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)

Remerciements

Au moment de solder les comptes de cette période charnière de mon parcours, il est autant d'usage qu'il m'est agréable d'évoquer les personnes qui ont rendu cette expérience sinon possible, du moins enrichissante et heureuse. Mes premières pensées vont à mes deux directeurs de thèse, Gérard Biau et Pascal Massart, sans les compétences et l'encadrement de qui l'enclenchement, bien sûr, mais aussi le bon déroulement et l'aboutissement de ce projet personnel n'auraient jamais eu lieu. Je n'aurais sans doute jamais été capable d'écrire correctement le manuscrit qui suit sans la patience et les conseils de Gérard, prodigués quelle que soit l'heure quand cela s'avérait nécessaire, et les discussions aussi stimulantes que matinales avec Pascal, dont les issues m'occupaient souvent pour des semaines. Au delà de cette aide indispensable, je veux surtout vous remercier pour m'avoir guidé dans un monde dont les rouages m'étaient totalement inconnu, et par dessus tout pour votre soutien et amitié indéfectibles durant ces quatre années.

I would thank Professor Linder for accepting to report this thesis, this was a pleasure and honour. Je tiens également à remercier Philippe Berthet d'avoir rapporté ce manuscrit, ses commentaires ont été précieux, ainsi que les deux derniers membres de ce jury, à savoir Stéphane Boucheron et Frédéric Chazal. La dernière partie de ce manuscrit n'aurait pu voir le jour sans l'aide opportune de Bertrand Michel en plein milieu du mois d'août, je tiens donc ici l'occasion de le remercier pour son assistance et sa gentillesse. En ce qui concerne le cadre et l'ambiance de travail à Orsay durant ces quatre années, ils n'auraient pu atteindre ce niveau sans le concours, par ordre décroissant de bureau, de Lionel, Emilien, Tristan, Giancarlo, Cagri, Morzi (malgré ses convictions politiques), Pierre-Antoine, Olivier, Vincent, Arthur, Valérie, et Elodie, sans compter ceux que j'ai probablement oublié. Dans un autre registre, je veux aussi remercier Valérie Lavigne dont l'efficacité m'a souvent dépêtré du bourbier administratif. La cotutelle de l'UPMC m'a donné l'occasion de rencontrer Baptiste, Benjamin, Cécile et Erwan, que je salue ici en espérant que les prochains séminaires AirBnb seront à la hauteur de la réputation des précédents. Au gré de ces séminaires et autres rencontres professionnelles j'ai pu faire la connaissance d'un bon nombre de sympathiques jeunes statisticiens, je pense par exemple à Andrès, Mélisande, Sébastien, Cristophe, Laure, Adil... Je profite de cette occasion pour leur témoigner le plaisir que j'ai à les côtoyer, syndrome de Stockholm/Saint Flour mis à part.

Ces remerciements me donnent aussi un prétexte pour remercier mes deux professeurs de collège et lycée, mesdames Lucas et Nedelec, qui par leur exigence m'ont donné le nécessaire coup de pied à l'orgueil, dont l'élan m'a emmené suffisamment loin pour me donner l'envie de continuer. Je remercie aussi mes parents, pour tout, et plus généralement ma famille, à commencer par mes frères Simon et Valentin, malgré leur effet délétère sur mon outil de travail, et mes sœurs Louise et Marie, qui m'ont supporté toutes ces années. Le cercle familial ne serait pas au complet sans mentionner mes grands-parents, auxquels je dédie ces travaux, Michaël, Martine, Olivalouis..., bref tous ceux qui font de chaque retour en Normandie une expérience inimitable et physiquement éprouvante, et que je remercie ici.

Vient maintenant le tour de la seconde famille, que je me suis choisie au fil des ans et qui m'est devenue nécessaire à bien des égards. Je remercie la bande des diots, Kim, Thibaut, Gauthier, Cong, Gary, Diego, Raphaël, pour notre expérience micro-communiste durant ces 4 ans d'école. Enfin les mots me manquent pour exprimer toute mon affection au canal historique, amis de la première heure, auxquels je dois beaucoup d'éléments constitutifs de mon bien-être, et qui m'empêchent de me prendre trop au sérieux. Ils sont Raphaël, Etienne, Perrine, Nico(s), Antoine, Thomas, Marie(s), avec une mention spéciale pour Romain et Virginie, avec qui j'ai partagé une des périodes les plus heureuses de mon existence.

Pour conclure, je remercie Lucie pour le monde que nous nous sommes créé, et dans lequel je me sens bien.

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.

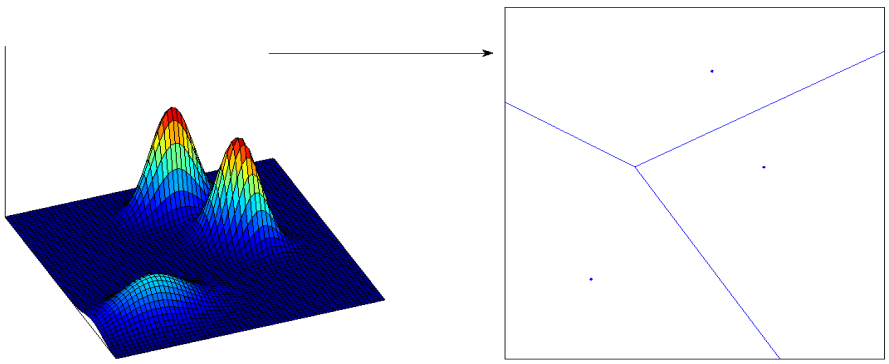

FIGURE 1.1 – Quantification d'une loi de mélange sur $\mathbb{R}^2$

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.

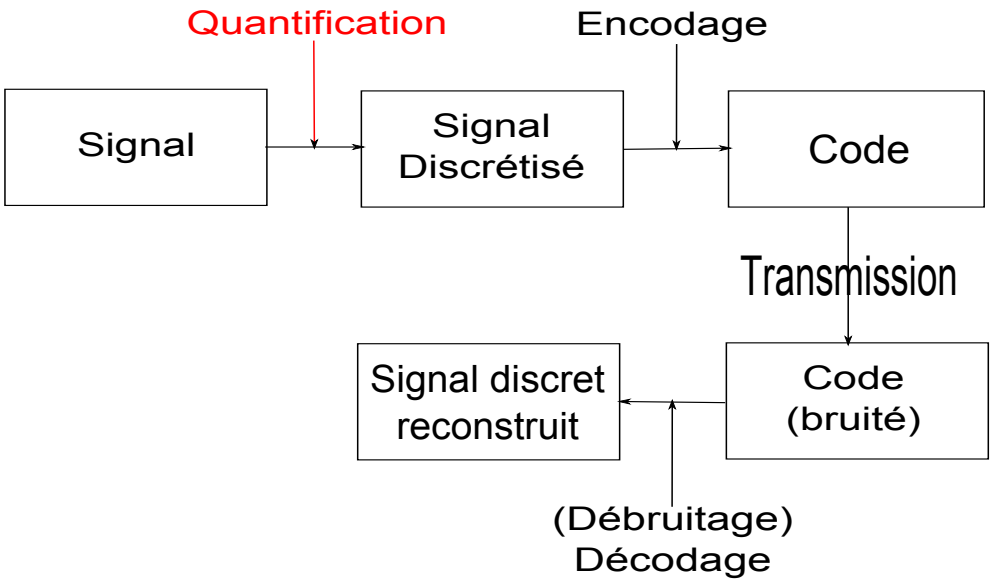

FIGURE 1.2 – Transmission du signal

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

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

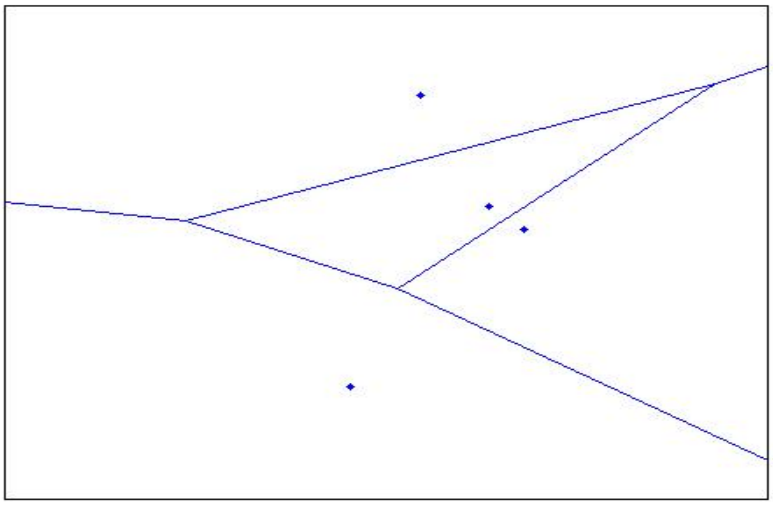

FIGURE 1.3 – Diagramme de Voronoi générique

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].

Chapitre 2

Fast rates for empirical vector quantization

Le Chapitre 2 présente les premiers résultats de convergence non asymptotique optimale de la perte pour le dictionnaire minimisant le risque empirique, ainsi qu'une étude des différentes conditions de la littérature concernant ce sujet. Il a fait l'objet d'un article, [Lev13], publié dans *Electronic Journal of Statistics*.

Sommaire

We consider the rate of convergence of the expected loss of empirically optimal vector quantizers. Earlier results show that the mean-squared expected distortion for any fixed probability distribution supported on a bounded set and satisfying some regularity conditions decreases at the rate $\mathcal{O}(\log n/n)$. We prove that this rate is actually $\mathcal{O}(1/n)$. Although these conditions are hard to check, we show that well-clustered distributions with continuous densities supported on a bounded set are included in the scope of this result.

2.1 Introduction

Empirical vector quantizer design is a way to answer the problem of identifying groupings of similar points that are relatively away from one another, or, in other

words, to partition the data into dissimilar groups of similar items. For a comprehensive introduction to this topic, the reader is referred to the monograph [GL00]. To isolate meaningful groups from a cloud of data is a topic of interest in many fields, from social science to biology. In fact this issue originates in the theory of signal processing in the late 40's, known as the quantization issue, or lossy data compression (a good introduction to this field can be found in [GG91]).

To be more precise, let P denote a probability distribution over the Euclidean space $\mathbb{R}^d$. A k-point quantizer Q, also called k-level quantizer in the case where $d = 1$, is a map from $\mathbb{R}^d$ to $\mathbb{R}^d$, whose image set is made of exactly k points, that is $\left|Q(\mathbb{R}^d)\right| = k$. By considering the preimages of these points, such a map partitions the whole space into k groups, and assigns each group a representative.

For any P-integrable function $f : \mathbb{R}^d \longrightarrow \mathbb{R}$, we will denote by Pf the integral of f with respect to P. To measure how well a quantizer Q performs in representing the source distribution P, we introduce the distortion

$$R(Q) := P\|x - Q(x)\|^2,$$

when $P\|x\|^2 < \infty$. This choice of distortion function is convenient, since it takes advantage of the underlying Euclidean structure. Note however that several authors deal with more general distortion functions (see, e.g., [GL00] or [Fis10]).

For a k-point quantizer Q with images $Q(\mathbb{R}^d) = \{c_1, \ldots, c_k\}$, we will call code points of Q the points $c_1, \ldots, c_k$, and codebook of Q an arbitrary vector $\mathbf{c}$ of $(\mathbb{R}^d)^k$, the components of which are the code points, e.g. $\mathbf{c} := (c_1, \ldots, c_k)$. Without loss of generality we restrict our attention to nearest neighbor quantizers, namely quantizers satisfying the condition $\|x - Q(x)\| = \min_{c \in Q(\mathbb{R}^d)} \|x - c\|$. Intuitively, a nearest neighbor quantizer sends any vector x to the nearest code point c_i to x. Note that a nearest neighbor quantizer is determined by its codebook $\mathbf{c}$ with ties arbitrarily broken. Since we only deal with continuous distributions, how ties are broken will not matter.

Throughout this paper, quantizers will be represented by their codebook. This choice will allow us to handle vectors rather than maps, in order to take advantage of the underlying Euclidean structure. From this point of view, the distortion function takes the form

$$R(\mathbf{c}) := P\|x - Q(x)\|^2 = P \min_{j=1,\ldots,k} \|x - c_j\|^2,$$

when $P\|x\|^2 < \infty$.

Let $X_1, \ldots, X_n$ be a independent and identically distributed sample with distribution P. The goal here is to find a codebook $\hat{\mathbf{c}}_n$, drawn from the data $X_1, \ldots, X_n$, whose distortion is as close as possible to the optimal distortion $R^* := \inf_{\mathbf{c} \in (\mathbb{R}^d)^k} R(\mathbf{c})$. To solve the problem, most approaches to date attempt to implement the principle of empirical error minimization in the vector quantization context. According to this principle, good code points can be found by searching for ones that minimize the empirical distortion over the training data, defined by

$$\hat{R}_n(\mathbf{c}) := \frac{1}{n} \sum_{i=1}^{n} (X_i - Q(X_i))^2 = \frac{1}{n} \sum_{i=1}^{n} \min_{j=1,\ldots,k} \|X_i - c_j\|^2.$$

The existence of such empirically optimal codebooks has been formally established for more general distortion functions in Theorem 4.12 of [GL00], following the approach of Lemma 8 in [Pol82c]. Denote by $\hat{\mathbf{c}}_n$ one of these empirically optimal codebooks. If the training data represents the source well, $\hat{\mathbf{c}}_n$ will hopefully also perform

near optimally on the real source. Roughly, this means that we expect $R(\hat{\mathbf{c}}_n) \approx R^*$. The problem of quantifying how good empirically designed codebooks are, compared to the truly optimal ones, has been extensively studied (see, e.g., [Lin02]).

To reach the latter goal, a standard route is to exploit the Wasserstein distance between the empirical distribution and the source distribution, to derive upper bounds on the average distortion of empirically optimal codebooks. Following this approach, it is proved in [Pol81] that, if $P\|x\|^2 < \infty$, then $R(\hat{\mathbf{c}}_n) - R^* \longrightarrow 0$ almost surely, as $n \to \infty$. Using techniques borrowed from statistical learning theory, it may be derived that (see, e.g., [LLZ94] or [BDL08]), provided that the support of P is bounded, $\mathbb{E}(R(\hat{\mathbf{c}}_n) - R^*) = \mathcal{O}(1/\sqrt{n})$, where the expectation is taken over the training sample $X_1, \ldots, X_n$. It has been established in [BLL98] that this rate is minimax over distributions supported on a finite set of points. More recently, the results exposed in [Ant05] improved the numerical constants mentioned in this minimax result, and also proved that the minimax rate over distributions over bounded sets with continuous densities is still $1/\sqrt{n}$.

However, faster rates individual convergence rates can be achieved, under certain conditions. For example, it was shown in [Cho94], following the main result of [Pol82b], that if the source distribution satisfies some regularity conditions, then $R(\hat{\mathbf{c}}_n) - R^* = \mathcal{O}_{\mathbb{P}}(1/n)$, where we recall that a sequence of random variables $Y_n = \mathcal{O}_{\mathbb{P}}(1/n)$ if, for all positive real number M, $\mathbb{P}(n|Y_n| \geq M) \to 0$ as $n \to \infty$. Nevertheless, this consistency result does not provide any information on how many training samples are needed to ensure that the average distortion of empirically optimal codebooks is close to the optimum. It has been established in [AGG05] that $\mathbb{E}(R(\hat{\mathbf{c}}_n) - R^*) = \mathcal{O}(\log n/n)$, under other conditions, paying a $\log n$ factor to derive a non-asymptotic bound. It is worth pointing out that the conditions cannot be checked in practice, and consequently remain of theoretical nature. Moreover, the rate of $\mathcal{O}(1/n)$ for the average distortion can be achieved when the source distribution is supported on a finite set of points (see, e.g., [AGG05]). Consequently, an open question is to know whether this optimal rate can be attained for more general distributions, and under what set of conditions.

In the present paper, we improve previous results of [AGG05], by getting rid of the $\log n$ factor, adding some minor regularity conditions on P. To this aim we use statistical learning arguments, and prove that the average distortion of empirically optimal codebooks decreases at the rate $\mathcal{O}(1/n)$, under certain conditions. To get this result we use techniques such as the localization principle (see, e.g., [BBM08] or [Kol06]). The condition we offer can be easily interpreted as a margin-type condition, similar to the ones exposed in [MN06], showing a clear connection between statistical learning theory and vector quantization.

Furthermore, we offer equivalences between different sets of regularity conditions which guarantee that the distortion of the empirically optimal codebook decreases at a fast rate. More precisely, we prove that conditions required in [Pol82b], conditions required in [AGG05], and conditions we required, are equivalent, in the case where P has a continuous density. It is worth pointing out that all conditions mentioned above are of theoretical nature, and remain hard to understand. We also give in this paper a more reader-friendly sufficient condition.

The paper is organized as follows. In Section 2 we introduce notation and definitions of interest. In Section 3 we offer our main results. These results are discussed in Section 2.4, and illustrated on examples such as Gaussian mixtures or quasi-finite distributions. Finally, proofs are gathered in Section 2.5.

2.2 The quantization problem

Throughout the paper, $X_1,\ldots,X_n$ is a sequence of independent $\mathbb{R}^d$-valued random variables with distribution P. To frame the quantization problem as a statistical learning one, we first have to consider quantization as a contrast minimization issue. To this aim we introduce the following notation. Consider a nearest neighbor quantizer with codebook $\mathbf{c} = (c_1,\ldots,c_n)$. The contrast function γ is defined as

$$\gamma : \left\{ \begin{array}{ccc} \left(\mathbb{R}^d\right)^k \times \mathbb{R}^d & \longrightarrow & \mathbb{R} \\ (\mathbf{c},x) & \longmapsto & \min\limits_{j=1,\ldots,k} \left\|x-c_j\right\|^2 \end{array} \right. .$$

Within this framework, the risk $R(\mathbf{c})$ takes the form $R(Q) = R(\mathbf{c}) = P\gamma(\mathbf{c},.)$, where $Pf(.)$ means integration of the function f with respect to P. Denote by P_n the empirical distribution that is induced on $\mathbb{R}^d$ by the n-sample $X_1,\ldots,X_n$, namely, for any measurable subset $A \subseteq \mathbb{R}^d$, $P_n(A) = |\{i|X_i \in A\}|$. Once P_n introduced, the empirical risk $\hat{R}_n(\mathbf{c})$ can be expressed as $P_n\gamma(\mathbf{c},.)$. Remark that an optimal $\mathbf{c}^*$ minimizes $P\gamma(\mathbf{c},.)$, whereas $\hat{\mathbf{c}}_n \in \arg\min_{\mathbf{c}\in(\mathbb{R}^d)^k} P_n\gamma(\mathbf{c},.)$. It is worth pointing out that, if $P\|x\|^2 < \infty$, then the existence of both $\hat{\mathbf{c}}_n$ and $\mathbf{c}^*$ are guaranteed by Theorem 4.12 in [GL00].

