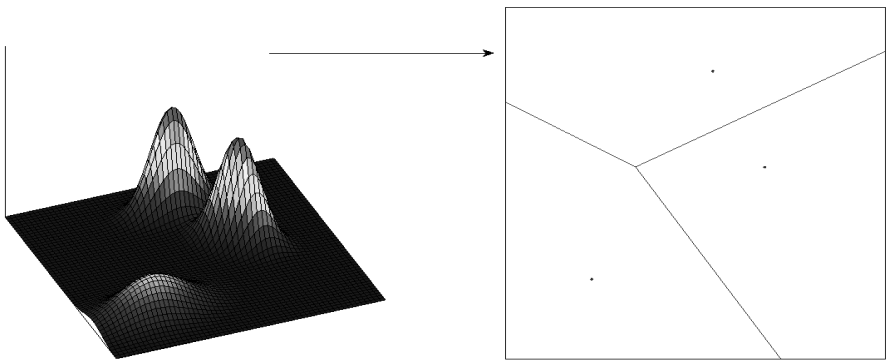

FIGURE 1.1 – Quantification d'une loi de mélange sur $\mathbb{R}^2$

Ce principe est illustré par le schéma 1.1, dont la figure de gauche représente une distribution de probabilité (de type mélange) sur $\mathbb{R}^2$, la figure de droite une suggestion de quantificateur. Tout au long de ce manuscrit, on se donnera un n-échantillon $X_1,\dots,X_n$ de variables indépendantes et identiquement distribuées selon la loi P. Le but poursuivi durant toute cette thèse sera de construire un quantificateur empirique $\hat{Q}$, à partir de l'échantillon $X_1,\dots,X_n$, qui représente au mieux la distribution source.

La quantification vectorielle a été originellement introduite dans les années 40 pour répondre à des problèmes de compression de signaux électriques. C'est effectivement la première idée d'application qui vient à l'esprit : un quantificateur permet de résumer une distribution de probabilité P, potentiellement complexe et occupant continûment l'espace $\mathbb{R}^d$, en un nombre fini de vecteurs.

Un quantificateur est totalement déterminé par k cellules $W_1,\dots,W_k$, formant une partition de $\mathbb{R}^d$, ainsi que par k éléments de $\mathbb{R}^d$, $c_1,\dots,c_k$, appelés points codes, via

$$Q(x) = \sum_{j=1}^{k} c_j \mathbb{1}_{x\in W_k}.$$

Pour une mesure de dissimilarité ϕ, on peut définir un risque associé au quantificateur Q, c'est à dire une manière de mesurer l'adéquation du quantificateur Q avec la distribution source P, via la formule

$$R_\phi(Q) = P\phi(x,Q(x)), \tag{1.1}$$

où par commodité Pf signifie l'intégration de la fonction f par rapport à la loi de P. Cette manière de construire un quantificateur à partir d'un échantillon et d'en mesurer la performance via un risque en prédiction (c'est à dire par rapport à une nouvelle donnée) nous place à l'interface de plusieurs domaines, correspondants à différents paradigmes, avec bien sûr des objets d'études communs.

1.1.1 Quantification et compression du signal

Du point de vue de la compression du signal, deux quantités sont souvent étudiées, ayant rapport avec ce problème de quantification. Premièrement, une bonne partie de la littérature sur le sujet s'intéresse au meilleur quantificateur possible, en ayant accès à la loi P directement, et non à un échantillon tiré suivant P. L'attention est essentiellement portée sur la dépendance de ce risque minimal en le nombre k de points codes que l'on s'autorise, ainsi qu'en la dimension d. Cette dépendance est connue d'un point de vue asymptotique de manière assez précise, pour diverses mesures de dissimilarités ϕ et hypothèses sur P (on peut citer dans ce domaine l'ouvrage de référence [GL00], ainsi que divers articles dans la même lignée, par exemple [DGLP04] et [GLP03], parmi beaucoup d'autres). Par exemple dans le cas où on choisit $\phi(x,y) = \|x-y\|^r$ et où P admet un moment d'ordre strictement plus grand que r et est absolument continue par rapport à la mesure de Lebesgue, le Théorème 6.2 de [GL00] donne un risque minimal asymptotique de type

$$\inf_Q P\|X - Q(X)\|^r \underset{k\to\infty}{\sim} k^{-\frac{r}{d}},$$

où la notation $\sim$ désigne l'équivalence en terme de suites.

Une application directe de ces résultats concernant la meilleure approximation possible de P via k points est l'intégration numérique. En effet, remplaçons la mesure de départ P par sa meilleure approximation sur k points, c'est à dire la mesure ayant pour support les points codes optimaux, avec pour masses respectives les poids des cellules optimales, et notons la δ_Q. On peut alors approcher, pour n'importe quelle fonction f, l'intégrale Pf par $\delta_Q f$. L'intérêt de connaître le risque minimum atteignable par un quantificateur pour la loi P permet de mesurer la précision de cette approximation d'intégrales. Cette application est expliquée en détail dans [Pag98].

L'autre quantité d'intérêt dans ce domaine est directement liée à la quantification vue comme une étape de la transmission du signal (une référence sur le sujet est [GG91]). En effet, si l'on se donne un signal continu à transmettre, la première étape est de le compresser en un nombre fini de vecteurs, pour ensuite encoder ces différentes possibilités de vecteurs, les transmettre (avec bruit éventuel), pour finalement décoder le signal. Le schéma 1.2, inspiré de [BG98], illustre ce processus de manière sans doute plus claire.