Let $\mathscr{M}$ denote the set of optimal codebooks, and let $\mathbf{c}^* \in \mathscr{M}$ be an optimal codebook, with code points c_i^*. The Voronoi cell V_i^*, also called quantization cell, is defined as the subset of $\mathbb{R}^d$ made of the points which are closer to c_i^* than any other c_j^*, i.e.

$$V_i^* := \left\{x \in \mathbb{R}^d | \quad \forall j \neq i \quad \|x-c_i^*\| \leq \|x-c_j^*\|\right\}.$$

It may be noted that many authors prefer to define the Voronoi cell V_i^* as the open set

$$\left\{x \in \mathbb{R}^d | \quad \forall j \neq i \quad \|x-c_i^*\| < \|x-c_j^*\|\right\}.$$

However, this choice of convention will not matter, since every boundary of an optimal Voronoi cell has zero P-measure, that is

$$P\left(\left\{x \in \mathbb{R}^d | \|x-c_i^*\| = \|x-c_j^*\|\right\}\right) = 0$$

for $i \neq j$ (see, e.g., Theorem 4.12 in [GL00]).

It is well known that any optimal codebook satisfies the centroid condition (see, e.g., Section 6.2 in [GG91]), which states that each optimal code point is chosen to minimize the distortion over its associated cell, or, in other words

$$P\left(\|x-c_i^*\|^2 \mathbf{1}_{x\in V_i^*}\right) = \inf_{c\in\mathbb{R}^d} P\left(\|x-c\|^2 \mathbf{1}_{x\in V_i^*}\right),$$

where, for any measurable subset $A \subset \mathbb{R}^d$, $\mathbf{1}_A$ denotes the indicator function of the set A. An immediate consequence of the centroid condition is that every c_i^* satisfies

$$c_i^* = \frac{P\left(x\mathbf{1}_{V_i^*}\right)}{p_i^*},$$

where $p_i^* := P(V_i^*)$ is nonzero, according to Theorem 4.1 in [GL00]. In the case where P has a density, it is proved in Lemma A of [Pol82b] that $\mathbf{c} \longmapsto P\gamma(\mathbf{c},.)$ is differentiable. In this case, it is easy to show that the centroid condition takes the form

$$\nabla P\gamma(\mathbf{c}^*,.) = 0,$$

where ∇f denotes the differential of f for any differentiable map $f : \left(\mathbb{R}^d\right)^k \longrightarrow \mathbb{R}$ (see, e.g., Section 6.2 of [GG91]).

Let $\mathbf{c} \in \left(\mathbb{R}^d\right)^k$ be a $k \times d$ vector, and let $\mathbf{c}^* \in \mathscr{M}$ be an optimal codebook. We introduce the loss, or distortion redundancy, to compare the performance of $\mathbf{c}$ and $\mathbf{c}^*$, namely

$$\ell(\mathbf{c}, \mathbf{c}^*) := R(\mathbf{c}) - R(\mathbf{c}^*) = P\left(\gamma(\mathbf{c}, .) - \gamma(\mathbf{c}^*, .)\right).$$

Throughout the paper we will use the following assumptions on the source distribution P. For $c \in \mathbb{R}^d$ and any positive real number $M > 0$, let $\mathscr{B}(c, M)$ denote the closed ball of radius M and center c in $\mathbb{R}^d$. To be precise

$$\mathscr{B}(c, M) = \left\{x \in \mathbb{R}^d \mid \quad \|x - c\| \leq M\right\}.$$

Assumption 2.1 (Peak Power Constraint). *The probability distribution P is such that $P(\mathscr{B}(0,1)) = 1$,*

For convenience we only consider distributions whose support is included in $\mathscr{B}(0,1)$. However, it is important to note that our results hold for distributions whose support is included in $\mathscr{B}(0, M)$, for an arbitrary M. In fact, a distribution supported on $\mathscr{B}(0, M)$ can easily be turned into a distribution supported on $\mathscr{B}(0,1)$, via an homothetic transformation. Therefore, we will only state results for distributions for which the support is included in $\mathscr{B}(0,1)$. Note that Assumption 2.1 is stronger than the requirement $P\|x\|^2 < \infty$, as it imposes that P is supported on a bounded subset of $\mathbb{R}^d$. However, it is likely that our results can be extended to the case where just the weaker assumption $P\|x\|^2 < \infty$ is required, using techniques such as in [MZ97] or [CP12].

The following regularity requirement, introduced in [Pol82b], was initially used to derive an asymptotic rate of convergence for the loss $\ell(\hat{\mathbf{c}}_n, \mathbf{c}^*)$.

Assumption 2.2 (Pollard's regularity condition). *The distribution P satisfies the following two conditions :*

1. *P has a continuous density f with respect to the Lebesgue measure on $\mathbb{R}^d$,*
2. *The Hessian matrix of $\mathbf{c} \longmapsto P\gamma(\mathbf{c}, .)$ is positive definite for all optimal codebooks $\mathbf{c}^*$.*

It may be noted that Condition 1 of Assumption 2.2 does not guarantee the existence of a second derivative for the function $\mathbf{c} \longmapsto P\gamma(\mathbf{c}, .)$. Nevertheless, if Assumption 2.1 and Condition 1 of Assumption 2.2 are satisfied, then it can be proved that $\mathbf{c} \longmapsto P\gamma(\mathbf{c}, .)$ is twice differentiable (see, e.g., Lemma C in [Pol82b]).

Let V_i be the Voronoi cell associated with c_i, for $i = 1, \ldots, k$. The Hessian matrix is composed of the following $d \times d$ blocks :

$$H(\mathbf{c})_{i,j} = \begin{cases} 2P(V_i) I_d - 2\sum_{\ell \neq i} r_{i\ell}^{-1} \sigma\left[f(x)(x - c_i)(x - c_i)^t \mathbf{1}_{\partial(V_i \cap V_\ell)}\right] & \text{for} \quad i = j \\ 2{r_{ij}}^{-1} \sigma\left[f(x)(x - c_i)(x - c_j)^t \mathbf{1}_{\partial(V_i \cap V_j)}\right] & \text{for} \quad i \neq j \end{cases}, \tag{2.1}$$

where $r_{ij} = \|c_i - c_j\|$, $\partial(V_i \cap V_j)$ denotes the possibly empty common face of V_i and V_j, and σ means integration with respect to the $(d-1)$-dimensional Lebesgue measure. For a proof of that statement, we refer to [Pol82b].

Assumption 2.2 is hard to check in general. However, there are some cases where it can be proved that $H(\mathbf{c}^*)$ is positive definite for every optimal codebook $\mathbf{c}^*$. For

example, it is proved in Corollary 2 of [AGG05] that, if $d = 1$ and P has a strictly log-concave density, then P satisfies Assumption 2.2. As will be shown in Corollary 2.1, this is also the case when the density is small enough at the boundaries of optimal Voronoi cells.

When Assumption 2.1 and Assumption 2.2 are satisfied, it has been derived in [Cho94] that $\ell(\hat{\mathbf{c}}_n, \mathbf{c}^*) = \mathcal{O}_{\mathbb{P}}(1/n)$. This result relies on the previous result of [Pol82b], which established the asymptotic normality of $\sqrt{n}(\hat{\mathbf{c}}_n - \mathbf{c}^*)$. To get this asymptotic result, conditions under which the distortion and the Euclidean distance are connected have been used, along with chaining arguments to bound from above a term which looks like a Rademacher complexity, constrained on an area around an optimal codebook. Note that a similar method has been employed in [Kol06] to apply the localization principle.

The following Assumption 2.3 is the assumption required to obtain our main result. It demands direct connections between the Euclidean distance, the loss $\ell(\mathbf{c}, \mathbf{c}^*)$ and the variance $\mathrm{Var}_P(\gamma(\mathbf{c},.) - \gamma(\mathbf{c}^*,.))$, taken with respect to P.

Assumption 2.3. *The distribution P satisfies the following two technical conditions :*

$$\text{(H1)} \qquad \exists A_1 > 0 \quad \forall \mathbf{c} \in \mathscr{B}(0,1) \quad \ell(\mathbf{c}, \mathbf{c}^*(\mathbf{c})) \geq A_1 \|\mathbf{c} - \mathbf{c}^*(\mathbf{c})\|^2,$$

where $\mathbf{c}^*(\mathbf{c}) \in \underset{\mathbf{c}^* \in \mathscr{M}}{\arg\min} \|\mathbf{c} - \mathbf{c}^*\|$, *and*

$$\text{(H2)} \exists A_2 > 0 \quad \forall \mathbf{c} \in \mathscr{B}(0,1) \quad \forall \mathbf{c}^* \in \mathscr{M} \quad \mathrm{Var}_P\left(\gamma(\mathbf{c},.) - \gamma(\mathbf{c}^*,.)\right) \leq A_2 \|\mathbf{c} - \mathbf{c}^*\|^2.$$

Notice that, contrary to Assumption 2.2, Assumption 2.3 does not require P to have a continuous density.

When considering several optimal codebooks $\mathbf{c}^*$ to be compared to a general $\mathbf{c}$, it is natural to choose one among the closest to $\mathbf{c}$, hence the choice of $\mathbf{c}^*(\mathbf{c})$. Furthermore, since for all $\mathbf{c}^*$ in $\mathscr{M}$, $\ell(\mathbf{c}, \mathbf{c}^*(\mathbf{c})) = \ell(\mathbf{c}, \mathbf{c}^*)$, we will write $\ell(\mathbf{c}, \mathbf{c}^*)$ without specifying which $\mathbf{c}^* \in \mathscr{M}$ is at stake.

It is worth pointing out that Corollary 1 in [AGG05] ensures that, if Assumption 2.1 is satisfied, then Assumption 2.2 implies Assumption 2.3. In the same paper, P is assumed to satisfy the following weaker Assumption 2.4.

Assumption 2.4 (Condition of Antos, Györfi and György). *There exists $A > 0$ such that*

$$\forall \mathbf{c} \in \mathscr{B}(0,1) \quad \mathrm{Var}_P(\gamma(\mathbf{c},.) - \gamma(\mathbf{c}^*(\mathbf{c}),.)) \leq A\,\ell(\mathbf{c}, \mathbf{c}^*).$$

Assumption 2.4 is at first sight weaker than Assumption 2.3, since it only requires a comparison between $\ell(\mathbf{c}, \mathbf{c}^*)$ and $\mathrm{Var}_P(\gamma(\mathbf{c},.) - \gamma(\mathbf{c}^*(\mathbf{c}),.))$, without comparing them to the intermediate $\|\mathbf{c} - \mathbf{c}^*(\mathbf{c})\|^2$.

It has been shown in Theorem 2 of [AGG05] that, if P satisfies Assumption 2.1 and Assumption 2.4, then $\mathbb{E}\ell(\hat{\mathbf{c}}_n, \mathbf{c}^*) = \mathcal{O}(\log(n)/n)$. As explained before the statement of Assumption 2.4, it has been proved in Corollary 1 of [AGG05] that, provided that Assumption 2.1 is satisfied, Assumption 2.2 implies Assumption 2.4. Consequently, if P denotes a distribution satisfying Assumption 2.1 and Assumption 2.2, the result exposed in [Cho94] states that $\ell(\hat{\mathbf{c}}_n, \mathbf{c}^*)$ converges in probability to 0 at the rate $1/n$, whereas Theorem 2 of [AGG05] indicates that $\ell(\hat{\mathbf{c}}_n, \mathbf{c}^*)$ converges to 0 at the rate $\log(n)/n$ in expectation. Therefore, a question of interest is to know whether these two rates are truly different, or whether the $\log(n)$ factor is artificial.

To be more precise, the main argument of the derivation of Theorem 2 of [AGG05] is a concentration inequality based on the fact that the variance and the expectation of the distortion are connected to get their result. Interestingly, this point of view has been developed in [BBM08] to get bounds on the classification risk of the SVM, using the localization principle. That is the approach that will be followed in the present paper.

2.3 Main results

The conditions we require to obtain our main result differ from the conditions required in [Pol82b], and from those proposed in Theorem 2 of [AGG05]. Consequently it is natural to make connections between these different sets of conditions clear. This is the aim of the following Proposition.

Proposition 2.1. *Let P be a distribution on $\mathbb{R}^d$. Then the following sets of conditions on P are equivalent :*

$$\left\{ \begin{array}{l} \textit{Assumption 2.1} \\ \textit{Assumption 2.2} \end{array} \right. \Leftrightarrow \left\{ \begin{array}{c} P \textit{ has a continuous density} \\ \mathscr{M} \textit{ is finite} \\ \textit{Assumption 2.1} \\ \textit{Assumption 2.3} \end{array} \right. ,$$

and

$$\left\{ \begin{array}{l} \textit{Assumption 2.1} \\ \textit{Assumption 2.2} \end{array} \right. \Leftrightarrow \left\{ \begin{array}{c} P \textit{ has a continuous density} \\ \mathscr{M} \textit{ is finite} \\ \textit{Assumption 2.1} \\ \textit{Assumption 2.4} \end{array} \right. .$$

Roughly, Proposition 2.1 states that, provided that P has a density which is continuous and whose support is bounded, all the conditions which ensure fast rates of convergence for the average distortion of the empirically optimal quantizer are equivalent. The following Theorem offers a new bound on the loss $\ell(\hat{\mathbf{c}}_n, \mathbf{c}^*)$, when P satisfies any of the three sets of conditions proposed in Proposition 2.1.

Theorem 2.1. *Suppose that P has a density f and $\mathscr{M}$ is a finite set. Assume that Assumption 2.1 and Assumption 2.3 are satisfied. Then, denoting by $\hat{\mathbf{c}}_n$ an empirical risk minimizer, we have*

$$\mathbb{E}\ell(\hat{\mathbf{c}}_n, \mathbf{c}^*) \leq \frac{C_0}{n},$$

where C_0 is a positive constant depending on P, k and d.

This result shows that a convergence rate of $1/n$ can be achieved in expectation, at the price of a few more conditions on the source distribution than those required in Theorem 2 of [AGG05]. To be more precise, this result only requires that Assumption 2.1 and Assumption 2.4 are satisfied. According to Proposition 2.1, provided that P has a continuous density and $\mathscr{M}$ is finite, the set of conditions required in Theorem 2 of [AGG05] turns out to be equivalent to the set of conditions required in [Pol82b] or to the the set of conditions mentioned in Theorem 2.1.

As illustrated by the proof in Section 2.5.3, the constant C_0 mentioned in Theorem 2.1 strongly depends on the constants A_1 and A_2 introduced in Assumption 2.3. Consequently, to understand how C_0 depends on k, d or P, the exact dependency of

A_1 and A_2 on P, k and d has to be known. Unfortunately, the existence of such an A_1 often derives from compactness arguments. Thus we are not able in this paper to explain how C_0 depends on the other parameters k, d and P.

The technical result, from which Theorem 2.1 is derived, is presented in Section 2.5.3, Theorem 2.3. It is based on a version of Talagrand's inequality mentioned in [Bou02] and its application to localization, following the approach of [MN06]. It is important to note that drawing connections between the Euclidean distance $\|\mathbf{c} - \mathbf{c}^*(\mathbf{c})\|$ and the loss $\ell(\mathbf{c}, \mathbf{c}^*(\mathbf{c}))$ is essential in the proof of Theorem 2.1, as it allows us to use chaining arguments as in [Pol82b].

The conditions required in Theorem 2.1 remain hard to check, and cannot be easily interpreted. In fact, Assumption 2.3 and Assumption 2.4 demands that the distribution P is such that a technical inequality is satisfied for every $\mathbf{c}$ in $\mathscr{B}(0, M)$, which cannot be checked in practice. Assumption 2.2 involves second derivatives of the distortion. Consequently, checking Assumption 2.2, even theoretically, remains a hard issue. Theorem 2.2 below offers a more interpretable condition regarding the L_∞-norm of the density f on the boundaries of optimal Voronoi cells, for the distribution P to satisfy Assumption 2.2. We recall that $\mathscr{M}$ denotes the set of all possible optimal codebooks $\mathbf{c}^*$.

Theorem 2.2. *Suppose that Assumption 2.1 is satisfied, $\mathscr{M}$ is finite, and P has a continuous density f. For an optimal codebook $\mathbf{c}^*$, denote by V_i^* the optimal Voronoi cell associated with the code point c_i^*. Let $N^* = \bigcup_{\mathbf{c}^* \in \mathscr{M}, i \neq j} \partial(V_i^* \cap V_j^*)$ denote the union of all possible boundaries of optimal Voronoi cells with respect to all possible optimal codebooks $\mathbf{c}^*$, and denote by Γ the Gamma function. At last, let $f_{|N^*}$ denote the restriction of the function f to the subset N^*, and define $B = \inf_{\mathbf{c}^* \in \mathscr{M}, i \neq j} \|c_i^* - c_j^*\|$.*

Suppose that

$$\|f_{|N^*}\|_\infty < \frac{\Gamma\left(\frac{d}{2}\right) B}{2^{d+3} \pi^{d/2}} \inf_{\mathbf{c}^* \in \mathscr{M}, i=1,\dots,k} P(V_i^*).$$

Then P satisfies Assumption 2.2.

The proof is given in Section 2.5.6. Remark that, for general distributions supported on $\mathscr{B}(0, M)$, we can state a similar Theorem, involving M^{d+1} in the right-hand side of the inequality in Theorem 2.2. Combining Theorem 2.2, Theorem 2.1, and the connections between different sets of conditions leads to the following corollary.

Corollary 2.1. *Suppose that Assumption 2.1 is satisfied and P has a continuous density. Then there exists an explicit constant $\kappa > 0$, depending only on k and d, such that, if $\|f_{|N^*}\|_\infty < \kappa$, then*

$$\mathbb{E}\ell(\hat{\mathbf{c}}_n, \mathbf{c}^*) = \mathcal{O}\left(\frac{1}{n}\right).$$

This corollary emphasizes the idea that, if P is well concentrated around its optimal code points, then some localization conditions can be satisfied and therefore it is a favorable case. The intuition behind this result is given by the extremal case where optimal Voronoi cells boundaries are empty with respect to P. This case is described in detail in Section 2.4. Moreover, the notion of a well-clustered distribution looks like margin-type conditions for the classification case, as described in

[MN06]. This confirms the intuition of an easy-to-quantize distribution, when the poles are well-separated.

Since the location of $\mathbf{c}^*$ are not easy to find for general distributions, the conditions required in Corollary 2.1 are not that simple to satisfy. However, the condition we offer in Theorem 2.2 is valid even if $d \geq 2$, and is not as technical as Assumption 2.2 or Assumption 2.4. Moreover, as will be described in Section 2.4.2, this condition is relevant in the case where P is a mixture distribution, and can be turned into a condition on parameters of the mixture which can be easily inferred from the training sample.

2.4 Examples and discussion

2.4.1 A toy example

In this Subsection we intend to understand which conditions on the density f can guarantee that the Hessian matrices H are positive definite. Some light is shed on the problem by the extremal case in which the density is zero at every boundary of optimal Voronoi cells. Indeed, in this case, equation 2.1 guarantees that the matrices H are diagonal matrices with positive elements, thus positive definite.

The following Proposition offers an intuitive example of such an extremal case.