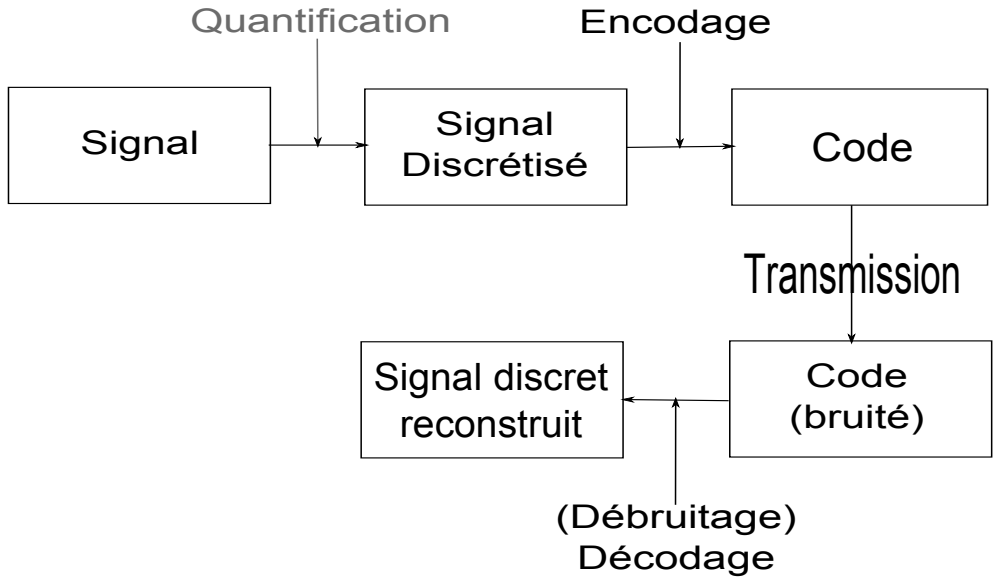

FIGURE 1.2 – Transmission du signal

Notons X le signal à transmettre, modélisé par une variable aléatoire sur $\mathbb{R}^d$. L'étape de quantification dans le processus de transmission fournit alors un quantificateur Q, de performance mesurée par la distorsion $\phi(X, Q(X))$. La construction théorique de quantificateurs performants, pour divers types de lois de X, a donné lieu à de nombreuses publications (par exemple [CLG89] ou [BG98]). Pour établir un parallèle entre ce problème et le nôtre, il convient de considérer l'échantillon $(X_1, \ldots, X_n)$ comme une donnée extérieure, ou préalable, à la quantification du signal X.

Supposons que l'on dispose d'un échantillon d'entraînement $(X_1, \ldots, X_n)$ indépendant et identiquement distribué, ayant pour distribution celle du signal à transmettre X, et que cette donnée préalable ait servi à construire un quantificateur $\hat{Q}$. Dans ce cas, l'étude de l'espérance de la distorsion $\phi(X, \hat{Q}(X))$ coïncide avec l'étude du risque en prédiction $R_\phi(\hat{Q})$ défini en (1.1). Il est intéressant de noter que ce point de vue a fourni les premiers résultats (voir [LLZ94] ou [MZ97]) sur le sujet qui nous

intéresse, à savoir d'étudier l'erreur en prédiction d'un quantificateur construit à partir d'un échantillon d'entraînement.

1.1.2 Quantification et classification non supervisée

L'autre grand point de vue sous lequel il est possible d'envisager la quantification vectorielle est celui de la classification non supervisée, d'où le vocabulaire "échantillon d'entraînement" est par ailleurs issu. En effet, séparer l'espace $\mathbb{R}^d$ en k zones $W_1,\dots,W_k$, comme expliqué précédemment, permet de classer toute donnée future dans l'une de ces k zones. Là encore il convient de distinguer deux sous-domaines au sein de cette approche.

Le premier est celui qui consiste à déterminer la meilleure manière possible de classer l'échantillon d'entraînement $(X_1,\dots,X_n)$, et non pas à étudier l'erreur en prédiction par rapport à une nouvelle donnée X tirée suivant la distribution P. Ce domaine, appelé clustering, est essentiellement orienté vers l'algorithmique, et propose des méthodes variées telles que le Between Cluster Sum of Squares Criterion (introduit dans [WT10]), ou le hierarchical clustering (pour lequel on trouvera une bonne introduction dans le chapitre 12 de [HTF09]). De telles méthodes de clustering peuvent être associées à des procédures de sélection de variables (comme dans [SB08] ou [CWLX14]), problème qui va nous intéresser à la fin de ce manuscrit.

Le second domaine est celui qui étudie l'erreur en prédiction (1.1) d'un quantificateur bâti à partir d'un échantillon d'entraînement. Ce paradigme de construction d'un meilleur prédicteur à partir de données d'entraînement trouve un très large écho dans le domaine de la classification supervisée (domaine dont on trouvera un aperçu complet dans [Vap00] ou plus concis dans [Lug02]). Le formalisme que nous allons adopter fait d'ailleurs très fortement référence à ce domaine, et beaucoup de résultats que nous allons présenter s'inspirent de résultats obtenus en classification supervisée. Les liens entre ces deux domaines ont déjà été exploités par différents auteurs (voir par exemple [Lin02], qui résume l'état de l'art sur les résultats obtenus à partir de méthodes de classification supervisée, ou [BDL08]). Certains de ces résultats seront détaillés dans la Section 1.2.

Notre but est donc, à partir de l'échantillon $(X_1,\dots,X_n)$, de minimiser en Q la fonction $P\phi(x,Q(x))$, en n'ayant pas accès à P. Pour ce faire nous allons adopter une stratégie de minimisation du risque empirique, consistant à remplacer la distribution P par la distribution empirique P_n, qui alloue la masse $1/n$ à chaque élément X_i de l'échantillon. Définissons donc le risque empirique $\hat{R}(Q)$, que l'on espère proche de $R(Q)$, par

$$\hat{R}(Q) = P_n\phi(x,Q(x)) = \frac{1}{n}\sum_{i=1}^{n}\phi(X_i,Q(X_i)).$$

La stratégie de quantification étudiée est alors $\hat{Q} = \arg\min \hat{R}(Q)$, à laquelle il sera fait référence par la suite sous la dénomination de quantificateur empirique. Le fait de pouvoir définir ce risque empirique comme l'intégrale par rapport à la mesure empirique d'une certaine fonction nous permet d'utiliser les outils généraux de l'estimation par minimisation de contraste (une introduction générale à ce domaine peut être trouvée dans [DGL96]).

Il est cependant bon de garder à l'esprit que ce n'est pas la seule méthode possible. Citons par exemple la stratégie du model-based clustering, consistant à approcher P (via P_n) par une loi de mélange en utilisant un critère de maximum de

vraisemblance, puis d'en déduire un partitionnement de l'espace via la règle du maximum à posteriori (un aperçu complet de ce domaine peut être trouvé dans [MP00] ou [FR02]).

Le choix de la mesure de dissimilarité ϕ dans la fonction de risque R_ϕ est important, car de ce choix dépendent beaucoup de propriétés géométriques de l'espace $\mathbb{R}^d$. Un choix classique est celui de la norme L_r, c'est à dire $\phi(x,y) = \|x-y\|^r$ (voir par exemple [GL00] ou [DGLP04]). Cependant il est possible de traiter le cas plus général où ϕ est définie comme une divergence de Bregman (sur ce sujet, on peut citer l'article [Fis10]). Dans ce manuscrit on s'intéressera uniquement à la mesure de dissimilarité définie par $\phi(x,y) = \|x-y\|^2$, donc par la norme euclidienne au carré, ce qui présente deux avantages. De prime abord, pour ce choix de mesure de dissimilarité, l'algorithme de construction effectif du quantificateur empirique est connu, sous le nom de k-means (voir par exemple [Llo82]), et est assez populaire. Ensuite, on s'apercevra dans le Chapitre 3 de ce manuscrit que ce choix permet de profiter au mieux de la structure euclidienne de $\mathbb{R}^d$, permettant ainsi d'obtenir des résultats plus fins que ceux que l'on aurait pu obtenir pour une norme L_r en toute généralité.

Dès lors, on ne s'intéressera plus qu'à la fonction de risque $R(Q) = P\|x-Q(x)\|^2$. Pour cette fonction de risque, on remarque que les quantificateurs optimaux sont de type plus proches voisins, c'est à dire s'écrivant sous la forme

$$x \mapsto \arg\min_{j=1,\dots,k} \|x-c_j\|^2,$$

pour un ensemble de vecteurs $c_1,\dots,c_k$. Par la suite, avec un léger abus de langage, on identifiera un vecteur $\mathbf{c} = (c_1,\dots,c_k)$ avec le quantificateur associé. Les points code c_j seront regroupés au sein d'un dictionnaire $\mathbf{c} = (c_1,\dots,c_k)$, c'est à dire un vecteur de dimension $k \times d$. On cherchera donc à minimiser le risque

$$R(\mathbf{c}) = P \min_{j=1,\dots,k} \|x-c_j\|^2 = P\gamma(\mathbf{c},.),$$

où la fonction γ représente la fonction de contraste $(\mathbf{c},x) \mapsto \min_{j=1,\dots,k} \|x-c_j\|^2$. N'ayant pas accès à P, on s'intéressera aux performances du dictionnaire empirique

$$\hat{\mathbf{c}}_n = \arg\min P_n\gamma(\mathbf{c},.) = \arg\min \frac{1}{n}\sum_{i=1}^n \min_{j=1,\dots,k} \|X_i-c_j\|^2. \tag{1.2}$$

Il est utile de préciser que, dès lors que $P\|x\|^2 < \infty$, de tels minimiseurs existent (voir par exemple [Pol81]). On notera $\mathbf{c}^*$ un minimiseur du vrai risque R, et les deux chapitres suivants de ce manuscrit seront essentiellement dédiés à l'étude de la perte associée au dictionnaire empirique, que l'on définit par

$$\ell(\hat{\mathbf{c}}_n, \mathbf{c}^*) = R(\mathbf{c}) - R(\mathbf{c}^*),$$

qui, rappelons le ici, est une quantité aléatoire en l'échantillon d'entraînement.

1.2 État de l'art

Le premier résultat théorique sur la performance des dictionnaires empiriques définis en (1.2) a été obtenu en 1981, dans l'article [Pol81], et confirme que la stratégie de minimisation du risque empirique converge asymptotiquement. Plus précisément, il est prouvé dans respectivement [Pol81] et [Pol82c], que, dès lors que P

admet un moment d'ordre 2, $\hat{\mathbf{c}}_n \to \mathbf{c}^*$, et que $\ell(\hat{\mathbf{c}}_n, \mathbf{c}^*) \to 0$, en probabilité, quand la taille de l'échantillon croît. Ce premier résultat a ouvert la voie à de nombreuses recherches sur la vitesse de cette convergence en la taille de l'échantillon, ainsi que sur l'influence des autres paramètres du problème sur cette vitesse (taille des dictionnaires k et dimension de l'espace d principalement). Les résultats portant sur ce sujet peuvent être classés en deux catégories : vitesses de convergence lentes et rapides en la taille de l'échantillon.

1.2.1 Vitesses lentes

Pour ce premier type de résultat, un résumé assez complet des connaissances sur le sujet peut être trouvé dans [Lin02]. Historiquement, la première vitesse de convergence non asymptotique provient du domaine de la transmission du signal, et garantit que, si le support de P est inclus dans la boule euclidienne $\mathscr{B}(0,M)$ de rayon M, alors

$$\mathbb{E}\ell(\hat{\mathbf{c}}_n, \mathbf{c}^*) \lesssim M^2 \sqrt{\frac{kd}{n}},$$

où le symbole $\lesssim$ signifie la majoration à un facteur constant près. Ce résultat a été obtenu en appliquant des méthodes ayant fait leurs preuves en classification supervisée par minimisation de contraste. Il est qualifié de vitesse faible car c'est aussi la vitesse type que l'on obtient par la théorie de Vapnik (voir par exemple [Vap82], ou dans un format plus récent [Lug02]), en la taille de l'échantillon n. L'influence de la dimension de l'espace des dictionnaires intervient via le terme $\sqrt{kd}$. Cette dépendance en racine de la dimension est aussi celle qui est attendue, en raisonnant par analogie avec la classification supervisée. Cependant, un résultat récent nous incite à repenser l'influence du terme de dimension. Plus précisément, on peut trouver dans [BDL08] le résultat de type vitesse lente suivant :

$$\mathbb{E}\ell(\hat{\mathbf{c}}_n, \mathbf{c}^*) \lesssim M^2 \frac{k}{\sqrt{n}}, \tag{1.3}$$

à condition que P ait un support inclus dans $\mathscr{B}(0,M)$. Ce résultat à première vue surprenant laisse croire que la dimension d de l'espace ne joue aucun rôle dans la vitesse de convergence du dictionnaire empirique, ce qui est contre intuitif du point de vue de la théorie de la minimisation de contraste. Ce résultat est en partie corroboré par la borne inférieure obtenue sur l'ensemble des distributions à support borné, fournie par le Théorème 1 de [BLL98],

$$\inf_{\hat{Q}} \sup_{P} \mathbb{E}\ell(\hat{Q}, \mathbf{c}^*) \gtrsim M^2 \sqrt{\frac{k^{1-\frac{4}{d}}}{n}}, \tag{1.4}$$

pour n assez grand (dépendant uniquement de k). En effet, la dépendance en la dimension de cette vitesse minimax est rapidement négligeable lorsque cette dimension augmente, ce qui plaide de nouveau pour une importance limitée de la dimension sur la vitesse de convergence de la perte. Par ailleurs, ce résultat confirme que la vitesse lente en la taille de l'échantillon $1/\sqrt{n}$ semble optimale, uniformément sur la classe des distributions bornées, ce qui est l'analogue des résultats obtenus dans [VC74] ou [Sim96], en classification supervisée sur les classes de distributions à dimension de Vapnik fixée.

1.2.2 Vitesses rapides

L'autre catégorie de résultat sur la vitesse de convergence de la perte $\ell(\hat{\mathbf{c}}_n, \mathbf{c}^*)$ est celle des vitesses de convergence dites rapides (en la taille de l'échantillon). Sur ce point les résultats sont un peu plus dispersés, et de natures à priori variées. Deux types de résultats sur les vitesses rapides ont été obtenus, sous des hypothèses différentes.