Proposition 2.2. *Let $z_1, \dots, z_k$ be vectors in $\mathbb{R}^d$. Let ρ be a positive number and $R = \inf_{i \neq j} \|z_i - z_j\|$ be the smallest possible distance between these vectors. Define the triangular function t on $\mathbb{R}^d$ as follows : $t(x_1, \dots, x_d) = (1-r)\mathbf{1}_{r \leq 1}$, where $r = \sqrt{x_1^2 + \dots + x_d^2}$ is the Euclidean norm of x. Then we define the distribution P_ρ and its density f_ρ as follows*

$$f_\rho(x) = \frac{1}{kN_\rho} \sum_{i=1}^{k} t\left(\frac{x - z_i}{\rho}\right),$$

where N_ρ is such that $P_\rho(\mathcal{B}(z_i, \rho)) = 1/k$, for $k = 1, \dots, k$.

Suppose that $\rho < R/2$. If $\left(\frac{R}{2} - 3\rho\right)^2 \geq \frac{2\rho^2 d(d+1)}{(d+2)(d+3)}$, then the optimal k-codebook is $(z_1, \dots, z_k)$.

The proof of Proposition 2.2, which is given in Section 2.5, is inspired from the proof of Step 3 in [BLL98]. It is interesting to note that Proposition 2.2 can be extended to the situation where we assume that the underlying distribution is supported on k small enough subsets. In this context, if each subset has a not too small P-measure, and if the subsets are far enough from each other, it can be proved in the same way that an optimal codebook has a code point in every small subset.

Let us now consider the distribution described in Proposition 2.2, with relevant values for ρ and R. We immediately see that, if $R/2 > \rho$, then every boundary of the Voronoi cells for the optimal codebook lies in a null-measured area. Thus, for this distribution,

$$H(\mathbf{c}^*) = \begin{pmatrix} \frac{1}{k} I_d & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \frac{1}{k} I_d \end{pmatrix},$$

which is clearly positive definite.

This short example illustrates the idea behind Theorem 2.2. Namely, if the density of the distribution is not too big at the boundaries of Voronoi cells associated with every optimal codebook, then the Hessian matrix H will roughly behave as a positive diagonal matrix. In this situation, Pollard's condition (Assumption 2.2) will hopefully be satisfied.

This most favorable case is in fact derived from the special case where the distribution is supported on k points. Here we spread the atoms into small balls to give a density to the distribution and match regularity conditions. It is proved in [AGG05] that, if the distribution has only a finite number of atoms, then the expected distortion $\ell(\hat{\mathbf{c}}_n, \mathbf{c}^*)$ is at most C/n, where C is a constant. Proposition 2.2 guarantees that the convergence rate of $\ell(\hat{\mathbf{c}}_n, \mathbf{c}^*)$ remains $1/n$ when the distribution P we offer in this Proposition is close enough to a distribution supported on k points.

Proposition 2.2 also illustrates the main difficulty in applying Theorem 2.2 : to locate the optimal codebooks $\mathbf{c}^*$. Although some results about geometrical properties of optimal codebooks have been obtained in [Tar95] in the special case where P is a strongly symmetric distribution, or in [Jun12] when k grows to infinity, there are few results about the exact location of optimal codebooks in general.

However, when the probability distribution P has k natural clusters, as in Proposition 2.2, it is possible to give an approximative location of the optimal codebooks of P. The following Section offers another example of such a well-clustered distribution.

2.4.2 Quasi-Gaussian mixture example

The aim of this Subsection is to apply our results to the Gaussian mixtures in dimension $d = 2$. The Gaussian mixture model is a typical and well-defined clustering example. However we will not deal with the clustering issue but rather with its theoretical background.

In general, a Gaussian mixture distribution $\tilde{P}$ is defined by its density

$$\tilde{f}(x) = \sum_{i=1}^{\tilde{k}} \frac{p_i}{2\pi\sqrt{|\Sigma_i|}} e^{-\frac{1}{2}(x-m_i)^t \Sigma_i^{-1}(x-m_i)},$$

where $\tilde{k}$ denotes the number of component of the mixture, and the p_i's denote the weights of the mixture, thus satisfy $\sum_{i=1}^{k} p_i = 1$. Moreover, the m_i's denote the means of the mixture, so that $m_i \in \mathbb{R}^2$, and the Σ_i's are the 2×2 variance matrices of the components.

We restrict ourselves to the case where the number of components $\tilde{k}$ is known, and match the size k of the codebooks. To ease the calculation, we make the additional assumption that every component has the same diagonal variance matrix $\Sigma_i = \sigma^2 I_2$. Note that a similar result to Proposition 2.3 can be derived for distributions with different variance matrices Σ_i, at the cost of more computing.

Since the distribution support of a Gaussian random variable is not bounded, we define the "quasi-Gaussian" mixture model as follows, truncating each Gaussian component. Let the density f of the distribution P be defined by

$$f(x) = \sum_{i=1}^{k} \frac{p_i}{2\pi\sigma^2 N_i} e^{-\frac{\|x-m_i\|^2}{2\sigma^2}} \mathbf{1}_{\mathcal{B}(0,1)},$$

where N_i denotes a normalization constant for each Gaussian variable.

To ensure this model to be close to the Gaussian mixture model, we assume that there exists a constant $\varepsilon \in [0,1]$ such that, for $i = 1,\dots,k$, $N_i \geq 1-\varepsilon$.

Denote by $\tilde{B} = \inf_{i\neq j}\|m_i - m_j\|$ the smallest possible distance between two different means of the mixture. To avoid boundary issues we suppose that, for all $i = 1,\dots,k$, $\mathscr{B}(m_i,\tilde{B}/3) \subset \mathscr{B}(0,1)$.

For such a model, Proposition 2.3 below offers a sufficient condition for P to be well-clustered.

Proposition 2.3. *Denote by $p_{min} = \min_{i=1,\dots,k} p_i$ and $p_{max} = \max_{i=1,\dots,k} p_i$. Suppose that*

$$\frac{p_{min}}{p_{max}} \geq \max\left(\frac{288k\sigma^2}{(1-\varepsilon)\tilde{B}^2(1-e^{-\tilde{B}^2/288\sigma^2})}, \frac{24k}{(1-\varepsilon)\sigma^2\tilde{B}(e^{\tilde{B}/72\sigma^2}-1)}\right).$$

Then P satisfies Assumption 2.2.

The inequality we propose as a condition in Proposition 2.3 can be decomposed as follows. If

$$\frac{p_{min}}{p_{max}} \geq \frac{288k\sigma^2}{(1-\varepsilon)\tilde{B}^2(1-e^{-\tilde{B}^2/288\sigma^2})},$$

then the optimal codebook $\mathbf{c}^*$ is close to the vector of means of the mixture $\mathbf{m} = (m_1,\dots,m_k)$. Knowing that, we can locate the boundaries of Voronoi cells associated with $\mathbf{c}^*$, and apply Theorem 2.2. This leads to the second term of the maximum in Proposition 2.3.

This condition can be interpreted as a condition on the polarization of the mixture. A favorable case for vector quantization seems to be when the poles of the mixtures are well-separated, which is equivalent to σ is small compared to $\tilde{B}$, when considering Gaussian mixtures. Proposition 2.3 just explained how σ has to be small compared to $\tilde{B}$, in order to satisfy Assumption 2.2, and therefore apply Corollary 2.1, for the loss $\ell(\hat{\mathbf{c}}_n,\mathbf{c}^*)$ to reach an improved convergence rate of $1/n$.

Notice that Proposition 2.3 can be considered as an extension of Proposition 2.2. In these two Propositions a key point is to locate $\mathbf{c}^*$, which is possible when the distribution P is well-clustered. The definition of a well-clustered distribution takes two similar forms when looking at Proposition 2.2 or Proposition 2.3. In Proposition 2.2 the good case is when every pole of the distribution is far enough from the other, separated by an empty area with respect to P, which ensures that the Hessian matrices $H(\mathbf{c}^*)$ are positive definite (in this case they are diagonal matrices). When slightly perturbing the framework of Proposition 2.2, it is quite natural to think that the Hessian matrices $H(\mathbf{c}^*)$ should remain positive definite. Proposition 2.3 is an illustration of this idea : the empty separation area between poles is replaced with an area where the density f is small compared to its value around the poles. The condition on σ and $\tilde{B}$ we offer in Proposition 2.3 gives a theoretical definition of a well-clustered distribution for quasi-Gaussian mixtures.

It is important to note that our result is valid when k is known and match exactly the number of components of the mixture. When the number of code points k is different from the number of components $\tilde{k}$ of the mixture, we have no general idea of where the optimal code points can be located.

Moreover, suppose that there is only one optimal codebook $\mathbf{c}^*$, up to re indexation, and that we are able to locate this optimal codebook $\mathbf{c}^*$. As explained in the proof of Proposition 2.3, the quantity at stake is in fact $B = \inf_{i\neq j}\|c_i^* - c_j^*\|$. In the

case where $\tilde{k} \neq k$, there is no simple relation between $\tilde{B}$ and B. Consequently, a condition like in Proposition 2.3 could not involve the natural parameter of the mixture $\tilde{B}$.

It is also worth pointing out that there exist cases where the set of optimal codebooks is not finite. For example, suppose that P is a truncated rotationally symmetric Gaussian distribution, and $k = 2$. Since every rotation of an optimal codebook leads to an other optimal codebook, there exists an infinite set of optimal codebooks. This ensures that at least one Hessian matrix $H(\mathbf{c}^*)$ cannot be positive definite, in fact none is positive definite.

The two assumptions $N_i \geq 1-\varepsilon$ and $\mathscr{B}(m_i, \tilde{B}/3) \subset \mathscr{B}(0,1)$ can easily be satisfied when P is constructed via an homothetic transformation. To see this, take a generic Gaussian mixture on $\mathbb{R}^2$, denote by $\bar{m}_i, i = 1,\dots,k$, its means and by $\bar{\sigma}^2$ its variance. For a given $\varepsilon > 0$, choose $M > 0$ such that, for all $i = 1,\dots,k$, $\int_{\mathscr{B}(0,M)} e^{-\|x-m_i\|^2/2\sigma^2} dx \geq 2\pi\sigma^2(1-\varepsilon)$ and $\mathscr{B}(m_i, \tilde{B}/3) \subset \mathscr{B}(0,M)$. Denote by P_0 the "quasi-Gaussian mixture" we obtain on $\mathscr{B}(0,M)$ for such an M. Then, applying an homothetic transformation with coefficient $1/M$ to P_0 provides a quasi-Gaussian mixture on $\mathscr{B}(0,1)$, with means $m_i = \bar{m}_i/M, i = 1,\dots,k$ and variance $\sigma^2 = \bar{\sigma}^2/M^2$. This distribution satisfies both $N_i \geq 1-\varepsilon$ and $\mathscr{B}(m_i, \tilde{B}/3) \subset \mathscr{B}(0,1)$.

2.5 Proofs

2.5.1 Proof of Proposition 2.1

Half of the equivalences of Proposition 2.1 derives from the following interesting lemma.

Lemma 2.1. *Suppose that Assumption 2.1 is satisfied, $\mathscr{M}$ is finite, and P has a continuous density. Then there exist two constants $C_- > 0$ and $C_+ > 0$ such that*

$$\forall \mathbf{c} \in \mathscr{B}(0,1)^k \quad C_- \|\mathbf{c} - \mathbf{c}^*(\mathbf{c})\|^2 \leq \mathrm{Var}_P\left(\gamma(\mathbf{c},.) - \gamma(\mathbf{c}^*(\mathbf{c}),.)\right) \leq C_+ \|\mathbf{c} - \mathbf{c}^*(\mathbf{c})\|^2.$$

Lemma 2.1 ensures that, if P is smooth enough, $\mathrm{Var}_P(\gamma(\mathbf{c},.) - \gamma(\mathbf{c}^*(\mathbf{c}),.)$ is equivalent to the squared Euclidean distance between $\mathbf{c}$ and $\mathbf{c}^*(\mathbf{c})$, which provides a direct connection between Assumption 2.3 and Assumption 2.4. Notice that we require the density of P to be smooth in order to apply Theorem 1 of [Bad77]. Improving the conditions of this Theorem could be a way to soften the smoothness requirements on P.

It is important to note that Assumption 2.1 too is crucial to make the result of Proposition 2.1 valid, since it allows us to turn local differentiation arguments into global properties. The other half of the equivalences of Proposition 2.1 follows from the following result, exposed in the proof of Corollary1 in [AGG05].

Lemma 2.2. *Suppose that P satisfies Assumption 2.1, then, there exists a constant $A_2 > 0$ such that for a fixed $\mathbf{c}^* \in \mathscr{M}$*

$$\sup_{\mathbf{c} \notin \mathscr{M}} \frac{\mathrm{Var}_P(\gamma(\mathbf{c},.) - \gamma(\mathbf{c}^*,.))}{\|\mathbf{c} - \mathbf{c}^*\|^2} \leq A_2.$$

Moreover, if P satisfies Assumption 2.1 and Assumption 2.2, and if P has a continuous density, then there exists a constant $A_1 > 0$ such that

$$\inf_{\mathbf{c} \notin \mathscr{M}} \frac{\ell(\mathbf{c}, \mathbf{c}^*(\mathbf{c}))}{\|\mathbf{c} - \mathbf{c}^*(\mathbf{c})\|^2} \geq A_1.$$

Equipped with these two lemmas, we are now in position to prove Proposition 2.1. It may be pointed out that, if P satisfies Assumption 2.2 and Assumption 2.1, then $\mathscr{M}$ is finite (see, e.g., [AGG05]). If not, due to a compactness argument it can be proved that $\mathscr{M}$ has an accumulation point, which ensures that the Hessian matrix H at this accumulation point cannot be positive definite.

Now suppose that P has a continuous density. Then, according to Lemma C in [Pol82b], $P\gamma(\mathbf{c},.)$ is differentiable twice at every point $\mathbf{c}$. Furthermore, if $\mathscr{M}$ is finite and if $\ell(\mathbf{c},\mathbf{c}^*(\mathbf{c})) \geq A_1\|\mathbf{c}-\mathbf{c}^*(\mathbf{c})\|^2$, then the Hessian matrices $H(\mathbf{c}^*)$ have to be positive definite for every $\mathbf{c}^*$ in $\mathscr{M}$. This leads to the following equivalence :

$$\left\{\begin{array}{c}\text{Assumption 2.1}\\ \text{Assumption 2.2}\end{array}\right. \Leftrightarrow \left\{\begin{array}{c}f \text{ has a continuous density}\\ \mathscr{M} \text{ is finite}\\ \text{Assumption 2.1}\\ \text{Assumption 2.3}\end{array}\right. .$$

The other equivalence relies on Lemma 2.1. Since Assumption 2.3 obviously implies Assumption 2.4, the direct implication is proved. Now suppose that P has a continuous density, $\mathscr{M}$ is finite, and satisfies Assumption 2.1 and Assumption 2.4. Since Assumption 2.1 is satisfied and $\mathscr{M}$ is finite, the first part of Lemma 2.2 provides us with a global $A_2 > 0$ such that

$$\mathrm{Var}_P(\gamma(\mathbf{c},.)-\gamma(\mathbf{c}^*(\mathbf{c}),.)) \leq A_2\|\mathbf{c}-\mathbf{c}^*\|^2.$$

Combining Lemma 2.1 with Assumption 2.4 ensures the existence of $A_1 > 0$ such that

$$\ell(\mathbf{c},\mathbf{c}^*(\mathbf{c})) \geq A_1\|\mathbf{c}-\mathbf{c}^*(\mathbf{c})\|^2.$$

Then, we can deduce that

$$\left\{\begin{array}{c}f \text{ has a continuous density}\\ \mathscr{M} \text{ is finite}\\ \text{Assumption 2.1}\\ \text{Assumption 2.3}\end{array}\right. \Leftrightarrow \left\{\begin{array}{c}f \text{ has a continuous density}\\ \mathscr{M} \text{ is finite}\\ \text{Assumption 2.1}\\ \text{Assumption 2.4}\end{array}\right. .$$

2.5.2 Proof of Lemma 2.1

Lemma 2.1 relies on the following technical lemma, which provides some differentiation arguments in order to connect $\mathrm{Var}_P(\gamma(\mathbf{c},.)-\gamma(\mathbf{c}^*,.))$ to the squared Euclidean distance $\|\mathbf{c}-\mathbf{c}^*\|^2$.

Lemma 2.3. *Let $\mathbf{c}^* \in \mathscr{M}$ be fixed. Let f be the real-valued function defined by*

$$f : \left\{\begin{array}{ccc}(\mathbb{R}^d)^k & \longrightarrow & \mathbb{R}\\ \mathbf{c} & \longmapsto & \mathrm{Var}_P(\gamma(\mathbf{c},.)-\gamma(\mathbf{c}^*,.))\end{array}\right. .$$

Then f is differentiable twice at the point $\mathbf{c}=\mathbf{c}^*$, and its Hessian matrix F is made of the following $d \times d$ blocks

$$F_{i,j} = \left\{\begin{array}{ccc}8\int_{V_i^*} f(x)(x-c_j^*)(x-c_j^*)^t & \text{if} & i=j\\ 0 & \text{if} & i\neq j\end{array}\right. \qquad i,j=1,\ldots,k.$$

Furthermore, the matrix F is positive definite.