1.2.2.1 Condition de régularité de Pollard et normalité asymptotique

En utilisant des méthodes de statistique asymptotique pour la M-estimation, ainsi que des arguments de type intégrale entropique de Dudley (voir par exemple [Dud67]), il est possible de prouver que $\sqrt{n}\|\hat{\mathbf{c}}_n - \mathbf{c}^*\|$ converge en loi, si P satisfait une condition introduite dans [Pol82b]. De manière informelle, cette condition requiert que P soit suffisamment régulière et localement quadratique non dégénérée autour des dictionnaires optimaux. Plus précisément, cette condition s'écrit

Condition 1.1 (Condition de régularité de Pollard). *Une distribution P à support borné satisfait la condition de régularité de Pollard si*

1. *P admet une densité continue f par rapport à la mesure de Lebesgue sur $\mathbb{R}^d$,*
2. *La matrice Hessienne de la fonction $\mathbf{c} \longmapsto P\gamma(\mathbf{c}, .)$ est définie positive aux dictionnaires optimaux $\mathbf{c}^*$.*

En utilisant la normalité asymptotique de $\sqrt{n}\|\hat{\mathbf{c}}_n - \mathbf{c}^*\|$, sous réserve que la condition 1.1 soit satisfaite, la vitesse de convergence asymptotique suivante peut alors être obtenue (voir [Cho94]).

$$\ell(\hat{\mathbf{c}}_n, \mathbf{c}^*) = \mathcal{O}_{\mathbb{P}}\left(\frac{1}{n}\right), \tag{1.5}$$

ce qui signifie que la suite $n\ell(\hat{\mathbf{c}}_n, \mathbf{c}^*)$ est bornée en probabilité. La vitesse de convergence $1/n$ semble donc être atteignable. Malheureusement, la nature asymptotique de ce résultat ne permet ni de borner la perte $\ell(\hat{\mathbf{c}}_n, \mathbf{c}^*)$ à n fixé, ni de discuter de l'influence des autres paramètres.

1.2.2.2 Condition de [AGG05] et vitesse non asymptotique

L'obtention de vitesses de convergence rapides et non asymptotiques requiert souvent une condition technique entre variance et perte des fonctions de contraste recentrées, dans le but d'appliquer une inégalité de concentration plus fine que celle des différences bornées (voir par exemple, le Théorème 5.1 de [Mas07]). Ce type de condition a été introduit pour le contexte de la quantification vectorielle dans [AGG05].

Condition 1.2 (Condition de [AGG05]). *Une distribution P à support borné satisfait la condition de [AGG05] si*

$$\exists A > 0 \forall \mathbf{c} \quad \ell(\mathbf{c}, \mathbf{c}^*) \geq A \operatorname{Var}(\gamma(\mathbf{c}, .) - \gamma(\mathbf{c}^*, .)). \tag{1.6}$$

Au contraire de la condition 1.1 de Pollard, la condition 1.6 ne fait pas d'hypothèse de régularité sur la distribution P. En revanche, cette condition est de nature technique, moins interprétable que la condition de régularité de Pollard. Comme

brièvement expliqué au dessus, ce type d'inégalité permet d'utiliser des inégalités de concentration prenant en compte la variance des processus recentrés, comme montré dans [MN06] pour le cadre de la classification supervisée. Cette heuristique a été appliquée pour la quantification vectorielle dans [AGG05], conduisant à la vitesse de convergence suivante.

$$\mathbb{E}\ell(\hat{\mathbf{c}}_n, \mathbf{c}^*) \leq C(A)\frac{\log(n)}{n}, \tag{1.7}$$

où $C(A)$ est une constante dépendant de manière implicite de la constante dans la condition 1.6, ainsi que des autres paramètres k et d. Cette vitesse de convergence est légèrement plus lente que la vitesse asymptotique (1.5). Elle est en revanche de nature non asymptotique. Néanmoins, la nature implicite de la constante $C(A)$ ne permet toujours pas de comprendre l'influence des autres paramètres.

1.3 Résumé des travaux

Au vu des résultats de convergence rapide précédemment cités, une question naturelle était de savoir si on pouvait obtenir une vitesse de convergence non asymptotique en $1/n$, et sous quelles conditions. Répondre à ce problème a donné lieu à deux articles, [Lev13] et [Lev14], correspondant respectivement aux Chapitres 2 et 3 du présent manuscrit. Ces deux chapitres peuvent être lus de manière indépendante. Les résultats présentés dans le Chapitre 3 laissant penser que la quantification vectorielle semble être indiquée pour des espaces de dimension d très grande, comme c'est le cas par exemple en classification de courbes, nous nous sommes intéressés en dernier lieu à une méthode de sélection de variables pour la quantification vectorielle en grande dimension. Les travaux relatifs à la sélection de variables pour la quantification vectorielle composent le dernier Chapitre 4 de ce manuscrit. Ce chapitre peut lui aussi être lu indépendamment du reste de ce manuscrit.

1.3.1 Vitesse non asymptotique optimale dans le cas régulier

Le Chapitre 2 présente deux types de résultats : une vitesse non asymptotique de convergence de la perte, optimale en la taille de l'échantillon n, ainsi qu'une première interprétation des conditions de cette convergence rapide sous la forme de condition de type marge, au sens de [MT99].

1.3.1.1 Équivalence des conditions existantes

En exploitant plus avant l'analogie formelle entre la condition 1.6 et les conditions utilisées pour obtenir des vitesses de convergence rapides en classification supervisée (voir par exemple [MN06]), nous avons introduit une autre condition, à savoir

$$\ell(\mathbf{c}, \mathbf{c}^*) \geq \kappa_0 \|\mathbf{c} - \mathbf{c}^*\|^2, \tag{1.8}$$

pour une constante positive κ_0. Cette condition est de la même nature que la condition 1.6, et semble même plus restrictive. En revanche, pour des distributions suffisamment régulières, la condition 1.1 de Pollard implique directement la condition

1.8. Ceci montre que, bien que de natures différentes, toutes les conditions citées semblent entretenir des liens de dépendance. Nous avons prouvé (dans la Proposition 2.1) que, si P admet une densité continue, alors les trois conditions, 1.1, 1.6 et 1.8 sont équivalentes.

1.3.1.2 Une première condition de marge

Nous avons ensuite porté nos investigations sur l'existence de conditions de type conditions de marge, au sens de [MT99], qui, à l'instar des conditions de marge en classification supervisée, garantiraient qu'une inégalité de type 1.6 soit satisfaite, tout en étant facilement interprétable. Pour rappel, les conditions de marge en classification supervisée sont de type

$$\mathbb{P}\left\{\left|2\eta(X)-1\right| \leq h\right\} \leq Bh^{\beta}, \tag{1.9}$$

pour $B > 0$ et $\beta > 1$, où η désigne la fonction de régression $\eta(x) = \mathbb{P}(Y = 1|X = x)$. De manière informelle, cette condition requiert que le poids du voisinage de la zone critique, c'est à dire dans ce cas d'indécision maximale (où $\eta = 1/2$), doit être suffisamment faible. Le Lemme 9 de [BJM06] prouve l'équivalence des conditions de type 1.9 avec des inégalités techniques de type

$$\mathrm{Var}(\gamma(t,.)-\gamma(t^*,.)) \leq \left(P(\gamma(t,.)-\gamma(t^*,.))\right)^{\alpha},$$

où $\gamma(t,x,y) = \mathbb{1}_{t(x)\neq y}$ est dans ce cas la fonction de contraste utilisée en classification supervisée, et α un exposant relié à l'exposant β de la condition 1.9.