Proof of Lemma 2.3. First we write

$$f(\mathbf{c}) = P\left(\gamma(\mathbf{c},.) - \gamma(\mathbf{c}^*,.)\right)^2 - \left(P\left(\gamma(\mathbf{c},.) - \gamma(\mathbf{c}^*,.)\right)^2\right).$$

Using almost the same argument as in Lemma A of [Pol82b], f is differentiable at every point $\mathbf{c}$ in $\mathscr{B}(0,1)^d$, with gradient

$$\begin{aligned}\nabla f(\mathbf{c}) &= 2P\left(\Delta(\mathbf{c},.)(\gamma(\mathbf{c},.) - \gamma(\mathbf{c}^*,.))\right) - 2P\Delta(\mathbf{c},.)P(\gamma(\mathbf{c},.) - \gamma(\mathbf{c}^*,.)) \\ &:= 2g_1(\mathbf{c}) - 2g_2(\mathbf{c}),\end{aligned}$$

where $\Delta(\mathbf{c},x)$ is the point wise gradient function defined as in [Pol82b]

$$\Delta(\mathbf{c},x) = -2((x-c_1)\mathbf{1}_{V_1},\ldots,(x-c_k)\mathbf{1}_{V_k}).$$

First we deal with g_1. Writing

$$g_1(\mathbf{c}) = \left(-2\int_{\mathbb{R}^d}\left(\|x-c_i\|^2 - \gamma(\mathbf{c}^*,x)\right)(x-c_i)\mathbf{1}_{V_i}(x)\right)_{i=1,\ldots,k},$$

we use Theorem 1 of [Bad77] to prove that g_1 is differentiable at every $\mathbf{c}$, with derivatives matrix denoted by H_1. Through computation like in Lemma C of [Pol82b], we get the following decomposition in $d\times d$ blocks for H_1 :

$$\begin{aligned}H_1(\mathbf{c})_{i,i} = 4\int_{V_i} f(x)(x-c_i)(x-c_i)^t dx + 2\int_{V_j} f(x)\left(\gamma(\mathbf{c},x) - \gamma(\mathbf{c}^*,x)\right)dx \\ - 2\sum_{p\neq i}\|c_i - \mathbf{c}_j\|^{-1}\int_{\partial(V_i\cap V_p)} f(x)\left(\gamma(\mathbf{c},x) - \gamma(\mathbf{c}^*,x)\right)(x-c_i)(x-c_i)^t dx,\end{aligned}$$

for diagonal blocks. For other blocks we have, with $i\neq j$,

$$H_1(\mathbf{c})_{i,j} = 2\|c_i - c_j\|^{-1}\int_{\partial(V_i\cap V_j)} f(x)\left(\gamma(\mathbf{c},x) - \gamma(\mathbf{c}^*,x)\right)(x-c_i)(x-c_j)^t dx.$$

Using the same argument (see, e.g., Theorem 1 of [Bad77]), g_2 is differentiable. Recalling that H denotes the Hessian matrix of $\mathbf{c}\longmapsto P\gamma(\mathbf{c},.)$, elementary calculation shows that, if H_2 denotes the matrix of derivatives of g_2,

$$H_2(\mathbf{c}) = P(\gamma(\mathbf{c},.) - \gamma(\mathbf{c}^*,.))H(\mathbf{c}) + P\Delta(\mathbf{c},.)(P\Delta(\mathbf{c},.))^t.$$

Hence we deduce that f is differentiable twice at point $\mathbf{c} = \mathbf{c}^*$, with Hessian matrix $F(\mathbf{c}^*) = H_1(\mathbf{c}^*) + H_2(\mathbf{c}^*)$. Taking $\mathbf{c} = \mathbf{c}^*$ in the above calculations leads to the result for the expression of F.

It remains to prove that F is positive definite. Let $\mathbf{h} = (h_1,\ldots,h_k)$ be a $k\times d$ vector, with $\mathbf{h}\neq 0$. We notice that

$$\mathbf{h}^t F\mathbf{h} = \sum_{i=1}^k h_i^t F_{i,i}h_i = 8\sum_{i=1}^k\int_{V_i^*} f(x)\langle h, x-c_i^*\rangle^2 dx.$$

Suppose that $h^tFh = 0$, then, for $i = 1,\ldots,k$, $h_i^tF_{i,i}h_i = 0$. Since $\mathbf{h}\neq 0$, we can assume without loss of generality that $h_1\neq 0$. We denote by $h_1^\perp$ the hyperplane in $\mathbb{R}^d$ orthogonal to h_1. Since $h_1^tF_{1,1}h_1 = 0$, we deduce that $P(V_1^*\setminus(c_1^* + h_1^\perp)) = 0$. Taking into account that P has a density, we get $P(V_1^*) = 0$, which is impossible, according to Theorem 4.1 of [GL00]. □

Now we turn to the proof of Lemma 2.1. Suppose that P has a continuous density f, $\mathscr{M}$ is finite, and P satisfies Assumption 2.1. Since P satisfies Assumption 2.1 and $\mathscr{M}$ is finite, the first part of Lemma 2.2 provides $C_+ > 0$ such that

$$\mathrm{Var}_P(\gamma(\mathbf{c},.) - \gamma(\mathbf{c}^*(\mathbf{c}),.)) \leq C_+ \|\mathbf{c} - \mathbf{c}^*\|^2.$$

Consequently, we only have to deal with the lower bound. To do this, suppose that P is such that

$$\inf_{\mathbf{c} \notin \mathscr{M}} \frac{\mathrm{Var}_P(\gamma(\mathbf{c},.) - \gamma(\mathbf{c}^*(\mathbf{c}),.))}{\|\mathbf{c} - \mathbf{c}^*(\mathbf{c})\|^2} = 0,$$

then there exists a sequence $(\mathbf{c}_n)_{n\geq 1}$, such that $\mathbf{c}_n \notin \mathscr{M}$ and

$$\frac{\mathrm{Var}_P(\gamma(\mathbf{c}_n,.) - \gamma(\mathbf{c}^*(\mathbf{c}_n),.))}{\|\mathbf{c}_n - \mathbf{c}^*(\mathbf{c}_n)\|^2} \longrightarrow 0,$$

as $n \longrightarrow \infty$. Since Assumption 2.1 is satisfied, we can assume without loss of generality that there exists $\mathbf{c}$ in $\mathscr{B}(0,1)^d$, such that $\mathbf{c}_n \longrightarrow \mathbf{c}$.

We have to prove that $\mathbf{c} \in \mathscr{M}$. To do this, suppose that $\mathbf{c} \notin \mathscr{M}$, and denote by $\mathbf{c}^*$ the closest optimal codebook to $\mathbf{c}$. Since $\mathbf{c} \notin \mathscr{M}$, there exists i such that $c_i^* \neq c_j$, for $j = 1, \ldots, k$. Furthermore, since $\mathscr{M}$ is a finite set, $\mathbf{c}^*(\mathbf{c}_n) = \mathbf{c}^*$ for n large enough. Since $\mathrm{Var}_P(\gamma(\mathbf{c}_n,.) - \gamma(\mathbf{c}^*(\mathbf{c}_n),.)) \longrightarrow 0$, we deduce that $\mathrm{Var}_P(\gamma(\mathbf{c},.) - \gamma(\mathbf{c}^*,.)) = 0$, which in turn leads to $\gamma(\mathbf{c},x) = \gamma(\mathbf{c}^*,x) + a$, , for a constant $a > 0$, P-almost surely in x. Let $x \in V_i^*$. We denote by G and G_j the following sets of points :

$$\begin{cases} G &= \left\{x \in V_i^* | \quad \gamma(\mathbf{c},x) = \gamma(\mathbf{c}^*,x) + a\right\}, \\ G_j &= \left\{x \in V_i^* | \quad \|x - c_j\|^2 = \|x - c_i^*\|^2 + a\right\}. \end{cases}$$

Formally we have $G \subset \bigcup_{j=1}^k G_j$, and $P(V_i^*) = P(G)$. However, since G_j is an affine space with dimension $d-1$ and P has a density, it follows that $P(G_j) = 0$ for all $j = 1, \ldots, k$. Hence we deduce that $P(G) = 0$, so that $P(V_i^*) = 0$, which is not possible for an optimal codebook $\mathbf{c}^*$ (see, e.g., Theorem 4.1 of [GL00]). Hence we deduce that $\mathbf{c} \in \mathscr{M}$.

Then we can assume that $\mathbf{c}_n \longrightarrow \mathbf{c}^*$, for some fixed $\mathbf{c}^* \in \mathscr{M}$, and, since $\mathscr{M}$ is a finite set, without loss of generality, $\mathbf{c}^*(\mathbf{c}_n) = \mathbf{c}^*$ for $n \geq 1$. According to Lemma 2.3, there exists $C_+ > 0$ such that

$$\mathrm{Var}_P(\gamma(\mathbf{c}_n,.) - \gamma(\mathbf{c}^*,.)) \geq C_+ \|\mathbf{c}_n - \mathbf{c}^*\|^2 + o(\|\mathbf{c}_n - \mathbf{c}^*\|^2),$$

and so

$$\frac{\mathrm{Var}_P(\gamma(\mathbf{c}_n,.) - \mathbf{c}^*(\mathbf{c}_n))}{\|\mathbf{c}_n - \mathbf{c}^*(\mathbf{c}_n)\|^2} \geq C_+ + o(1),$$

which leads to a contradiction.

2.5.3 Proof of Theorem 2.1

The proof strongly relies on the localization principle and its application, as exposed in [BBM08]. We start with the following definition.

Definition 2.1. *Let Φ be a real-valued function. Φ is called a sub-α function if and only if Φ is non-decreasing and the map $x \mapsto \Phi(x)/x^\alpha$ is non-increasing.*

The next Theorem is an adaptation of Theorem 6.1 in [BBM08]. For the sake of clarity its proof is given in Subsection 2.5.4.

Theorem 2.3. *Let $\mathscr{F}$ be a class of bounded measurable functions such that there exist $b>0$ and $\omega:\mathscr{F}\longrightarrow\mathbb{R}^+$ satisfying*

(i) $\forall f\in\mathscr{F}\quad \|f\|_\infty\le b$,

(ii) $\forall f\in\mathscr{F}\quad \mathrm{Var}_P(f)\le\omega(f)$.

Let K be a positive constant, Φ a sub-α function, $\alpha\in[1/2,1[$. Then there exists a constant $C(\alpha)$ such that, if D is a constant satisfying $D\le 6KC(\alpha)$, and δ^* is the unique solution of the equation $\Phi(\delta)=\delta/D$, the following holds. Assume that

$$\forall\delta\ge\delta^*\qquad \mathbb{E}\left(\sup_{\omega(f)\le\delta}|(P-P_n)f|\right)\le\Phi(\delta).$$

Then, for all $x>0$, with probability larger than $1-e^{-x}$,

$$\forall f\in\mathscr{F}\quad Pf-P_nf\le K^{-1}\left(\omega(f)+\left(\frac{6KC(\alpha)}{D}\right)^{\frac{1}{1-\alpha}}\delta^*+\frac{(9K^2+16Kb)x}{4n}\right).$$

As explained in the proof, an optimal choice for $C(\alpha)$ is

$$C(\alpha)=\inf_{x>1}\left(1+x^\alpha\left(\frac{1}{2}+\frac{1}{x^{1-\alpha}-1}\right)\right).$$

This Theorem provides a sharp concentration inequality in the case where it is possible to control the maximal deviation between P and P_n over a set of functions whose variance with respect to P is constrained within a ball. The main point is to find a suitable control function for the variance of the process. Here the interesting set is

$$\mathscr{F}=\left\{\gamma(\mathbf{c},.)-\gamma(\mathbf{c}^*,.),\mathbf{c}\in\mathscr{B}(0,1)^k,\mathbf{c}^*\in\mathscr{M}\right\}.$$

Since P satisfies Assumption 2.3, the relevant control function for the variance of the process $\gamma(\mathbf{c},.)-\gamma(\mathbf{c}^*,.)$ is $\omega(\mathbf{c},\mathbf{c}^*)=A_2\|\mathbf{c}-\mathbf{c}^*\|^2$, where A_2 is defined in Assumption 2.3.

Thus it remains to bound from above the quantity

$$\mathbb{E}\left(\sup_{\mathbf{c}^*\in\mathscr{M},A_2\|\mathbf{c}-\mathbf{c}^*\|^2\le\delta}\left|(P_n-P)(\gamma(\mathbf{c}^*,.)-\gamma(\mathbf{c},.))\right|\right).$$

This is done in the following Proposition.

Proposition 2.4. *Suppose that P has a density and satisfies Assumption 2.1. Furthermore we assume that $\mathscr{M}$ is finite. Then*

$$\mathbb{E}\left(\sup_{\mathbf{c}^*\in\mathscr{M},A_2\|\mathbf{c}-\mathbf{c}^*\|^2\le\delta}\left|(P_n-P)(\gamma(\mathbf{c}^*,.)-\gamma(\mathbf{c},.))\right|\right)\le\sqrt{\delta}\frac{\Xi}{\sqrt{n}},$$

where Ξ is a constant depending on k, d, and P.

Since we assume that $\mathscr{M}$ is finite, P has a density and Assumption 2.3 is satisfied, we can apply Theorem 2.3, with $\omega(\mathbf{c},\mathbf{c}^*)=A_2\|\mathbf{c}-\mathbf{c}^*\|^2$, $b=8$, and $\Phi(\delta)=\sqrt{\delta}\Xi/\sqrt{n}$. Noticing that the solution of the equation $\delta=\Phi(\delta)/D$ is Ξ^2D^2/n, for an arbitrary $D>0$, we get the following result.

Lemma 2.4. *Suppose that P has a density, satisfies Assumption 2.1 and Assumption 2.3, and $\mathscr{M}$ is finite. Let $D > 0$. For all $\mathbf{c}^* \in \mathscr{M}$, $x > 0$ and $D \leq 6KC(1/2)$, we have, with probability larger than $1 - e^{-x}$,*

$$(P - P_n)(\gamma(\mathbf{c},.) - \gamma(\mathbf{c}^*,.)) \leq K^{-1}A_2\|\mathbf{c} - \mathbf{c}^*\|^2 + \frac{36KC(1/2)^2\Xi^2}{n} + \frac{2K+32}{n}x.$$

Take $\mathbf{c}^* = \mathbf{c}^*(\mathbf{c})$, a nearest optimal codebook to $\mathbf{c}$, and use (H2) to connect $\|\mathbf{c} - \mathbf{c}^*(\mathbf{c})\|^2$ to $\ell(\mathbf{c}, \mathbf{c}^*(\mathbf{c}))$. Choosing $K = 2A_1A_2$, $D = 6KC(1/2)$, we get, with probability larger than $1 - e^{-x}$,

$$1/2(P - P_n)(\gamma(\hat{\mathbf{c}}_n,.) - \gamma(\mathbf{c}^*(\hat{\mathbf{c}}_n),.)) \leq \frac{C_1}{n} + \frac{C_2}{n}x,$$

for some constants $C_1 > 0$ and $C_2 > 0$. Since $P_n(\gamma(\hat{\mathbf{c}}_n,.) - \gamma(\mathbf{c}^*(\hat{\mathbf{c}}_n),.)) \leq 0$, taking expectation leads to, for all $\mathbf{c}^* \in \mathscr{M}$,

$$\mathbb{E}\ell(\hat{\mathbf{c}}_n, \mathbf{c}^*) \leq \frac{C_0}{n},$$

for a constant $C_0 > 0$ depending only on k, d, and P.

2.5.4 Proof of Theorem 2.3

This proof is a modification of the proof of Theorem 6.1 in [BBM08]. For $\delta \geq 0$, set

$$\Omega_\delta = \sup_{f \in \mathscr{F}} \left| (P - P_n)\frac{f}{\omega(f) + \delta} \right|.$$

We start with a modified version of the so-called peeling lemma :

Lemma 2.5. *Under the assumptions of Theorem 2.3, there exists a constant $C(\alpha)$ depending only on α such that, for all $\delta > 0$,*

$$\mathbb{E}(\Omega_\delta) \leq C(\alpha)\frac{\Phi(\delta)}{\delta}.$$

Furthermore, we have $C(\alpha) \underset{\alpha \to 1}{\longrightarrow} \infty$.

Proof of Lemma 2.5. Let $x > 1$ be a real number. We may write

$$\sup_{f \in \mathscr{F}} \left| (P - P_n)\frac{f}{\omega(f) + \delta} \right| \leq \sup_{\omega(f) \leq \delta} \left| (P - P_n)\frac{f}{\omega(f) + \delta} \right| + \sum_{k \geq 0} \sup_{\delta x^k < \omega(f) \leq \delta x^{k+1}} \left| (P - P_n)\frac{f}{\omega(f) + \delta} \right|.$$

Since $\sup_{\delta x^k < \omega(f) \leq \delta x^{k+1}} |(P - P_n)f| \geq 0$, and $\omega(f) + \delta > 0$, taking expectation on both sides leads to

$$\mathbb{E}(\Omega_\delta) \leq \frac{\Phi(\delta)}{\delta} + \sum_{k \geq 0} \frac{\Phi(\delta x^{k+1})}{\delta(1 + x^k)}.$$

Recalling that Φ is a sub-α function, we may write $\Phi(\delta x^{k+1}) \leq x^{\alpha(k+1)}\Phi(\delta)$. Hence we get

$$\begin{aligned} \mathbb{E}(\Omega_\delta) &\leq \frac{\Phi(\delta)}{\delta} + \frac{\Phi(\delta)}{\delta} \sum_{k \geq 0} \frac{x^{\alpha(k+1)}}{1 + x^k} \\ &\leq \frac{\Phi(\delta)}{\delta}\left(1 + x^\alpha\left(\frac{1}{2} + \frac{1}{x^{1-\alpha} - 1}\right)\right). \end{aligned}$$

Taking $C(\alpha) = \inf_{x > 1}\left(1 + x^\alpha\left(\frac{1}{2} + \frac{1}{x^{1-\alpha}-1}\right)\right)$ proves the result. □

We are now in a position to prove Theorem 2.3. Using the Talagrand's inequality for a supremum of bounded variables offered in [Bou02], we have, with probability larger than $1-e^{-x}$,

$$\Omega_\delta \le \mathbb{E}(\Omega_\delta) + \sqrt{\frac{x}{2\delta n}} + 2\sqrt{\frac{xb\mathbb{E}(\Omega_\delta)}{n\delta}} + \frac{bx}{3\delta n}.$$

Using Lemma 2.5 and the inequality $2ab \le a^2+b^2$,

$$\Omega_\delta \le \frac{2C(\alpha)\Phi(\delta)}{\delta} + \sqrt{\frac{x}{2\delta n}} + \frac{4}{3}\frac{bx}{\delta n}.$$

Let δ^* be the solution of $\Phi(\delta) = \frac{\delta}{D}$. If $\delta \ge \delta^*$, then $\frac{\Phi(\delta)}{\delta} \le \left(\frac{\delta^*}{\delta}\right)^{1-\alpha}\frac{1}{D}$. For such an δ we have

$$\Omega_\delta \le \beta_1\delta^{-(1-\alpha)} + \beta_2\delta^{-1/2} + \beta_3\delta^{-1},$$

with

$$\begin{cases} \beta_1 &= \frac{2C(\alpha)(\delta^*)^{1-\alpha}}{D}, \\ \beta_2 &= \sqrt{\frac{x}{2n}}, \\ \beta_3 &= \frac{4bx}{3n}. \end{cases}$$

We want to find a suitable δ such that $\delta \ge \delta^*$ and $\Omega_\delta \le 1/K$. To this aim, it suffices to see that if $\delta \ge (3K\beta_1)^{\frac{1}{1-\alpha}} + (3K\beta_2)^2 + 3K\beta_3$, and $\delta \ge \delta^*$, then $\Omega_\delta \le 1/K$ using the previous upper bound on Ω_δ.

It remains to check that the condition $(3K\beta_1)^{\frac{1}{1-\alpha}} + (3K\beta_2)^2 + 3K\beta_3 \ge \delta^*$ holds. To see this just recall that

$$(3K\beta_1)^{\frac{1}{1-\alpha}} = \delta^* \times \left(\frac{6KC(\alpha)}{D}\right)^{\frac{1}{1-\alpha}}.$$

Thus, we deduce that, if $D \le 6KC(\alpha)$, the choice $\delta = (3K\beta_1)^{\frac{1}{1-\alpha}} + (3K\beta_2)^2 + 3K\beta_3$ guarantees $\Omega_\delta \le K^{-1}$ and, consequently, with probability larger than $1-e^{-x}$,

$$\begin{aligned} Pf - P_n f &\le |(P-P_n)f| \\ &\le \left|(P-P_n)\frac{f}{\omega(f)+\delta^*}\right| \times (\omega(f)+\delta^*) \\ &\le \Omega_{\delta^*}(\omega(f)+\delta^*) \\ &\le \frac{1}{K}\left(\omega(f) + \left(\frac{6KC(\alpha)}{D}\right)^{\frac{1}{1-\alpha}}\delta^* + \frac{(9K^2+16Kb)x}{4n}\right). \end{aligned}$$

2.5.5 Proof of Proposition 2.4

Following the approach of [Pol82b], we notice that, for any $\mathbf{c} \in (\mathbb{R}^d)^k$ and $\mathbf{c}^* \in \mathscr{M}$, P-almost surely in x,

$$\gamma(\mathbf{c},x) = \gamma(\mathbf{c}^*,x) + \left\langle \mathbf{c}-\mathbf{c}^*, \Delta(\mathbf{c}^*,x)\right\rangle + \|\mathbf{c}-\mathbf{c}^*\| R(\mathbf{c}^*,\mathbf{c}-\mathbf{c}^*,x),$$

where, with use of the notation in [Pol82b],

$$\begin{cases} \Delta(\mathbf{c}^*,x) = -2((x-c_1^*)\mathbf{1}_{V_1^*},\dots,(x-c_k^*)\mathbf{1}_{V_k^*}), \\ R(\mathbf{c}^*,\mathbf{c}-\mathbf{c}^*,x) = \sum_{i,j=1,\dots,k} \mathbf{1}_{V_i^*}\mathbf{1}_{V_j}\|\mathbf{c}-\mathbf{c}^*\|^{-1}\left[2(c_i-c_j)^t x + \|c_i^*\|^2 \right. \\ \qquad\qquad\qquad\qquad\qquad\qquad \left. -2(c_i^*)^t c_i + \|c_j\|^2\right]. \end{cases}$$

We recall that V_i^* denotes the Voronoi cell associated with the code point $\mathbf{c}_i^*$, where c_i^* is a coordinate of $\mathbf{c}^*$, and that $\mathbf{1}_{V_i^*}(x)$ takes the value 1 if $x \in V_i^*$, 0 elsewhere.

Splitting the expectation in two parts, we obtain

$$\begin{aligned}
&\mathbb{E}\left(\sup_{\mathbf{c}^*\in\mathscr{M},\|\mathbf{c}-\mathbf{c}^*\|^2\le\delta}\left|(P_n-P)(\gamma(\mathbf{c}^*,.)-\gamma(\mathbf{c},.))\right|\right)\\
&\qquad\le \mathbb{E}\left(\sup_{\mathbf{c}^*\in\mathscr{M},\|\mathbf{c}-\mathbf{c}^*\|^2\le\delta}\left|(P_n-P)\left\langle-(\mathbf{c}-\mathbf{c}^*),\Delta(\mathbf{c}^*,.)\right\rangle\right|\right)\\
&\qquad+\sqrt{\delta}\,\mathbb{E}\left(\sup_{\mathbf{c}^*\in\mathscr{M},\|\mathbf{c}-\mathbf{c}^*\|^2\le\delta}\left|(P_n-P)(-R(\mathbf{c}^*,\mathbf{c}-\mathbf{c}^*,.))\right|\right)\\
&\qquad:= A+B. \qquad (2.2)
\end{aligned}$$

2.5.5.1 Term A : Complexity of the model

Term A in inequality (2.2) is at first sight the dominant term in the expression $\Phi(\delta)$. The upper bound we obtain below is rather accurate, due to the finite-dimensional Euclidean space structure. Indeed, we have to bound a scalar product when the vectors are contained in a ball, thus it is easy to see that the largest value of the product matches in fact the largest value of the coordinates of the gradient term. We recall that $\mathscr{M}$ denotes the finite set of optimal codebooks. Let $\mathbf{x}=(x_1,\dots,x_k)$ be a vector in $(\mathbb{R}^d)^k$. We denote by x_{j_r} the r-th coordinate of x_j, and name it the (j,r)-th coordinate of $\mathbf{x}$. Moreover, denote by $e_{j,r}$ the vector whose (j,r)-th coordinate is 1, and other coordinates are 0.

Taking into account that every $\mathbf{c}^*$ in $\mathscr{M}$ satisfies the centroid condition, that is $P\Delta(\mathbf{c}^*,.)=0$, we may write

$$\begin{aligned}
&\sup_{\mathbf{c}^*\in\mathscr{M},\|\mathbf{c}-\mathbf{c}^*\|\le\sqrt{\delta}}\left|\left\langle\mathbf{c}-\mathbf{c}^*,(P_n-P)(-\Delta(\mathbf{c}^*))\right\rangle\right|\\
&\qquad=\sup_{\mathbf{c}^*\in\mathscr{M},j=1,\dots,k,r=1,\dots,d}\left|\frac{1}{n}\sum_{i=1}^{n}(X_i-c_j^*)\mathbf{1}_{V_j^*}(X_i)\right|_r\times\sqrt{\delta}.\\
&\qquad=\sup_{\mathbf{c}^*\in\mathscr{M},j=1,\dots,k,r=1,\dots,d,\varepsilon=\pm1}\left\langle\varepsilon\sqrt{\delta}e_{j,r},P_n(\Delta(\mathbf{c}^*,.))\right\rangle
\end{aligned}$$

Therefore we can reduce the set of $\mathbf{c}$'s and $\mathbf{c}^*$'s of interest to a finite set we denote by $\mathscr{H}_{\mathscr{M}}$, which contains $|\mathscr{M}|2^{kd}$ elements. Taking into account that, for every $\mathbf{c}^*$ in $\mathscr{M}$, $P\Delta(\mathbf{c}^*,.)=0$, and that, for every fixed $\mathbf{c}$ and $\mathbf{c}^*$, the quantity $\langle\mathbf{c}-\mathbf{c}^*,P_n(\Delta(\mathbf{c}^*,.))\rangle$ is a sub-Gaussian random variable with variance $16\delta/n$, we get, by a maximal inequality (see, e.g., Lemma 2.3 in [Mas07]) :

$$\begin{aligned}
&\mathbb{E}\left(\sup_{\mathbf{c}^*\in\mathscr{M},\|\mathbf{c}-\mathbf{c}^*\|^2\le\delta}\left|(P_n-P)\left\langle-(\mathbf{c}-\mathbf{c}^*),\Delta(\mathbf{c}^*,.)\right\rangle\right|\right)\\
&\qquad=\mathbb{E}\left(\sup_{(\mathbf{c},\mathbf{c}^*)\in\mathscr{H}_{\mathscr{M}}}\left|(P_n-P)\left\langle-(\mathbf{c}-\mathbf{c}^*),\Delta(\mathbf{c}^*,.)\right\rangle\right|\right)\\
&\qquad\le\sqrt{2\frac{16\delta}{n}\log(|\mathscr{H}_{\mathscr{M}}|)}\\
&\qquad\le 4\sqrt{2kd\log(2|\mathscr{M}|)}\frac{\sqrt{\delta}}{\sqrt{n}}.
\end{aligned}$$

Therefore, the expected dominant term involves the complexity of the model in a way which is proportional to the square root of the complexity. In our case, this complexity is the dimension of the codebook space.

2.5.5.2 Bound on B

To bound the second term in inequality (2.2), we follow the approach of [Pol82b], using complexity arguments such as Dudley's entropy integral.

Let $\mathscr{F}$ be a set of functions defined on $\mathscr{X}$ with envelope F. Let $S \subset \mathscr{X}$ be a finite set and f a function. We denote $\|f\|_{l^2(S)} = \left(1/n \sum_{x \in S} f^2(x)\right)^{1/2}$, where $n = |S|$, and by $N_F(\varepsilon, S, \mathscr{F})$ the smallest integer m such that there exist $\phi_1, \dots, \phi_m$, m functions on $\mathscr{X}$ satisfying $\min_{i=1,\dots,m} \|f - \phi_i\|^2_{l^2(S)} \leq \varepsilon^2 \|F\|^2_{l^2(S)}$. Let also $H(\varepsilon)$ be defined by $H(\varepsilon) = \sup_{|S|<\infty} \log N_F(\varepsilon, S, \mathscr{F})$, and $m(\varepsilon) = e^{H(\varepsilon)}$, so that for any subset $S \subset \mathscr{X}$ there exists a $\varepsilon \|F\|_{l^2(S)}$-chaining of $\mathscr{F}$ with at most $m(\varepsilon)$ elements.

It is proved in [Pol82b], using a result proposed in Theorem 9 of [Pol82a], that, for the class of functions

$$\mathscr{F} = \left\{ R(., \mathbf{c}^*, \mathbf{c} - \mathbf{c}^*), \mathbf{c}^* \in \mathscr{M}, \mathbf{c} \in \mathscr{B}(0,1)^k \right\},$$

there exist $C > 0$ depending on k and d such that $F(x) = C(1 + \|x\|)$ is an envelope for $\mathscr{F}$. Furthermore, for this envelope, we have

$$H(\varepsilon) \leq \log(A) - W \log(\varepsilon),$$

where A is a positive constant, and W depends only on the pseudo-dimension of $\mathscr{F}$. We will use a classical chaining argument to bound term B. Let $\tilde{\mathbf{c}}$ denote the pair $(\mathbf{c}, \mathbf{c}^*) \in (\mathscr{B}(0,1))^k \times \mathscr{M}$. For practical, let $f_{\tilde{\mathbf{c}}}$ denote the function $R(., \mathbf{c}^*, \mathbf{c} - \mathbf{c}^*)$. We set $\varepsilon_0 = 1$ and $\varepsilon_j = 2^{-j} \varepsilon_0$.

Let $X_1, \dots, X_n$ be fixed, and denote by S_n the random set $\{X_1, \dots, X_n\}$. For any $f_{\tilde{\mathbf{c}}}$, let $f_{\tilde{\mathbf{c}}_j}$ be a function such that $\|f_{\tilde{\mathbf{c}}} - f_{\tilde{\mathbf{c}}_j}\|^2_{l^2(S_n)} \leq \varepsilon_j^2 \|F\|^2_{l^2(S_n)}$. Making use of the result of [Pol82b] mentioned above, we may write $\left| \left\{ f_{\tilde{\mathbf{c}}_j} | \mathbf{c} \in \mathscr{B}(0,1)^k, \mathbf{c}^* \in \mathscr{M} \right\} \right| \leq m(\varepsilon_j) \leq A \varepsilon_j^{-W}$.

Since Assumption 2.1 holds, F is bounded from above by a constant C_F. By dominated convergence Theorem we have $f_{\tilde{\mathbf{c}}_j} \xrightarrow[j \to \infty]{L^1, a.s} f_{\tilde{\mathbf{c}}}$, and thus

$$(P_n - P) f_{\tilde{\mathbf{c}}} = (P_n - P) f_{\tilde{\mathbf{c}}_0} + \sum_{j=1}^{\infty} (P_n - P)(f_{\tilde{\mathbf{c}}_j} - f_{\tilde{\mathbf{c}}_{j-1}}).$$

Therefore

$$\mathbb{E}\left(\sup_{\mathbf{c}^* \in \mathscr{M}, \|\mathbf{c} - \mathbf{c}^*\| \leq \sqrt{\delta}} |(P_n - P) f_{\tilde{\mathbf{c}}}| \right) \leq \mathbb{E}\left(\sup_{\mathbf{c}^* \in \mathscr{M}, \|\mathbf{c} - \mathbf{c}^*\| \leq \sqrt{\delta}} \left| (P_n - P) f_{\tilde{\mathbf{c}}_0} \right| \right) + \sum_{j>0} \mathbb{E}\left(\sup_{\mathbf{c}^* \in \mathscr{M}, \|\mathbf{c} - \mathbf{c}^*\| \leq \sqrt{\delta}} \left| (P_n - P)(f_{\tilde{\mathbf{c}}_j} - f_{\tilde{\mathbf{c}}_{j-1}}) \right| \right).$$

Here we use a symmetrization inequality to bound from above the last term with a Rademacher complexity. Symmetrization inequalities were introduced in [GZ84], however we rather use the approach developed in Section 2.2 of [Kol06]. In fact,

introducing some Rademacher random variables σ ($\sigma = \pm 1$ with probability 1/2), we get, for the first term :

$$\begin{aligned}
\mathbb{E}\left(\sup_{\mathbf{c}^*\in\mathscr{M},\|\mathbf{c}-\mathbf{c}^*\|\le\sqrt{\delta}} \left|(P_n-P)f_{\tilde{\mathbf{c}}_0}\right|\right) &\le 2\mathbb{E}_X\mathbb{E}_\sigma\left(\sup_{\mathbf{c}^*\in\mathscr{M},\|\mathbf{c}-\mathbf{c}^*\|\le\sqrt{\delta}} \frac{1}{n}\sum_{i=1}^n \sigma_i f_{\tilde{\mathbf{c}}_0}(X_i)\right) \\
&\le 2\sqrt{2}\mathbb{E}_X\left(\sqrt{\sup_{\mathbf{c}^*,\mathbf{c}} \|f_{\tilde{\mathbf{c}}_0}\|^2_{l^2(S_n)}\log(m(\varepsilon_0))}\right) \\
&\le 2\sqrt{2}\mathbb{E}_X\left(\sqrt{\|F\|^2_{l^2(S_n)}\log(m(\varepsilon_0))}\right) \\
&\le 2\sqrt{2}\mathbb{E}_X\left(\sqrt{C_F^2\log(m(\varepsilon_0))}\right) \\
&\le \frac{\kappa_A}{\sqrt{n}},
\end{aligned}$$

where κ_A depends on k, d and P. In the second line of this inequality, we used the maximal inequality for random processes depending only on Rademacher variables given in Lemma 2.3 of [Mas07].

It remains to bound the second term. Using the same approach (symmetrization and maximal inequality for Rademacher variables) we get, for every $j > 0$,

$$\begin{aligned}
&\mathbb{E}\left(\sup_{\mathbf{c},\mathbf{c}^*}\left|(P_n-P)(f_{\tilde{\mathbf{c}}_j}-f_{\tilde{\mathbf{c}}_{j-1}})\right|\right) \\
&\quad\le 2\mathbb{E}_X\left(\sqrt{\frac{2}{n}\log(m(\varepsilon_j)m(\varepsilon_{j-1}))\sup_{\mathbf{c},\mathbf{c}^*}\|f_{\tilde{\mathbf{c}}_j}-f_{\tilde{\mathbf{c}}_{j-1}}\|^2_{l^2(S_n)}}\right).
\end{aligned}$$

However $\|f_{\tilde{\mathbf{c}}_j}-f_{\tilde{\mathbf{c}}}\|_{l^2(S_n)} \le \varepsilon_j\|F\|_{l^2(S_n)}$, consequently

$$\begin{aligned}
\|f_{\tilde{\mathbf{c}}_j}-f_{\tilde{\mathbf{c}}_{j-1}}\|^2_{l^2(S_n)} &\le 4\varepsilon^2_{j-1}\|F\|^2_{l^2(S_n)} \\
&\le 4C_F^2\varepsilon^2_{j-1}.
\end{aligned}$$

Comparing a sum with an integral, we obtain

$$\sum_{j>0}\mathbb{E}\left(\sup_{\mathbf{c}^*\in\mathscr{M},\|\mathbf{c}-\mathbf{c}^*\|\le\sqrt{\delta}}\left|(P_n-P)(f_{\tilde{\mathbf{c}}_j}-f_{\tilde{\mathbf{c}}_{j-1}})\right|\right) \le \frac{32}{\sqrt{n}}\int_0^{\varepsilon_1}\sqrt{\log(m(\varepsilon))}d\varepsilon,$$

which, by assumption on $m(\varepsilon)$, can be bounded from above by $\frac{\kappa_B}{\sqrt{n}}$, where κ_B depends on k, d and P.

We are now in position to prove Proposition 2.4. From the two above Subsections we deduce that

$$\mathbb{E}\left(\sup_{\mathbf{c}^*\in\mathscr{M},\|\mathbf{c}-\mathbf{c}^*\|^2\le A_2\delta}\left|(P_n-P)(\gamma(\mathbf{c}^*,.)-\gamma(\mathbf{c},.))\right|\right) \le \sqrt{\delta}\frac{\Xi}{\sqrt{n}}.$$

This concludes the proof.

2.5.6 Proof of Theorem 2.2

Let $\mathbf{x} = (x_1,\dots,x_k)$ be a $k\times d$ vector, $V_1^*,\dots,V_k^*$ the Voronoi cells associated with an optimal codebook $\mathbf{c}^*$. We state here a sufficient condition for the Hessian matrix $H(\mathbf{c}^*)$ to be positive. Denote $r_{ij} = \|c_i^* - c_j^*\|$. It holds

$$\langle H\mathbf{x},\mathbf{x}\rangle = \sum_{i=1}^k\left[\langle H_{i,i}x_i,x_i\rangle + \sum_{j\ne i}\langle H_{i,j}x_j,x_i\rangle\right].$$

Recalling the expression of $H_{i,i}$ and $H_{i,j}$ given in equation 2.1,

$$H_{i,j} = \begin{cases} 2P(V_i) - 2\sum_{\ell \neq i} r_{i\ell}^{-1} \sigma\left[f(x)(x-c_i)(x-c_i)^t \mathbf{1}_{\partial(V_i \cap V_\ell)}\right] & \text{for} \quad i = j \\ 2r_{ij}{}^{-1} \sigma\left[f(x)(x-c_i)(x-c_j)^t \mathbf{1}_{\partial(V_i \cap V_j)}\right] & \text{for} \quad i \neq j \end{cases},$$

we may write, for $i = 1, \ldots, k$,

$$\begin{aligned} \langle H_{i,i} x_i, x_i \rangle + \sum_{j \neq i} \langle H_{i,j} x_j, x_i \rangle = \quad & 2P(V_i^*)\|x_i\|^2 \\ & - 2x_i^t \left(\sum_{j \neq i} r_{i,j}^{-1} \int_{\partial(V_i^* \cap V_j^*)} f(u)(u-c_i^*)(u-c_i^*)^t du \right) x_i \\ & + 2x_i^t \sum_{j \neq i} r_{i,j}^{-1} \left(\int_{\partial(V_i^* \cap V_j^*)} f(u)(u-c_i^*)(u-c_j^*)^t du \right) x_j. \end{aligned}$$