L'analogue de la zone $\eta = 1/2$ en quantification est la frontière du diagramme de Voronoi associé au dictionnaire optimal, c'est à dire

$$\begin{aligned} N^* &:= \bigcup_{j=1}^{k} \partial W_j(\mathbf{c}^*) \\ &= \bigcup_{j=1}^{k} \left\{x | \exists r \forall s \quad \|x-c_j^*\| = \|x-c_r^*\| \leq \|x-c_s^*\|\right\}. \end{aligned} \tag{1.10}$$

Le schéma 1.3 ci-dessous donne une illustration de diagramme de Voronoi associé à un dictionnaire. Le Chapitre 2 fournit un premier résultat dans le but d'établir des conditions de type marge en quantification. Plus précisément, nous avons prouvé dans le cas des distributions à densités continues, que, si P satisfait l'inégalité

$$\sup_{x\in N^*} |f(x)| \leq C(k,d,P),$$

où $C(k,d,P)$ est une quantité dépendant des différents paramètres , alors la condition 1.8 était satisfaite. Cette approche permet de traiter des distributions naturellement polarisées en k zones, comme les mélanges gaussiens, auxquels cette condition peut s'appliquer (voir la Proposition 2.3 du Chapitre 2).

1.3.1.3 Vitesse non asymptotique optimale

Le Chapitre 2 fait état d'un premier résultat sur la convergence rapide, quand la condition 1.8 est satisfaite, à savoir

$$\mathbb{E}\ell(\hat{\mathbf{c}}_n, \mathbf{c}^*) \leq \kappa_0 M^2 \frac{C(k,d,P)}{n}, \tag{1.11}$$

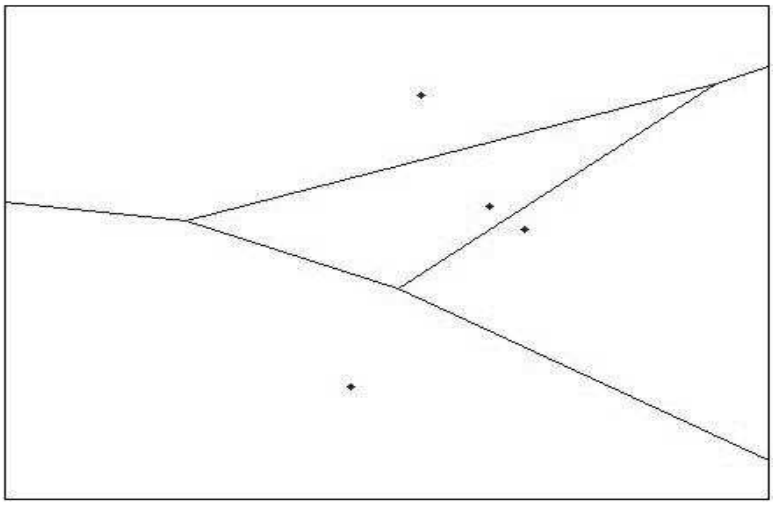

FIGURE 1.3 – Diagramme de Voronoi générique

où $C(k,d,P)$ est une constante non explicite en ses différents paramètres (voir le Théorème 2.1). Ce résultat est établi en adaptant les techniques de localisation utilisées en classification supervisée au cadre de la quantification vectorielle (voir [Kol06] pour une introduction à ces techniques). La vitesse de convergence (1.11) est non asymptotique, comme en (1.7), et atteint la vitesse de la convergence asymptotique en $1/n$ du résultat (1.5).

1.3.2 Condition de marge et influence de la dimension

Comme souligné dans le paragraphe précédent, le résultat de convergence fourni par le Chapitre 3 ne permet pas d'expliquer l'influence des autres paramètres du problème de quantification (k, d, ...) sur la vitesse de cette convergence. Dans le Chapitre 3, plusieurs résultats explicites en les différents autres paramètres sont présentés, ainsi qu'une nouvelle condition de type marge, plus générale que celle obtenue au chapitre précédent.

1.3.2.1 Une condition générale de marge

Le Chapitre 3 poursuit, de la même manière que le Chapitre 2, le but de déterminer une vraie condition de marge pour la quantification, mais en utilisant des techniques légèrement différentes. Comme attendu, il s'avère que l'on peut donner une condition portant sur le voisinage de la zone d'indécision maximale, suffisante pour la satisfaction de la condition 1.8. Cette condition peut s'écrire

$$\exists r_0 > 0 \quad \forall t \leq r_0 \quad \mathbb{P}\{d(X,N^*) \leq t\} \leq a(P)t, \tag{1.12}$$

où $a(P)$ est une constante fixée au préalable et dépendant essentiellement de la taille k des dictionnaires (une définition précise pourra être trouvée en 3.1). Il est intéressant de souligner que, contrairement à la condition 1.9 en classification supervisée, seul l'exposant 1 intervient dans la majoration du poids du voisinage de

N^*. Ce détail est dû au fait que la fonction de contraste γ, dans le cadre de la quantification, est intrinsèquement liée à la distance euclidienne au carré, fixant ainsi l'exposant adéquat pour une majoration du poids du voisinage de N^*. On peut aussi remarquer le fait que la condition 1.12 ne requiert qu'un contrôle local (mais néanmoins explicite) du poids du voisinage. Cette souplesse s'avèrera nécessaire pour traiter certains exemples, comme les mélanges gaussiens dans la Section 3.3.3 de ce manuscrit. Néanmoins, cet aspect local entraîne quelques complications techniques pour l'obtention d'une constante globale et explicite dans la condition 1.8. Cette difficulté sera surmontée par l'introduction de paramètres globaux naturels dans le cadre de la quantification, tel le facteur de séparation ε, défini comme la perte minimale des minimiseurs locaux et non globaux du risque, dont on trouvera le détail dans le Chapitre 3, en 3.2 plus précisément.

Enfin, la condition 1.12, contrairement à la première condition de type marge introduite dans le Chapitre 2, ne requiert aucune régularité de la part de la distribution P. Cette dernière remarque permet de présenter un cadre commun pour les deux grandes classes de distributions satisfaisant des conditions de convergence rapide, à savoir les distributions à densité continue satisfaisant la condition 1.1 de Pollard et les distributions satisfaisant la condition 1.6.

1.3.2.2 Influence des paramètres sur la vitesse de convergence

La condition de marge 1.12 une fois définie, nous sommes en mesure de calculer des vitesses de convergence pour la perte $\ell(\mathbf{c}, \mathbf{c}^*)$, d'une part non asymptotiques, mais aussi explicites en les différents autres paramètres introduits précédemment. Sans entrer dans les détails, ces deux bornes fournissent les vitesses de convergence suivantes

$$\ell(\mathbf{c}, \mathbf{c}^*) \lesssim \kappa_0 M^2 \frac{kd\sqrt{\log(kd)}}{n}, \tag{1.13}$$

ainsi que

$$\ell(\mathbf{c}, \mathbf{c}^*) \lesssim \kappa_0 M^2 \frac{k}{n}, \tag{1.14}$$

qu'on pourra retrouver dans le Théorème 3.1. La borne (1.13) est en accord avec les résultats de classification supervisée, où un terme de dimension de l'espace des paramètres considérés (ici kd) est usuellement présent (voir par exemple, en estimation de densité [BBM99]). De plus, le facteur constant de l'inégalité (1.13) est connu.

La seconde borne, (1.14) est plus surprenante, car elle est totalement indépendante de la dimension d de l'espace euclidien considéré. De fait, la borne (1.13) est obtenue en utilisant des outils de chaînage, combinés à des intégrales entropiques de type Dudley (voir, par exemple, [Dud67] ou [Pol82a]), faisant intervenir naturellement la dimension de l'espace des dictionnaires via le cardinal des recouvrements de petite taille de l'ensemble des dictionnaires. En revanche, la borne (1.14) contourne l'argument de chaînage, imitant en cela le résultat (1.3) pour les vitesses lentes. Il est intéressant de souligner le fait que le facteur constant de (1.14) est inconnu, car provenant du principe de generic chaining introduit en [Tal05].

Le fait que (1.14) ne dépende pas de la dimension nous a poussé à envisager la quantification dans un espace de Hilbert séparable, donc de dimension non nécessairement finie. Pour ce nouveau cadre de travail, beaucoup de résultats simples

en dimension finie, comme l'existence de dictionnaires optimaux, se révèlent non triviaux. En nous appuyant sur les résultats de [Fis10] et [GLP07], qui traitent le cadre plus général de la quantification dans les espaces de Banach, le chapitre 3 traite le cas de la dimension infinie, et prouve que, si la condition 1.12 est satisfaite, alors la borne (1.14) reste valide.

1.3.2.3 Borne inférieure sur la vitesse de convergence

En dernier lieu, nous établissons dans le Chapitre 3 une borne inférieure minimax analogue à (1.4) sur la vitesse de convergence minimale sur l'ensemble des distributions satisfaisant une condition de marge. La borne inférieure obtenue est du même ordre de grandeur en la taille de l'échantillon n que les bornes supérieures (1.14) et (1.13). Plus précisément, si $\mathscr{D}(\varepsilon)$ désigne l'ensemble des distributions à support borné satisfaisant une condition de marge de rayon r_0 et ε séparées, alors la Proposition 3.3, que l'on trouvera dans le Chapitre 3, montre que

$$\sup_{P\in\mathscr{D}(\varepsilon)} \mathbb{E}\ell(\hat{\mathbf{c}}_n,\mathbf{c}^*) \gtrsim \kappa_0 M^2 \frac{k^{-\left(1+\frac{4}{d}\right)}}{\sqrt{n}}, \tag{1.15}$$

quand $\varepsilon \sim 1/\sqrt{n}$. Ce résultat indique que les dépendances en la taille de l'échantillon n des vitesses de convergence énoncées en (1.14) et (1.13), sous la condition de marge 1.12, semblent être du bon ordre de grandeur. En revanche, la comparaison de la borne inférieure (1.15) avec la borne (1.14) révèle des différences concernant l'influence de k et d : s'il semble avéré que d ne joue presque aucun rôle dans ces deux résultats, une question ouverte reste néanmoins de savoir quelle est l'influence réelle de k sur la perte $\ell(\hat{\mathbf{c}}_n,\mathbf{c}^*)$. Enfin, la portée de la borne inférieure 3.3 est amoindrie du fait qu'elle est valide uniquement pour un régime particulier du paramètre ε, et non à paramètre ε fixé.

1.3.3 k-means et sélection de variables

Bien que théoriquement applicable en dimension infinie, et donc potentiellement adaptée à la classification de courbes, deux détails suggèrent que l'implémentation effective de la stratégie de quantification par minimisation du risque empirique, via l'algorithme des k-means (voir [Llo82]), nécessite une étape de réduction du nombre de variables, comme expliqué dans [AF12]. Premièrement, d'un point de vue pratique, il est impossible de stocker un nombre infini de coefficients. Ensuite, il est intéressant de préciser que la borne (1.14) dépend de la taille du support de P en norme 2. Soit alors M tel quel $Supp(P) \subset \mathscr{B}_2(0,M)$. Si l'on suppose que chaque coordonnée est bornée par une constante, notée M_∞, ce qui est une hypothèse courante en classification non supervisée, il apparaît que $M \leq dM_\infty$, ce qui redonne une dépendance en la dimension dans nos deux vitesses de convergence. Pour ces deux raisons, nous nous sommes intéressés dans la dernière partie de cette thèse à une méthode combinée de quantification et sélection de variables.

Le problème de la sélection de variables en classification non supervisée est un domaine actif, avec autant d'approches différentes que de types de résultat. Beaucoup de méthodes, telle le Penalized Between Cluster Sum of Squares (voir [WT10] ou [CWLX14]) évaluent la performance de leurs algorithmes via la probabilité de bien classer à posteriori l'échantillon d'entraînement. Pour ce type de procédures, plusieurs résultats théoriques sur la classification à posteriori ont pu être donnés

(voir [CWLX14] par exemple), sous des hypothèses d'indépendance des différentes variables. Ce type d'approche se concilie mal avec le problème de quantification que nous étudions, d'une part parce que le risque de quantification $R(\mathbf{c})$ est un risque en prédiction, portant sur une nouvelle donnée potentielle de loi P, d'autre part parce que l'hypothèse de coordonnées indépendantes semble trop restrictive et de peu d'intérêt pour les résultats théoriques que nous avons précédemment établis.

D'autres approches concilient la réduction du nombre de variables d'intérêt et la mesure de la performance en prédiction. On peut par exemple citer les procédures model-based pénalisées (voir par exemple [Mey13] et [MM13]), qui consistent à approcher la distribution P par un modèle de mélange gaussien, en pénalisant les modèles où les moyennes des composantes sont de support étendu. Ces procédures garantissent généralement que la densité sélectionnée est proche d'une densité de compromis entre l'approximation de la loi et la taille du support des moyennes (voir [MM13]), au sens de la distance de Hellinger. En revanche peu de résultats théoriques ont été donnés sur la convergence des moyennes vers des vecteurs de support réduit, à notre connaissance.