The support of P is included in $\mathscr{B}(0,1)$, thus we can replace $\partial(V_i^* \cap V_j^*)$ with $\partial(V_i^* \cap V_j^*) \cap \mathscr{B}(0,1)$ in the equations above. However, to lighten notation, we will omit the indication and implicitly assume that every set we consider is contained in $\mathscr{B}(0,1)$. Let $p_{i,j} = \int_{\partial(V_i^* \cap V_j^*)} f(u) du$ be the $d-1$-dimensional P-measure of the boundary between V_i^* and V_j^*. Recalling that the underlying norm is the Euclidean norm, even for matrices, we may write

$$\begin{aligned} \langle H_{i,i} x_i, x_i \rangle + \sum_{i \neq j} \langle H_{i,j} x_j, x_i \rangle \geq \quad & 2P(V_i)\|x_i\|^2 \\ & - 2\|x_i\|^2 \left\| \sum_{j \neq i} r_{i,j}^{-1} \int_{\partial(V_i^* \cap V_j^*)} f(u)(u-c_i^*)(u-c_i^*)^t du \right\| \\ & - 2\|x_i\| \left\| \sum_{j \neq i} r_{i,j}^{-1} \left(\int_{\partial(V_i^* \cap V_j^*)} f(u)(u-c_i^*)(u-c_j^*)^t du \right) x_j \right\|, \end{aligned}$$

with

$$\begin{aligned} & \left\| \sum_{j \neq i} r_{i,j}^{-1} \left(\int_{\partial(V_i^* \cap V_j^*)} f(u)(u-c_i^*)(u-c_j^*)^t du \right) x_j \right\| \\ & \quad \leq \sum_{j \neq i} r_{i,j}^{-1} \left\| \left(\int_{\partial(V_i^* \cap V_j^*)} f(u)(u-c_i^*)(u-c_j^*)^t du \right) x_j \right\| \\ & \quad \leq \sum_{j \neq i} r_{i,j}^{-1} \left(\int_{\partial(V_i^* \cap V_j^*)} f(u)\|u-c_i^*\|\|u-c_j^*\| du \right) \|x_j\| \\ & \quad \leq \sum_{j \neq i} r_{i,j}^{-1} p_{i,j} 4\|x_j\|. \end{aligned}$$

Next,

$$\langle H_{i,i} x_i, x_i \rangle + \sum_{j \neq i} \langle H_{i,j} x_j, x_i \rangle \geq \left(2P(V_i^*) - \frac{8}{B} \sum_{i \neq j} p_{i,j} \right) \|x_i\|^2 - \frac{8}{B} \sum_{j \neq i} p_{i,j} \|x_i\| \|x_j\|,$$

where we recall that $B = \inf_{i \neq j, \mathbf{c}^* \in \mathscr{M}} \|c_i^* - c_j^*\|$. Making use of the inequality $2\|x_i\|\|x_j\| \leq \|x_i\|^2 + \|x_j\|^2$, and summing with respect to i leads to

$$\langle H\mathbf{x}, \mathbf{x} \rangle \geq \sum_{i=1}^{k} \left(2P(V_i) - \frac{16}{B} \sum_{j \neq i} p_{i,j} \right) \|x_i\|^2.$$

The last step is to derive bounds for $p_{i,j}$ from the conditions on f. Denote $\lambda = \|f\|_\infty$, we see that

$$\sum_{j\neq i} p_{i,j} = \int_{\partial V_i^*} f(u)du.$$

V_i^* is a regular convex set included in $\mathscr{B}(c_i^*, 2)$. Therefore, by a direct application of Stokes Theorem, the surface of ∂V_i^* is smaller than the surface of $\mathscr{S}_{d-1}(c_i^*, 2)$ (the sphere of radius 2). Consequently

$$\sum_{j\neq i} p_{i,j} \leq \lambda \frac{2\pi^{d/2}}{\Gamma(d/2)} 2^{d-1}.$$

It follows that $\lambda < \frac{B\Gamma(d/2)}{2^{d+3}\pi^{d/2}} \inf_{i=1,\dots,k} P(V_i^*)$ is enough to ensure that the Hessian matrix $H(\mathbf{c}^*)$ is positive definite.

2.5.7 Proof of Proposition 2.2

We consider a distribution on $\mathbb{R}^d$, distributed over small balls away from one another, and whose density inside each ball is a small cone, for continuity reasons. Denote by V_i the Voronoi cell associated with z_i in $(z_1,\dots,z_k)$. Let Q be a k-quantizer, Q^* the expected optimal quantizer which maps V_i to z_i for all i. Denote finally, for all $i = 1,\dots,k$, $R_i(Q) = \int_{V_i} \|x - Q(x)\|^2 dx$ the contribution of the i-th Voronoi cell to the risk of Q.

Let S denote the surface of the unit ball in $\mathbb{R}^d$. Taking into account that $N_\rho = \frac{kS\rho^d}{d(d+1)}$ we have

$$\begin{aligned} R_i(Q^*) &= \frac{1}{kN_\rho} \int_0^\rho Sr^{d+1}\left(1 - \frac{r}{\rho}\right) dr \\ &= \frac{\rho^2 d(d+1)}{k(d+3)(d+2)}. \end{aligned}$$

Let i be an integer between 1 and k. Let $m_i^{in} = \left|Q(\mathscr{B}_d(z_i,\rho)) \cap V_i\right|$ be the number of images of V_i sent by Q inside V_i, and let $m_i^{out} = \left|Q(\mathscr{B}_d(z_i,\rho)) \cap V_i^c\right|$ be the number of images of V_i sent outside V_i. The three situations of interest are the following ones :

→ If $m_i^{in} = 1$ and $m_i^{out} = 0$, it is clear that $R_i(Q) \geq R_i(Q^*)$.

→ If $m_i^{in} \geq 2$ and $m_i^{out} = 0$, then $R_i(Q) \geq R_i(Q^*) - \frac{\rho^2 d(d+1)}{k(d+2)(d+3)} = 0$.

→ At last, suppose that $m_i^{out} \geq 1$. Then there exists $x \in \mathscr{B}_d(z_i,\rho)$ such that $Q(x) \notin V_i$. Since Q is a nearest neighbor quantizer, for such an x we have

$$\begin{cases} \|Q(x) - x\| \leq \inf\limits_{c \in Q(\mathscr{B}_d(z_i,\rho))} \|x - c\|, \\ \|Q(x) - x\| \geq d(z_i, V_i^c) - \rho \geq \dfrac{R}{2} - \rho. \end{cases}$$

Let $c \in Q(\mathscr{B}_d(z_i,\rho))$. Then

$$\begin{aligned} \|c - z_i\| &\geq \|c - x\| - \rho \\ &\geq \|Q(x) - x\| - \rho \\ &\geq \frac{R}{2} - 2\rho. \end{aligned}$$

Then, we deduce that, for every $y \in \mathscr{B}_d(z_i,\rho)$ and codepoint $c \in Q(\mathscr{B}_d(z_i,\rho))$, $\|y-c\| \geq \frac{R}{2} - 3\rho$. Therefore

$$\begin{aligned} R_i(Q) &\geq \frac{\left(\frac{R}{2}-3\rho\right)^2}{k} \\ &\geq R_i(Q^*) + \frac{1}{k}\left(\left(\frac{R}{2}-3\rho\right)^2 - \frac{\rho^2 d(d+1)}{(d+2)(d+3)}\right). \end{aligned}$$

Now suppose that $m_i^{in} \geq 2$. Then at least two code points of Q lies in V_i. Therefore, there exists j such that no code point of Q lies in V_j, so that $m_j^{out} \geq 1$. We straightforward deduce that the number of cells V_i for which $m_i^{in} \geq 2$ is smaller than the number of cells for which $m_j^{out} \geq 1$.

Taking into account all contributions of Voronoi cells, we get

$$\begin{aligned} R(Q) &= \sum_{\{i;m_i^{in}\geq 2, m_i^{out}=0\}} R_i(Q) + \sum_{\{i;m_i^{out}\geq 1\}} R_i(Q) + \sum_{\{i;m_i^{in}=1, m_i^{out}=0\}} R_i(Q) \\ &\geq R(Q^*) + \sum_{\{i;m_i^{in}\geq 2, m_i^{out}=0\}} \frac{1}{k}\left(\left(\frac{R}{2}-3\rho\right)^2 - \frac{2\rho^2 d(d+1)}{(d+2)(d+3)}\right), \end{aligned}$$

from which we deduce a sufficient condition to get $R(Q) \geq R(Q^*)$.

2.5.8 Proof of Proposition 2.3

We begin with a lemma which ensures that every possible optimal code point c_i^* is close to at least one mean m_j of the mixture, in the case where the ratio p_{min}/p_{max} is large enough.

Lemma 2.6. *Let $\mathbf{c}^*$ be an optimal codebook. Suppose that*

$$\frac{p_{min}}{p_{max}} \geq \frac{288 k\sigma^2}{(1-\varepsilon)\tilde{B}^2(1-e^{-\tilde{B}^2/288\sigma^2})}.$$

Then, for every $j = 1,\dots,k$, there exists $i \in \{1,\dots k\}$ such that $\|m_j - c_i^*\| \leq \frac{\tilde{B}}{6}$.

Proof of Lemma 2.6. Denote by $\mathbf{m}$ the codebook $(m_1,\dots,m_k)$, and by M_i the Voronoi cell associated with m_i. We bound from above the quantity $P\gamma(\mathbf{m},.)$:

$$\begin{aligned} P\gamma(\mathbf{m},.) &= \sum_{i=1}^k \frac{p_i}{2\pi\sigma^2 N_i} \int_{M_i} \|x-m_i\|^2 e^{-\|x-m_i\|^2/2\sigma^2} dx \\ &\leq \sum_{i=1}^k \frac{p_i}{2\pi\sigma^2 N_i} \int_{\mathbb{R}^2} \|x-m_i\|^2 e^{-\|x-m_i\|^2/2\sigma^2} dx \\ &\leq \frac{2k p_{max}\sigma^2}{(1-\varepsilon)}. \end{aligned}$$

Let $\mathbf{c}$ be a codebook such that there exists j satisfying, for all $i = 1,\dots,k$, $\|m_j - c_i\| > \tilde{B}/6$. We will prove that $P\gamma(\mathbf{c},.) > P\gamma(\mathbf{m},.)$, which implies that $\mathbf{c} \notin \mathscr{M}$. In fact we have, for all $i = 1,\dots,k$ and for all $x \in \mathscr{B}(m_j,\tilde{B}/12)$, $\|x-c_i\| > \tilde{B}/12$. Hence, a lower bound

for $P\gamma(\mathbf{c},.)$ is

$$\begin{aligned}
P\gamma(\mathbf{c},.) &\geq \int_{\mathscr{B}(m_j,\tilde{B}/12)} \min_{i=1,\dots,k} \|x-c_i\|^2 f(x)dx \\
&> \frac{\tilde{B}^2}{144} \sum_{i=1}^{k} \frac{p_i}{2\pi\sigma^2 N_i} \int_{\mathscr{B}(m_j,\tilde{B}/12)} e^{-\|x-m_i\|^2/2\sigma^2} dx \\
&> \frac{\tilde{B}^2 p_j}{288\pi\sigma^2 N_j} \int_{\mathscr{B}(m_j,\tilde{B}/12)} e^{-\|x-m_j\|^2/2\sigma^2} dx \\
&> \frac{p_{min}\tilde{B}^2}{144}\left(1-e^{-\tilde{B}^2/288\sigma^2}\right) \\
&> P\gamma(\mathbf{m},.).
\end{aligned}$$

Hence we deduce that every optimal codebook has a code point close to every mean m_j of the mixture, of at most $\tilde{B}/6$. □

Suppose that the ratio p_{min}/p_{max} satisfies the assumption of Proposition 2.3. In particular p_{min}/p_{max} satisfies the assumption of Lemma 2.6. Then we deduce that, up to a re indexation, for every $\mathbf{c}^* \in \mathscr{M}$, $\|c_i^* - m_i\| \leq \tilde{B}/6$. We conclude that $2\tilde{B}/3 \leq B \leq 4\tilde{B}/3$.

Since, for all $i = 1,\dots,k$, $\mathscr{B}(c_i^*,B/2) \subset V_i^*$, it is easy to see that $\mathscr{B}(m_i,B/4) \subset \mathscr{B}(c_i^*,B/2) \subset V_i^*$, which leads to $N^* \subset \left(\bigcup_{i=1}^{k} \mathscr{B}(m_i,B/4)\right)^c$. Consequently, in order to apply Theorem 2.2, we just have to prove that

$$\|f_{\left|\left(\bigcup_{i=1}^{k} \mathscr{B}(m_i,B/4)\right)^c}\|_\infty \leq \frac{\Gamma(1)B}{2^5\pi} \inf_{i=1,\dots,k} P(\mathscr{B}(m_i,B/4)).$$

First we derive a lower bound for the right-hand side. For every $i = 1,\dots,k$,

$$\begin{aligned}
P(\mathscr{B}(m_i,B/4)) &\geq \frac{p_i}{N_i}\frac{1}{2\pi\sigma^2} \int_{\mathscr{B}(0,B/4)} e^{-\frac{\|x\|^2}{2\sigma^2}} dx \\
&\geq \frac{p_i}{N_i}\frac{1}{2\pi\sigma^2} \times 2\pi \int_0^{B/4} r e^{-\frac{r^2}{2\sigma^2}} dr \\
&\geq p_{min}\left(1-e^{-\frac{B^2}{32\sigma^2}}\right).
\end{aligned}$$

Then, we deal with the left-hand side. Let x be at distance from every m_i of at least $B/4$. Then

$$\begin{aligned}
f(x) &\leq \sum_{i=1}^{k} \frac{p_i}{N_i}\frac{1}{2\pi\sigma^2} e^{-\frac{B^2}{32\sigma^2}} \\
&\leq \frac{k p_{max}}{2\pi\sigma^2(1-\varepsilon)} e^{-\frac{B^2}{32\sigma^2}}.
\end{aligned}$$

The rest of the proof follows from straightforward computation, using the assumption of Proposition 2.3 and the relationship between B and $\tilde{B}$: $2\tilde{B}/3 \leq B \leq 4\tilde{B}/3$.

Remark. A careful reader should have noticed that the k factor is suboptimal in the previous inequality. In fact we are able in this case to bound from above $f(x)$ with $\frac{1}{2\pi\sigma^2(1-\varepsilon)} e^{-\frac{B^2}{32\sigma^2}}$. However, this bound does not involve p_{max}, and so involve a condition not on the ratio of extremal proportions of the mixture, but rather on the

minimal proportion of the mixture, which is less natural. Moreover, the p_{max}-free bound is valid only in the equal variance case, namely when the variance σ_i^2 of any element of the mixture is the same. In general it is not the case and a condition as in Proposition 2.3 for that kind of mixture would naturally involve the ratio p_{min}/p_{max}.

Acknowledgement

The author would like to thank two referees for valuable comments and suggestions.

Chapitre 3

Non asymptotic bounds for vector quantization

Le second chapitre de ce manuscrit présente des résultats dans la continuité de ceux exposés dans le Chapitre 3 : les bornes données sur la vitesse de convergence sont maintenant explicites en les différents autres paramètres du problème, ce qui permet de discuter de leur influence, et une condition de type marge plus générale que celle énoncée en (2.2) est présentée. Par ailleurs, la majorité des résultats que l'on trouvera dans le Chapitre 3 sont donnés pour le cas où $\mathscr{H}$ est un espace de Hilbert séparable, donc de dimension potentiellement infinie, ce qui les rend applicables pour la classification de courbes comme dans [AF12]. Ce chapitre a fait l'objet d'une publication, [Lev14], en cours de soumission.

Sommaire

Recent results in quantization theory show that the convergence rate for the mean-squared expected distortion of the empirical risk minimizer strategy, for any fixed probability distribution satisfying some regularity conditions, is $\mathcal{O}(1/n)$, where n is the sample size (see, e.g., [CL06] or Theorem 2.1 in Chapter 2). However, the

dependency of the average distortion on other parameters is not known, and these results are valid for distributions over finite dimensional Euclidean spaces.

This paper deals with the general case of distributions over separable, possibly infinite dimensional, Hilbert spaces. A condition is proposed, which may be thought of as a margin condition (see, e.g., [MT99]), under which a non asymptotic upper bound on the expected distortion rate of the empirically optimal quantizer is derived. The dependency of the distortion on other natural parameters of the quantization issue is then discussed, in particular through a minimax lower bound.

3.1 Introduction

Quantization, also called lossy data compression in information theory, is the problem of replacing a probability distribution with an efficient and compact representation, that is a finite set of points. To be more precise, let $\mathscr{H}$ denote a separable Hilbert space, and let P denote a probability distribution over $\mathscr{H}$ and k a positive integer. A so-called k-points quantizer Q is a map from $\mathscr{H}$ to $\mathscr{H}$, whose image set is made of exactly k points, that is $|Q(\mathscr{H})| = k$. For such a quantizer, every image point $c_i \in Q(\mathscr{H})$ is called a code point, and the vector composed of the code points $(c_1, \dots, c_k)$ is called a codebook, denoted by $\mathbf{c}$. By considering the preimages of its code points, a quantizer Q partitions the separable Hilbert space $\mathscr{H}$ into k groups, and assigns each group a representative. General references on the subject are to be found in [GL00], [GG91] and [Lin02] among others.

The quantization theory was originally developed as a way to answer signal compression issues in the late 40's (see, e.g., [GG91]). However, unsupervised classification is also in the scope of its application. Isolating meaningful groups from a cloud of data is a topic of interest in many fields, from social science to biology. Classifying points into dissimilar groups of similar items is more interesting as the amount of accessible data is large. In many cases data need to be preprocessed through a quantization algorithm in order to be exploited.

If the distribution P has a finite second moment, the performance of a quantizer Q is measured by the risk, or distortion

$$R(Q) := P\|x - Q(x)\|^2,$$

where Pf means integration of the function f with respect to P. The choice of the squared norm is convenient, since it takes advantages of the Hilbert space structure of $\mathscr{H}$. Nevertheless, it is worth pointing out that several authors deal with more general distortion functions. For further information on this topic, the interested reader is referred to [GL00] or [Fis10].

In order to minimize the distortion introduced above, it is clear that only quantizers of the type $x \mapsto \arg\min_{c_1,\dots,c_k} \|x - c_i\|^2$ are to be considered. Such quantizers are called nearest-neighbor quantizers. With a slight abuse of notation, $R(\mathbf{c})$ will denote the risk of the nearest-neighbor quantizer associated with a codebook $\mathbf{c}$.

Provided that P has a bounded support, there exist optimal codebooks minimizing the risk R (see, e.g., Corollary 3.1 in [Fis10] or Theorem 1 in [GLP07]). The aim is to design a codebook $\hat{\mathbf{c}}_n$, according to an n-sample drawn from P, whose distortion is as close as possible to the optimal distortion $R(\mathbf{c}^*)$, where $\mathbf{c}^*$ denotes an optimal codebook.

To solve this problem, most approaches to date attempt to implement the principle of empirical risk minimization in the vector quantization context (see, e.g., [LLZ94]). Let $X_1,\dots,X_n$ denote an independent and identically distributed sample with distribution P. According to this principle, good code points can be found by searching for ones that minimize the empirical distortion over the training data, defined by

$$\hat{R}_n(\mathbf{c}) := \frac{1}{n}\sum_{i=1}^{n} \|X_i - Q(X_i)\|^2 = \frac{1}{n}\sum_{i=1}^{n} \min_{j=1,\dots,k} \|X_i - c_j\|^2.$$

If the training data represents the source well, then $\hat{\mathbf{c}}_n$ will hopefully also perform near optimally on the real source, that is $\ell(\hat{\mathbf{c}}_n,\mathbf{c}^*) = R(\hat{\mathbf{c}}_n) - R(\mathbf{c}^*) \approx 0$. The problem of quantifying how good empirically designed codebooks are, compared to the truly optimal ones, has been extensively studied, as for instance in [Lin02] in the finite dimensional case.