Il est enfin intéressant de remarquer qu'en pratique, les procédures de quantification de type minimisation de risque empirique sont souvent utilisées après l'application préalable d'une procédure empirique de sélection de variables. Les critères appliqués pour cette étape de sélection font l'objet d'études via simulations et exemples, on peut en trouver un certain nombre dans [SB08]. Pour donner une idée des procédures utilisées en pratique, on peut citer l'exemple de la classification de courbes de consommation EDF par décomposition dans une base d'ondelettes, seuillage des coefficients, et quantification vectorielle via l'algorithme des k-means, présentée dans [ABCP13]. Le critère utilisé pour seuiller les coefficients d'ondelettes, donc pour sélectionner les variables d'intérêt, prend en compte le ratio entre la variance de la loi marginale et la variance totale de la loi. En bref, si $\hat{\sigma}_p$ représente la variance empirique marginale de la coordonnée p, la variable p sera éliminée si $\hat{\sigma}_p^2/\hat{\sigma}^2$ reste en dessous d'un certain seuil. Bien qu'utilisée en pratique, et confirmée par des exemples d'applications probants, aucune garantie théorique n'est donnée pour ce type de procédure empirique.

Ces procédures de sélection de variables peuvent être englobées dans le paradigme plus général consistant à chercher des points code dans un sous-espace de dimension réduite. Une méthode très populaire pour atteindre cet objectif consiste à combiner ACP et k-means sur le sous-espace obtenu. Deux exemples actuels illustrant cette heuristique sont les algorithmes RKM et FKM (respectivement Reduced K-Means et Factorial K-Means), introduits dans [DeS] et [VK01], consistant à trouver un sous-espace de dimension déterminée au préalable, et des points codes appartenant à ce sous-espace, minimisant la distorsion et la distorsion de la distribution projetée respectivement. Á l'instar des méthodes de sélection de variables évoquées précédemment, quelques résultats théoriques sur l'erreur de classification de l'échantillon d'entraînement ont été prouvés (voir, par exemple, [TCKV10]), sous des conditions de nature géométrique sur le support de la distribution. Cependant, des résultats théoriques en prédiction ont aussi été démontrés pour ces deux algorithmes, prouvant que les dictionnaires ainsi construits convergent presque sûrement vers des dictionnaires optimaux au sens de chacun de ces deux problèmes (voir, respectivement, [Ter12] et [Ter13]). Bien qu'aucune vitesse de convergence n'aie été démontrée à ce jour, il est fort probable que les techniques utilisées dans [LLZ94] ou [BDL08] puissent être employées dans ce cas pour obtenir un analogue

des vitesses obtenues dans la Section 1.2.1. Le principal défaut de ces méthodes est qu'il faut fixer au préalable la dimension du sous-espace recherché, ce qui n'est pas le cas de la procédure que nous allons introduire. Enfin, la notion de variable pertinente devient plus difficile à établir dès lors que les sous-espaces choisis ne sont pas forcément perpendiculaires aux axes des coordonnées.

1.3.3.1 Introduction de la procédure de sélection de variables et propriétés des dictionnaires empiriques régularisés

Le Chapitre 4 de ce manuscrit introduit une procédure de réduction du nombre de variables et de quantification simultanée, en sélectionnant un dictionnaire suivant le critère

$$\hat{\mathbf{c}}_{n,\lambda} \in \arg\min_{\mathbf{c}} P_n \gamma(\mathbf{c},.) + \lambda I(\mathbf{c}), \tag{1.16}$$

où lambda est un paramètre à déterminer et $I(\mathbf{c})$ un terme de pénalité destiné à privilégier les dictionnaires ayant un faible support, c'est à dire ayant beaucoup de k-coordonnées $(c_1^{(p)},\dots,c_k^{(p)})$ nulles. Cette procédure avait déjà donné lieu à un article, [SWF12], où un résultat de convergence asymptotique sous des conditions très restrictives sur P avait été établi. Dans un premier temps, nous avons étudié une procédure classique de group-Lasso, comme dans [Bac08], où

$$I_L(\mathbf{c}) = \sum_{p=1}^{d} \sqrt{c_1^{(p)2} + \ldots + c_k^{(p)2}}. \tag{1.17}$$

La forme de la fonction de pénalité $I(\mathbf{c})$ est calibrée pour que les composantes de chaque point code au sein d'un même dictionnaire s'annulent en même temps. Pour définir l'importance de la composante p pour la quantification, il est nécessaire d'introduire les quantités suivantes

$$\begin{cases} \sigma_p^2 &= P^{(p)}\|x\|^2, \\ \hat{\sigma}_p^2 &= P_n^{(p)}\|x\|^2, \\ R_p^* &= \min_{\mathbf{c}} P^{(p)}\gamma(\mathbf{c},.), \\ \hat{R}_p^* &= \min_{\mathbf{c}} P_n^{(p)}\gamma(\mathbf{c},.), \end{cases} \tag{1.18}$$

où $P^{(p)}$ représente la distribution marginale de P suivant la coordonnée p. En exploitant les conditions de Karush-Kuhn-Tucker (dont on trouvera un énoncé dans [BV04]), il apparaît que les coordonnées dont la différence $\hat{\sigma}_p^2 - \hat{R}_p$ est grande sont nulles dans l'estimateur $\hat{\mathbf{c}}_{n,\lambda}$ (voir Proposition 4.1 pour un énoncé plus précis). On peut remarquer que ce critère est assez proche des critères empiriques mentionnés plus haut, consistant à éliminer les variables dont la variance empirique ne dépasse pas un certain seuil. Ce critère peut cependant poser des problèmes d'échelle : une coordonnée p dont l'amplitude ne serait pas suffisante se retrouve automatiquement éliminée, indépendamment de ses performances prédictives. Pour pallier ce défaut, nous avons introduit dans le Chapitre 4 une deuxième pénalité de type Weighted group-Lasso, à savoir

$$\hat{I}_{WL}(\mathbf{c}) = \sum_{p=1}^{d} \hat{\sigma}_p \left(\sqrt{c_1^{(p)2} + \ldots + c_k^{(p)2}} \right).$$

Les mêmes conditions de Karush-Kuhn-Tucker confirment cette fois ci que les coordonnées dont le ratio $\hat{R}_p/\hat{\sigma}_p^2$ est grand sont éliminées, résolvant ainsi le problème, d'échelle mentionné plus haut (de même que précédemment, on trouvera un énoncé complet de ce résultat en l'espèce de la Proposition 4.2). En revanche, cette pénalité dépend de l'échantillon d'entraînement $X_1,\dots,X_n$, ce qui en complique l'étude théorique.

1.3.3.2 Résultats théoriques de convergence et de sélection de variables

Pour ces deux types de pénalité, le Chapitre 4 présente trois types de résultats. Dans un premier temps, en envisageant la pénalisation Lasso comme de la sélection de modèles parmi les boules de norme 1 (approche dont on trouvera une aperçu complet dans [MM11]), des résultats concernant l'erreur en prédiction des estimateurs $\hat{\mathbf{c}}_{n,\lambda}$ sont donnés par les Théorèmes 4.1 et 4.3, de type

$$\ell(\hat{\mathbf{c}}_{n,\lambda},\mathbf{c}^*) \leq \inf_{r>0} \inf_{I(\mathbf{c})\leq r} \left(\ell(\mathbf{c},\mathbf{c}^*) + K\lambda r\right),$$

avec grande probabilité, quand $\lambda \gtrsim k\log(d)/\sqrt{n}$ et pour une constante K. Ces résultats garantissent que, sur n'importe quelle boule L_1 de rayon R, $\hat{\mathbf{c}}_{n,\lambda}$ est aussi performant en termes de risque R que le meilleur dictionnaire sur cette boule L_1, à un terme $K\lambda R$ près.

D'autre part, la borne inférieure en $\log(d)/n$ sur la constante de pénalisation s'avère être assez usuelle dans beaucoup d'autres modèles, notamment de régression linéaire, comme expliqué dans [vdG08]. L'obstacle majeur à l'extension des résultats de type Lasso du contexte des modèles linéaires généralisés à la quantification par la méthode de minimisation du risque empirique pénalisé est le fait que la fonction de contraste que nous utilisons n'est pas linéaire.

En revanche, bien que non linéaire, la fonction de contraste γ peut s'écrire comme un minimum de fonctions linéaires, via

$$\gamma(\mathbf{c},x) = \|x\|^2 + \min_{j=1,\dots,k} \langle -2x, c_j\rangle + \|c_j\|^2.$$

Cette remarque est au cœur de l'obtention des vitesses de convergence qui ne dépendent pas de la dimension, telle celle donnée par le Théorème 2.1 de [BDL08] ou (1.14). En effet, l'idée principale de la preuve de (1.14) est qu'il est possible de majorer la complexité de Rademacher (qui est le terme dépendant usuellement de la dimension de l'espace des paramètres) associée aux fonctions de contraste γ par une complexité Gaussienne, mais associée à des fonctions de contraste linéaires. Ce point technique permet alors d'adapter les résultats obtenus en régression pénalisée pour les modèles linéaires généralisés (voir [vdG08]) au cadre de la quantification par minimisation du risque empirique.

Par conséquent, si P satisfait une condition de marge de type 1.8, il est alors possible de garantir (voir les Théorèmes 4.5 et 4.8) que, pour un choix de $\lambda \gtrsim \sqrt{k\log(kd)\log(n)/n}$, avec forte probabilité,

$$\lambda I(\hat{\mathbf{c}}_{n,\lambda} - \mathbf{c}_\lambda^*) \leq \left(3R(\mathbf{c}_\lambda^*) + \frac{K\lambda^2}{\kappa_0} I_0(\mathbf{c}_\lambda^*)\right) \vee \lambda^2, \tag{1.19}$$

où $\mathbf{c}_\lambda^*$ est défini comme un minimiseur du terme de droite dans (1.19), avec $I_0(\mathbf{c}) = \left|\{p|\mathbf{c}^{(p)} \neq 0\}\right|$ si on choisit $I(\mathbf{c}) = I_L(\mathbf{c})$, et $I_0(\mathbf{c}) = \sum_{\{p|\mathbf{c}^{(p)}\neq 0\}} \sigma_p^2$ si $I(\mathbf{c}) = \hat{I}_{WL}(\mathbf{c})$. Dans

les deux cas, le dictionnaire $\mathbf{c}^*_\lambda$ réalise un compromis entre performance prédictive et taille de l'ensemble des coordonnées non nulles.

Plus précisément, les coordonnées non nulles de $\mathbf{c}^*_\lambda$ peuvent être caractérisées pour les deux types de pénalité proposées, ce qui fait l'objet des Propositions 4.3 et 4.5. Dans le cas où $I(\mathbf{c}) = I_L(\mathbf{c})$, les coordonnées telles que $\sigma_p^2 - R_p^* \lesssim \lambda^2$ sont éliminées, ce qui pose le même problème d'échelle que pour le critère empirique, à savoir que les variables de faible amplitude seront systématiquement éliminées. Comme précédemment, si la pénalité choisie est $\hat{I}_{WL}(\mathbf{c})$, alors les coordonnées telles que $1 - R_p^*/\sigma_p^2 \lesssim \lambda^2$ seront nulles dans le dictionnaire compromis, ce qui répond au problème d'échelle.

Ces deux résultats de consistence vont de pair avec des résultats en prédiction sur $\ell(\hat{\mathbf{c}}_{n,\lambda}, \mathbf{c}^*)$ du même ordre de grandeur que la borne sur l'écart entre $\hat{\mathbf{c}}_{n,\lambda}$ et le dictionnaire compromis $\mathbf{c}^*_\lambda$, énoncés dans les Théorèmes 4.5 et 4.8, à savoir

$$\ell(\hat{\mathbf{c}}_{n,\lambda}, \mathbf{c}^*) \leq K\left(\frac{\lambda^2 I_0(\mathbf{c}^*)}{\kappa_0} \vee \lambda^2\right), \tag{1.20}$$

où comme expliqué plus haut $I_0(\mathbf{c}^*)$ caractérise la taille du support de $\mathbf{c}^*$ pour les deux choix de pénalité, donc est de l'ordre de la dimension d ou de σ^2. Comme dans le cas de la régression via des modèles linéaires généralisés, ces résultats en prédiction peuvent être substantiellement améliorés en réeffectuant une procédure de quantification empirique (non pénalisée) sur l'ensemble des coordonnées sélectionnées. Un développement possible serait alors de collecter l'ensemble de la trajectoire de régularisation pour les différentes valeurs de λ, et au sein de ce sous ensemble de variables appliquer une procédure de sélection de modèle classique en pénalisant par un terme de dimension pour ces différents sous ensembles, comme proposé dans [Mey12].

Il est aussi possible de nuancer ces résultats en soulignant le fait que la borne inférieure fournie par la théorie pour le facteur de régularisation λ est déterminée à une constante numérique inconnue près, issue des méthodes de generic chaining présentées dans [Tal05], et appliquées pour le Lasso pour des modèles linéaires généralisés dans [vdG13]. Par conséquent une étape de calibration des constantes de pénalisation semble inévitable dans le but d'une implémentation effective. Plusieurs techniques de calibration semblent possibles, en suivant les méthodes proposés dans le cadre de la régression linéaire via le Lasso, par exemple en sélectionnant la constante de pénalisation par validation croisée comme proposé dans [Cha12].