In the case where $\mathscr{H} = \mathbb{R}^d$, for some $d > 0$, it has been proved in [LLZ94] that $\mathbb{E}\ell(\hat{\mathbf{c}}_n,\mathbf{c}^*) = \mathscr{O}(1/\sqrt{n})$, provided that P has a bounded support. This result has been extended to the case where $\mathscr{H}$ is a separable Hilbert space in [BDL08]. However, this upper bound has been tightened whenever the source distribution satisfies additional assumptions, in the finite dimensional case only.

When $\mathscr{H} = \mathbb{R}^d$, for the special case of finitely supported distributions, it is shown in [AGG05] that $\mathbb{E}\ell(\hat{\mathbf{c}}_n,\mathbf{c}^*) = \mathscr{O}(1/n)$. There are much more results in the case where P is not assumed to have a finite support.

In fact, different sets of assumptions have been introduced in [AGG05], [Pol82b] or in (2.2) exposed in Chapter 2, to derive fast convergence rates for the distortion in the finite dimensional case. To be more precise, it is proved in [AGG05] that, if P satisfies a technical inequality for every codebook $\mathbf{c}$, namely

$$\ell(\mathbf{c},\mathbf{c}^*) \geq a \operatorname{Var}\left(\min_{j=1,\dots,k} \|X - c_j\|^2 - \min_{j=1,\dots,k} \|X - c_j^*\|^2 \right), \tag{3.1}$$

for some $a > 0$, then $\mathbb{E}\ell(\hat{\mathbf{c}}_n,\mathbf{c}^*) \leq C(k,d,P)\log(n)/n$, where $C(k,d,P)$ depends on the natural parameters k and d, but also on the technical parameter a. However, in the continuous density and unique minimum case, it has been proved in [Cho94], following the approach of [Pol82b], that, provided that the Hessian matrix of $\mathbf{c} \mapsto R(\mathbf{c})$ is positive definite at the optimal codebook, $n\ell(\hat{\mathbf{c}}_n,\mathbf{c}^*)$ converges in distribution to a law, depending on the Hessian matrix. As proved in Chapter 2, the technique used in [Pol82b] can be slightly modified to derive a non-asymptotic bound of the type $\mathbb{E}\ell(\hat{\mathbf{c}}_n,\mathbf{c}^*) \leq C/n$ in this case, for some unknown $C > 0$.

As shown in Chapter 2, these different sets of assumptions turn out to be equivalent in the continuous density case to a technical condition, similar to that used in [MN06] to derive fast rates of convergence in the statistical learning framework.

Thus, a question of interest is to know whether some margin type conditions can be derived for the source distribution to satisfy the technical condition mentioned above, as has been done in the statistical learning framework in [MT99]. This paper provides a condition, which can clearly be thought of as a margin condition in the quantization framework, under which the condition (3.1) is satisfied, where the technical constant a has an explicit expression in term of natural parameters of the quantization issue, such as the smallest distance between two optimal code points. It is worth mentioning that this margin condition does not require $\mathscr{H}$ to have a finite dimension, or P to have a continuous density. In the finite dimensional case,

this condition does not require either that there exists a unique optimal codebook, as required in [Pol82b], hence seems easier to check.

Moreover, a non asymptotic bound of the type $\mathbb{E}\ell(\hat{\mathbf{c}}_n, \mathbf{c}^*) \leq C(k,P)/n$ is derived for distributions satisfying this margin condition, where $C(k,P)$ is explicitly given in terms of natural parameters of the quantization issue. This bound is also valid in the case where $\mathscr{H}$ has an infinite dimension. This point may be of interest for curve quantization, as done in [AF12].

In addition, a minimax lower bound is given which allows to discuss the influence of the different parameters mentioned in the upper bound. It is worth pointing out that this lower bound is valid over a set of probability distributions with uniformly bounded continuous densities and unique optimal codebook, such that the minimum eigenvalue of the second derivative matrices, at the optimal codebook, is uniformly lower bounded. This result refines the previous minimax bounds obtained in [Ant05] or [BLL98].

The paper is organized as follows. In Section 3.2 some notation and definition are introduced, along with some basic results for quantization in a Hilbert space. The so-called margin condition is then introduced, and the main results are exposed in Section 3.3 : firstly an oracle inequality on the loss is stated, along with a minimax result. Then it is shown that Gaussian mixtures are in the scope of the margin conditions. Finally, proofs are gathered in Section 3.4, and technical proofs in Section 3.5.

3.2 Notation and Definitions

Throughout the paper, for $M > 0$ and a in $\mathscr{H}$, $\mathscr{B}(a,M)$ will denote the closed ball with center a and radius M. With a slight abuse of notation, P is said to be M-bounded if its support is included in $\mathscr{B}(0,M)$. Furthermore, it will also be assumed that the support of P contains more than k points.

To frame the quantization issue as an empirical risk minimization issue, the following contrast function γ is introduced as

$$\gamma : \begin{cases} (\mathscr{H})^k \times \mathscr{H} & \longrightarrow & \mathbb{R} \\ (\mathbf{c}, x) & \longmapsto & \min_{j=1,\dots,k} \|x - c_j\|^2 \end{cases},$$

where $\mathbf{c} = (c_1, \dots, c_k)$ denotes a codebook, that is a kd-dimensional vector if $\mathscr{H} = \mathbb{R}^d$. Throughout the paper, only the case $k \geq 2$ will be considered. The risk $R(\mathbf{c})$ then takes the form $R(\mathbf{c}) = R(Q) = P\gamma(\mathbf{c}, .)$, where we recall that Pf denotes the integration of the function f with respect to P. Similarly, the empirical risk $\hat{R}_n(\mathbf{c})$ can be defined as $\hat{R}_n(\mathbf{c}) = P_n\gamma(\mathbf{c}, .)$, where P_n is the empirical distribution associated with $X_1, \dots, X_n$, in other words $P_n(A) = 1/n\,|\{i | X_i \in A\}|$, for every measurable subset $A \subset \mathscr{H}$.

It is worth pointing out that, if P is M-bounded, for some $M > 0$, then there exist such minimizers $\hat{\mathbf{c}}_n$ and $\mathbf{c}^*$ (see, e.g., Corollary 3.1 in [Fis10]). In the sequel the set of minimizers of the risk $R(.)$ will be denoted by $\mathscr{M}$. Since every permutation of labels of an optimal codebook provides an optimal codebook, $\mathscr{M}$ contains more than $k!$ elements. To address the issue of a large number of optimal codebooks, $\bar{\mathscr{M}}$

is introduced as a set of codebooks which satisfies

$$\begin{cases} \forall \mathbf{c}^* \in \mathscr{M} \quad \exists \bar{\mathbf{c}} \in \bar{\mathscr{M}} & \{c_1^*, \dots, c_k^*\} = \{\bar{c}_1, \dots, \bar{c}_k\}, \\ \quad \forall \bar{\mathbf{c}}^1, \bar{\mathbf{c}}^2 \in \bar{\mathscr{M}} & \{\bar{c}_1^1, \dots, \bar{c}_k^1\} \neq \{\bar{c}_1^2, \dots, \bar{c}_k^2\}. \end{cases}$$

In other words, $\bar{\mathscr{M}}$ is a subset of the set of optimal codebooks which contains every element of $\mathscr{M}$, up to a permutation of the labels, and in which two different codebooks have different sets of code points. It may be noticed that $\bar{\mathscr{M}}$ is not uniquely defined. However, when $\mathscr{M}$ is finite, all the possible $\bar{\mathscr{M}}$ have the same cardinality.

Let $c_1, \dots, c_k$ be a sequence of code points. A central role is played by the set of points which are closer to c_i than to any other c_j's. To be more precise, the Voronoi cell, or quantization cell associated with c_i is the closed set defined by

$$V_i(\mathbf{c}) = \left\{ x \in \mathscr{H} | \quad \forall j \neq i \quad \|x - c_i\| \leq \|x - c_j\| \right\}.$$

It may be noted that $(V_1(\mathbf{c}), \dots, V_k(\mathbf{c}))$ does not form a partition of $\mathscr{H}$, since $V_i(\mathbf{c}) \cap V_j(\mathbf{c})$ may be non empty. To address this issue, a Voronoi partition associated with $\mathbf{c}$ is defined as a sequence of subsets $(W_1(\mathbf{c}), \dots, W_k(\mathbf{c}))$ which forms a partition of $\mathscr{H}$, and such that for every $i = 1, \dots, k$,

$$\bar{W}_i(\mathbf{c}) = V_i(\mathbf{c}),$$

where $\bar{W}_i(\mathbf{c})$ denotes the closure of the subset $W_i(\mathbf{c})$. The open Voronoi cell is defined the same way by

$$\overset{o}{V}_i(\mathbf{c}) = \left\{ x \in \mathscr{H} | \quad \forall j \neq i \quad \|x - c_i\| < \|x - c_j\| \right\}.$$

Given a Voronoi partition $W(\mathbf{c}) = (W_1(\mathbf{c}), \dots, W_k(\mathbf{c}))$, the following inclusion holds, for i in $\{1, \dots, k\}$,

$$\overset{o}{V}_i(\mathbf{c}) \subset W_i(\mathbf{c}) \subset V_i(\mathbf{c}),$$

and the risk $R(\mathbf{c})$ takes the form

$$R(\mathbf{c}) = \sum_{i=1}^{k} P\left(\|x - c_i\|^2 \mathbb{1}_{W_i(\mathbf{c})}(x) \right),$$

where $\mathbb{1}_A$ denotes the indicator function associated with A. Whenever $(W_1, \dots, W_k)$ are fixed subsets such that $P(W_i) \neq 0$, for every $i = 1, \dots, k$, it is clear that

$$P(\|x - c_i\|^2 \mathbb{1}_{W_i(\mathbf{c})}(x)) \geq P(\|x - \eta_i\|^2 \mathbb{1}_{W_i(\mathbf{c})}(x)),$$

with equality only if $c_i = \eta_i$, where η_i denotes the conditional expectation of P over the subset $W_i(\mathbf{c})$, that is

$$\eta_i = \frac{P(x \mathbb{1}_{W_i(\mathbf{c})}(x))}{P(W_i(\mathbf{c}))}.$$

Moreover, it is proved in Proposition 1 of [GLP07] that, for every Voronoi partition $W(\mathbf{c}^*)$ associated with an optimal codebook $\mathbf{c}^*$, and every $i = 1, \dots, k$, $P(W_i(\mathbf{c}^*)) \neq 0$. Consequently, any optimal codebook satisfies the so-called centroid condition (see, e.g., Section 6.2 of [GG91]), that is

$$\mathbf{c}_i^* = \frac{P(x \mathbb{1}_{W_i(\mathbf{c}^*)}(x))}{P(W_i(\mathbf{c}^*))}.$$

As a remark, the centroid condition ensures that, for every $\mathbf{c}^*$ in $\mathscr{M}$ and $i \neq j$,

$$\begin{aligned} P(V_i(\mathbf{c}^*) \cap V_j(\mathbf{c}^*)) &= P\left(\left\{x \in \mathscr{H} | \quad \forall i' \quad \|x - c_i^*\| = \|x - c_j^*\| \leq \|x - c_{i'}^*\|\right\}\right) \\ &= 0. \end{aligned}$$

A proof of this statement can be found in Proposition 1 of [GLP07]. According to this remark, it is clear that, for every optimal Voronoi partition $(W_1(\mathbf{c}^*), \ldots, W_k(\mathbf{c}^*))$,

$$\begin{cases} P(W_i(\mathbf{c}^*)) & = & P(V_i(\mathbf{c}^*)), \\ P_n(W_i(\mathbf{c}^*)) & \underset{a.s.}{=} & P_n(V_i(\mathbf{c}^*)). \end{cases} \tag{3.2}$$

The following quantities are of importance in the bounds exposed in Section 3.1 :

$$\begin{cases} B & = & \inf_{\mathbf{c}^* \in \mathscr{M}, i \neq j} \|c_i^* - c_j^*\|, \\ p_{min} & = & \inf_{\mathbf{c}^* \in \mathscr{M}, i = 1, \ldots, k} P(V_i(\mathbf{c}^*)). \end{cases}$$

It is worth noting here that $B \leq 2M$ whenever P is M-bounded, and $p_{min} \leq 1/k$. If $\mathscr{M}$ is finite, it is clear that p_{min} and B are strictly positive. The following Proposition ensures that this statement remains true when $\mathscr{M}$ is not assumed to be finite.

Proposition 3.1. *Suppose that P is M-bounded. Then both B and p_{min} are strictly positive quantities.*

A proof of Proposition 3.1 is given in Section 3.4. The role of the boundaries between optimal Voronoi cells may be compared to the role played by the critical value 1/2 for the regression function in the statistical learning framework. To draw this comparison, the following set is introduced, for any $\mathbf{c}^* \in \mathscr{M}$,

$$N(\mathbf{c}^*) = \bigcup_{i \neq j} V_i(\mathbf{c}^*) \cap V_j(\mathbf{c}^*).$$

Next, the critical region N^* is defined as

$$N^* = \bigcup_{\mathbf{c}^* \in \mathscr{M}} N(\mathbf{c}^*).$$

This region seems to be of importance when considering the conditions under which the empirical risk minimization strategy for the quantization issue achieves faster rates of convergence, as exposed in Theorem 2.1 of Chapter 2. However, to fully draw the comparison between the margin conditions for the statistical learning issue (see, e.g., [MT99]) and quantization, the neighborhood of this region has to be introduced. For this purpose the t-neighborhood of the critical region N^* is defined as

$$N_t^* = \left\{x \in \mathscr{H} | \quad d(x, N^*) \leq t\right\}.$$

Intuitively, if $P(N_t^*)$ is small enough, then the source distribution P is concentrated around its optimal codebook, and may be thought of as a slight modification of the probability distribution with finite support made of an optimal codebook $\mathbf{c}^*$. To be more precise, let us introduce the following key assumption :

Definition 3.1 (Margin condition)**.** *Denote by $p(t) = P(N_t^*)$. Then P satisfies a margin condition with radius r_0 if and only if*

i) P is M-bounded,
ii) for all $0 \leq t \leq r_0$,

$$p(t) \leq \frac{B p_{min}}{128 M^2} t. \tag{3.3}$$

Note that, since $p(2M) = 1$, $p_{min} \leq 1/k$, $k \geq 2$ and $B \leq 2M$, (3.3) implies that $r_0 < 2M$. It is worth pointing out that Definition 3.1 does not require P to have a density or a unique optimal codebook, up to relabeling, contrary to the conditions introduced in [Pol82b].

It may be mentioned that the margin condition introduced here only requires a local control of the weight of the neighborhood of the critical region N^*. The parameter r_0 may be thought of as a gap size around N^*, as illustrated by the following example :

Example 1 : Assume that there exists $r > 0$ such that $p(x) = 0$ if $x \leq r$ (for instance if P is supported on k points). Then P satisfies (3.3), with radius r.

It is also worth pointing out that the condition mentioned in [MT99] requires a control of the weight of the neighborhood of the critical value 1/2 with a polynomial function with degree larger than 1. In the quantization framework, the special role played by the exponent 1 leads to only consider linear controls of the weight function. This point is explained by the following example :

Example 2 : Assume that P is M-bounded, and that there exists $Q > 0$ and $q > 1$ such that $p(x) \leq Q x^q$. Then P satisfies (3.3), with

$$r_0 = \left(\frac{p_{min} B}{128 M^2 Q} \right)^{1/(q-1)}.$$

In the case where P has a density and $\mathscr{H} = \mathbb{R}^d$, the condition (3.3) can be thought of as a generalization of the condition mentioned in (2.2) of Chapter 2, which requires the density of the distribution to be small enough over the critical region N^*. In fact, provided that P has a continuous density, a uniform bound on the density over N^* provides a local control of $p(t)$ with a polynomial function of degree 1. This idea is developed in the following example :

Example 3(Continuous densities, $\mathscr{H} = \mathbb{R}^d$) : Assume that $\mathscr{H} = \mathbb{R}^d$, P has a continuous density f and is M-bounded, and that $\mathscr{M}$ is finite. In this case, $p(t)$ is differentiable at 0, with derivative

$$p'(0) = \int_{N^*} f(u) d\lambda_{d-1}(u),$$

where λ_{d-1} denotes the $(d-1)$ dimensional Lebesgue measure, considered over the $(d-1)$ dimensional space N^*. Therefore, if P satisfies

$$\int_{N^*} f(u) d\lambda_{d-1}(u) < \frac{B p_{min}}{128 M^2}, \tag{3.4}$$

then there exists $r_0 > 0$ such that P satisfies (3.3). It can easily be deduced from (3.4) that a uniform bound on the density located at the critical region N^* can provide a sufficient condition for a distribution P to satisfy a margin condition. Such a result has to be compared to (2.2) in Chapter 2, where it was required that

$$\| f_{|N^*} \|_\infty \leq \frac{\Gamma\left(\frac{d}{2}\right) B}{2^{d+5} M^{d+1} \pi^{d/2}} p_{min},$$

where Γ denotes the Gamma function, and $f_{|N^*}$ denotes the restriction of f to the set N^*. Note however that the uniform bound mentioned above ensures that the Hessian matrices of the risk function R, at optimal codebooks, are positive definite. This does not necessarily implies that (3.3) is satisfied.

Another interesting parameter of the quantization issue is the following separation factor, which quantifies the difference between optimal codebooks and local minima of the risk.

Definition 3.2. *Denote by $\tilde{\mathscr{M}}$ the set of local minima of the map $\mathbf{c} \longmapsto P\gamma(\mathbf{c},.)$. Let $\varepsilon > 0$, then P is said to be ε-separated if*

$$\inf_{\mathbf{c} \in \tilde{\mathscr{M}} \cap \mathscr{M}^c} \ell(\mathbf{c}, \mathbf{c}^*) = \varepsilon. \tag{3.5}$$

It may be noticed that local minima of the risk function satisfy the centroid condition. Whenever $\mathscr{H} = \mathbb{R}^d$, P has a density and $P\|x\|^2 < \infty$, it can be proved, using Lemma A of [Pol82b], that the set of minima of R coincides with the set of codebooks satisfying the centroid condition, also called stationary points. However, this result cannot be extended to non continuous distributions, as proved in Example 4.11 of [GL00].

The main results of the present paper are based on the following Proposition, which connects the margin condition stated in Definition 3.1 to the condition introduced in Theorem 2 of [AGG05]. It is recalled here that only the case $k \geq 2$ is considered.

Proposition 3.2. *Assume that P satisfies a margin condition with radius r_0, and is ε-separated. Then, for all $\mathbf{c} \in \mathscr{B}(0, M)$,*

i) there exists $\mathbf{c}^*(\mathbf{c}) \in \mathscr{M}$ such that

$$\|\mathbf{c} - \mathbf{c}^*(\mathbf{c})\| = \arg \inf_{\mathbf{c}^* \in \mathscr{M}} \|\mathbf{c} - \mathbf{c}^*\|,$$

ii) $\|\mathbf{c} - \mathbf{c}^*(\mathbf{c})\|^2 \leq \kappa_0 \ell(\mathbf{c}, \mathbf{c}^*)$,

where $\kappa_0 = 4kM^2 \left(\frac{1}{\varepsilon} \vee \frac{64M^2}{p_{min} B^2 r_0^2} \right)$.

Moreover, if $\mathscr{H} = \mathbb{R}^d$, then $\mathscr{M}$ is finite.

As mentioned in [CL06] or in Chapter 2, the connection between the loss and the squared distance can be thought of as a technical margin condition. It is worth pointing out that the dependency of κ_0 on different parameters of the quantization issue is known. This fact allows us to roughly discuss how κ_0 should scale with the parameters k, d and M, in the finite dimensional case. According to Theorem 6.2 of [GL00], $R(\mathbf{c}^*)$ scales like $k^{-2/d}$, at least in the density case. Furthermore, it is likely that $r_0 \sim B$ (see, e.g., the distributions exposed in Section 3.3.2). Considering that $\varepsilon \sim R(\mathbf{c}^*) \sim k^{-2/d}$, $r_0 \sim B \sim Mk^{-1/d}$, and $p_{min} \sim 1/k$ leads to

$$\kappa_0 \sim k^{2+4/d}.$$

At first sight κ_0 does not scale with M, and seems to decrease with the dimension, at least in the finite dimensional case. However, there is no result on how κ_0 should scale in the infinite dimensional case.

It is worth mentioning that, if $\mathscr{H} = \mathbb{R}^d$, P has a unique optimal codebook up to relabeling, and has a continuous density, Proposition 3.2 ensures that the second derivative matrix of R at the optimal codebook is positive definite, with minimum eigenvalue larger than $p_{min}/2$. This is the condition required in [Pol82b] for

$n\ell(\hat{\mathbf{c}}_n, \mathbf{c}^*)$ to converge in distribution. This Proposition allows us to derive explicit upper bounds on the excess risk in the following Section.

3.3 Results

3.3.1 Risk bound

The main result of this Chapter is the following :

Theorem 3.1. *Let k be larger than 2. Assume that $\mathscr{M}$ is finite, P satisfies a margin condition with radius r_0, and is ε-separated. Let κ_0 be defined as*

$$\kappa_0 = 4kM^2\left(\frac{1}{\varepsilon} \vee \frac{64M^2}{p_{min}B^2r_0^2}\right).$$

If $\hat{\mathbf{c}}_n$ is an empirical risk minimizer, then, with probability larger than $1-e^{-x}$,

$$\ell(\hat{\mathbf{c}}_n, \mathbf{c}^*) \leq C_0'\kappa_0 \frac{\left|\bar{\mathscr{M}}\right|^2 M^2 k}{n} + \kappa_0 \frac{144M^2}{n}x + \frac{64M^2}{n}x, \tag{3.6}$$

where C_0' is an absolute constant.

Moreover, if $\mathscr{H} = \mathbb{R}^d$, with probability larger than $1-e^{-x}$,

$$\ell(\hat{\mathbf{c}}_n, \mathbf{c}^*) \leq C_0\kappa_0 \frac{M^2kd\left(\log(4|\bar{\mathscr{M}}|\sqrt{kd}) + 1\right)}{n} + \kappa_0 \frac{144M^2}{n}x + \frac{64M^2}{n}x, \tag{3.7}$$

where C_0 is an absolute constant.

In addition, if $\mathscr{H} = \mathbb{R}^d$, then, with probability larger than $1-2e^{-x}$,

$$\ell(\hat{\mathbf{c}}_n, \mathbf{c}^*) \leq C_0''\kappa_0 \frac{|\mathscr{M}|^2 R(\mathbf{c}^*)}{n} + \kappa_0^3 \frac{C_1}{n^2} + \kappa_0 \frac{C_2}{n}x + \frac{C_3}{n}x, \tag{3.8}$$

where C_0'' is an absolute constant, C_1 is a combination of square roots of polynomial functions in k, $\log(k)$, d, B and M, C_2 is polynomial in k and M, C_3 is polynomial in M and $\sqrt{k}$.

This result is in line with Theorem 2.1 in Chapter 2 or Theorem 1 in [CL06], concerning the dependency on the sample size n of the loss $\ell(\hat{\mathbf{c}}_n, \mathbf{c}^*)$. The main advance lies in the dependency on other parameters of the loss of $\hat{\mathbf{c}}_n$, which provides a non-asymptotic bound for the excess risk.

At first sight, in the finite dimensional case, (3.6) seems to outperform (3.7) when d is large. However the dependency on the number of optimal codebooks is dramatically worse in (3.6) than in (3.7). This difference can be explained by the two different methods used to derive these bounds.

In fact, most of the proof of (3.7) relies on the application of Dudley's entropy bound. As exposed in Section 2.5.5, this technique was already the main argument in [Pol82b] or [CL06], and makes a classical dimension factor kd appear. This result slightly improves the asymptotic bound exposed in [Pol82b], since it offers an explicit calculation of the metric entropy used to derive this result.

As suggested in [BDL08], the use of metric entropy techniques to derive bounds on the convergence rate of the distortion may be suboptimal, as it does not take advantage of the Hilbert space structure of the squared distance based quantization.

This issue can be addressed using a more general chaining technique based on comparison with Gaussian vectors, such as the generic chaining principle developed in [Tal05]. The second upper bound (3.6) is derived that way.

The third upper bound mentioned in Theorem 3.1 may be thought of as a semi asymptotic result. Indeed, inequality (3.8) ensures that, in the finite dimensional case,

$$\limsup_{n\to\infty} n\mathbb{E}\ell(\hat{\mathbf{c}}_n, \mathbf{c}^*) \lesssim |\mathscr{M}|^2 R(\mathbf{c}^*).$$

In fact, when considering only the first term of the right side of (3.8), the dependency of the expected distortion on the parameters k and d seems to be more accurate. However, the dependency on k and d of terms with higher order of $1/n$ in (3.8) are dramatically worse than in (3.7) or (3.6). Therefore it is likely that, whenever d is large compared to n, the result (3.8) is outperformed by (3.7) and (3.6).

Another interesting point is that Theorem 3.1 does not require that P has a density or is distributed over points, contrary to the requirements of the previous bounds in [Pol82b], [AGG05] or [CL06] which achieved the optimal rate of $\mathscr{O}(1/n)$. Up to our knowledge, the more general result on this topic is to be found in Theorem 2 of [AGG05], which derives a convergence rate of $\mathscr{O}(\log(n)/n)$ without any requirement on the regularity of the distribution P. It may also be noted that, in the finite dimensional case, contrary to the results exposed in [Pol82b], Theorem 3.1 does not require that $\bar{\mathscr{M}}$ contains a single element. According to Proposition 3.2, only (3.3) has to be proved for P to satisfy the assumptions of Theorem 3.1. Since proving that $|\bar{\mathscr{M}}| = 1$ may be difficult, even for simple distributions, it seems easier to check the assumptions of Theorem 3.1 than the assumptions required in [Pol82b]. An illustration of this point is given in Section 3.3.3.

It is also worth mentioning that the dependency in ε surprisingly turns out to be sharp when $\varepsilon \sim n^{-1/2}$, as will be shown in Proposition 3.3. In fact, tuning this separation factor is the core of the demonstration of the minimax results in [BLL98] or [Ant05].

3.3.2 Minimax lower bound

Theorem 1 in [BLL98] ensures that the minimax convergence rate over the M-bounded distributions of any empirically designed codebook can be bounded from below by $\Omega(1/\sqrt{n})$. A question of interest is to know whether this lower bound can be refined when considering only distributions satisfying some fast convergence condition. A partial answer is given by Corollary 2 in [Ant05], where it is proved that the minimax rate over distributions with uniformly bounded continuous densities, unique optimal codebook (up to relabeling), and such that the second derivative matrices at the optimal codebook $H(\mathbf{c}^*)$ are positive definite, is still $\Omega(1/\sqrt{n})$. According to [Ant05], a natural question is to know whether a uniform upper bound of the type $o(1/\sqrt{n})$ may be derived, with the additional requirement that the minimum eigenvalue of the second derivative matrices $H(\mathbf{c}^*)$ is uniformly bounded from below.

This Subsection is devoted to obtaining a minimax lower bound on the excess risk over a set of distributions with continuous densities, unique optimal codebook, and satisfying the margin condition defined in Definition 3.1, in which some parameters, such as p_{min} are fixed or uniformly lower-bounded. Since, in this case, the minimum eigenvalues of $H(\mathbf{c}^*)$ are larger than $p_{min}/2$, such a minimax lower bound provides an answer to the question mentioned above.

Throughout this Subsection, only the case $\mathscr{H} = \mathbb{R}^d$ is considered, and $\hat{\mathbf{c}}_n$ will denote an empirically designed codebook, that is a map from $(\mathbb{R}^d)^n$ to $(\mathbb{R}^d)^k$. Let k be an integer such that $k \geq 3$, and $M > 0$. For simplicity, k is assumed to be divisible by 3. Let us introduce the following quantities :

$$\begin{cases} m &= \frac{2k}{3}, \\ \Delta &= \frac{5M}{32m^{1/d}}. \end{cases}$$

To focus on the dependency on the separation factor ε, the quantities involved in Definition 3.1 are fixed as :

$$\begin{cases} B &= \Delta, \\ r_0 &= \frac{7\Delta}{16}, \\ p_{min} &\geq \frac{1}{2k}. \end{cases} \tag{3.9}$$

Denote by $\mathscr{D}(\varepsilon)$ the set of probability distributions which are ε-separated, have a continuous density and a unique optimal codebook, and which satisfy a margin condition with parameters defined in (3.9). The minimax result is the following :

Proposition 3.3. *Assume that $k \geq 3$. Then, for any empirically designed codebook,*

$$\sup_{P \in \mathscr{D}(c_1/\sqrt{n})} \mathbb{E}\ell(\hat{\mathbf{c}}_n, \mathbf{c}^*) \geq c_0 M^2 \frac{\sqrt{k^{1-\frac{4}{d}}}}{\sqrt{n}},$$

where c_0 is an absolute constant, and

$$c_1 = \frac{(5M)^2}{4(32m^{\frac{1}{4}+\frac{1}{d}})^2}.$$

Proposition 3.3 is in line with the previous minimax lower bounds obtained in Theorem 1 of [BLL98] or Theorem 4 in [Ant05]. In fact, the classes of distributions used in both these results satisfy a uniform margin condition, without specification of the separation factor. Proposition 3.3, as well as these two previous results, emphasizes the fact that fixing the parameters of the margin condition uniformly over a class of distributions does not guarantee an optimal uniform convergence rate. This shows that a uniform separation assumption is needed to derive a fast uniform convergence rate over a set of distributions.

Furthermore, as mentioned above, Proposition 3.3 also proves that the minimax distortion rate over the set of distributions with continuous densities, unique optimal codebook, and such that the minimum eigenvalues of the Hessian matrices $H(\mathbf{c}^*)$ are uniformly lower bounded by $1/4k$, is still $\Omega(1/\sqrt{n})$.

This minimax lower bound has to be compared to the upper risk bound obtained in Theorem 3.1 for the empirical risk minimizer $\hat{\mathbf{c}}_n$ over the set of distributions $\mathscr{D}(c_1/\sqrt{n})$. To be more precise, Theorem 3.1 ensures that, provided that n is large enough,

$$\sup_{P \in \mathscr{D}(c_1/\sqrt{n})} \mathbb{E}\ell(\hat{\mathbf{c}}_n, \mathbf{c}^*) \leq \frac{g(k,d,M)}{\sqrt{n}},$$

where $g(k,d,M)$ depends only on k, d and M. In other words, the dependency of the upper bounds stated in Theorem 3.1 on ε turns out to be sharp whenever $\varepsilon \sim n^{-\frac{1}{2}}$. Unfortunately, Proposition 3.3 can not be easily extended to the case where $\varepsilon \sim n^{-\alpha}$, with $0 < \alpha < 1/2$. Consequently an open question is whether the upper bounds stated in Theorem 3.1 remains accurate with respect to ε in this case.

3.3.3 Quasi-Gaussian mixture example

The aim of this Subsection is to illustrate the results offered in Section 3.3 with Gaussian mixtures in dimension $d = 2$. The Gaussian mixture model is a typical and well-defined clustering example. However we will not deal with the clustering issue but rather with its theoretical background.

In general, a Gaussian mixture distribution $\tilde{P}$ is defined by its density

$$\tilde{f}(x) = \sum_{i=1}^{\tilde{k}} \frac{\theta_i}{2\pi\sqrt{|\Sigma_i|}} e^{-\frac{1}{2}(x-m_i)^t \Sigma_i^{-1}(x-m_i)},$$

where $\tilde{k}$ denotes the number of components of the mixture, and the θ_i's denote the weights of the mixture, which satisfy $\sum_{i=1}^{k} \theta_i = 1$. Moreover, the m_i's denote the means of the mixture, so that $m_i \in \mathbb{R}^2$, and the Σ_i's are the 2×2 variance matrices of the components.

We restrict ourselves to the case where the number of components $\tilde{k}$ is known, and match the size k of the codebooks. To ease the calculation, we make the additional assumption that every component has the same diagonal variance matrix $\Sigma_i = \sigma^2 I_2$. Note that a similar result to Proposition 3.4 can be derived for distributions with different variance matrices Σ_i, at the cost of more computing.

Since the support of a Gaussian random variable is not bounded, we define the "quasi-Gaussian" mixture model as follows, truncating each Gaussian component. Let the density f of the distribution P be defined by

$$f(x) = \sum_{i=1}^{k} \frac{\theta_i}{2\pi\sigma^2 N_i} e^{-\frac{\|x-m_i\|^2}{2\sigma^2}} \mathbb{1}_{\mathscr{B}(0,M)},$$

where N_i denotes a normalization constant for each Gaussian variable.

Let ε be defined as $\varepsilon = 1 - \min_{i=1,\dots,k} N_i$. Roughly, the model proposed above will be close to the Gaussian mixture model when ε is small. Let $\tilde{B}$ denote the smallest possible distance between two different means of the mixture, that is $\tilde{B} = \inf_{i \neq j} \|m_i - m_j\|$. To avoid boundary issues we assume that, for all $i = 1, \dots, k$, $\mathscr{B}(m_i, \tilde{B}/3) \subset \mathscr{B}(0, M)$.

It is worth noticing that the assumption $\mathscr{B}(m_i, \tilde{B}/3) \subset \mathscr{B}(0, M)$ can easily be satisfied as soon as M is chosen large enough. For such a model, Proposition 3.4 below offers a sufficient condition for P to satisfy a margin condition.

Proposition 3.4. *Let $\theta_{min} = \min_{i=1,\dots,k} \theta_i$, and $\theta_{max} = \max_{i=1,\dots,k} \theta_i$. Assume that*

$$\frac{\theta_{min}}{\theta_{max}} \geq \max\left(\frac{2048 k \sigma^2}{(1-\varepsilon)\tilde{B}^2 (1 - e^{-\tilde{B}^2/2048\sigma^2})}, \frac{2048 k^2 M^3}{(1-\varepsilon) 7\sigma^2 \tilde{B} (e^{\tilde{B}^2/32\sigma^2} - 1)} \right). \tag{3.10}$$

Then P satisfies a margin condition with radius $\frac{\tilde{B}}{8}$.

It is worth mentioning that P has a continuous density, and that, according to Proposition 3.2, the second derivative matrices of the risk function, at the optimal codebooks, must be positive definite. Thus, P might be in the scope of the result in [Pol82b]. However, there is no elementary proof of the fact that $|\bar{\mathscr{M}}| = 1$, whereas $\mathscr{M}$ is finite is guaranteed by Proposition 3.2. This shows that the margin condition

given in Definition 3.1 may be easier to check than the condition offered in [Pol82b]. The condition (3.10) can be decomposed as follows. If

$$\frac{\theta_{min}}{\theta_{max}} \geq \frac{2048k\sigma^2}{(1-\varepsilon)\tilde{B}^2(1-e^{-\tilde{B}^2/2048\sigma^2})},$$

then every optimal codebook $\mathbf{c}^*$ must be close to the vector of means of the mixture $\mathbf{m} = (m_1, \ldots, m_k)$. Therefore, it is possible to approximately locate N^*, and to derive an upper bound on the weight function $p(t)$ defined in Definition 3.1. This leads to the second term of the maximum in (3.10).

This condition can be interpreted as a condition on the polarization of the mixture. A favorable case for vector quantization seems to be when the poles of the mixtures are well separated, which is equivalent to σ is small compared to $\tilde{B}$, when considering Gaussian mixtures. Proposition 3.4 gives details on how σ has to be small compared to $\tilde{B}$, in order to satisfy the requirements of Proposition 3.2. This ensures that the loss $\ell(\hat{\mathbf{c}}_n, \mathbf{c}^*)$ reaches an improved convergence rate of $1/n$.

It may be noticed that Proposition 3.4 offers almost the same condition than Proposition 2.3 in Chapter 2. In fact, since the Gaussian mixture distributions have a continuous density, making use of (3.4) in Example 3 ensures that the margin condition for Gaussian mixtures is equivalent to a bound on the density over the critical region N^*.

It is important to note that this result is valid when k is known and match exactly the number of components of the mixture. When the number of code points k is different from the number of components $\tilde{k}$ of the mixture, we have no general idea of where the optimal code points can be located.

Moreover, suppose that there exists only one optimal codebook $\mathbf{c}^*$, up to relabeling, and that we are able to locate this optimal codebook $\mathbf{c}^*$. As mentioned in Proposition 3.2, the key quantity is in fact $B = \inf_{i \neq j} \|c_i^* - c_j^*\|$. In the case where $\tilde{k} \neq k$, there is no simple relation between $\tilde{B}$ and B. Consequently, a condition like in Proposition 3.4 could not involve the natural parameter of the mixture $\tilde{B}$.

It is also worth pointing out that there exist cases where the set of optimal codebooks is not finite. For example, assume that P is a truncated rotationally symmetric Gaussian distribution, and $k = 2$. Since every rotation of an optimal codebook leads to another optimal codebook, there exists an infinite set of optimal codebooks. Since, in this case, $N^* = \mathbb{R}^2$, condition (3.3) can not be satisfied.

3.4 Proofs

3.4.1 Proof of Proposition 3.1

The lower bound on B follows from a compactness argument for the weak topology on $\mathscr{H}$, stated in the following lemma. For the sake of completeness, it is recalled that a sequence c_n of elements in $\mathscr{H}$ weakly converges to c, denoted by $c_n \rightharpoonup_{n\to\infty} c$, if, for every continuous linear real-valued function f, $f(c_n) \to_{n\to\infty} f(c)$. Moreover, a function ϕ from $\mathscr{H}$ to $\mathbb{R}$ is weakly lower semi-continuous if, for all $\lambda \in \mathbb{R}$, the level sets $\{c \in \mathscr{H} | \quad \phi(c) \leq \lambda\}$ are closed for the weak topology.

Lemma 3.1. *Let $\mathscr{H}$ be a separable Hilbert space, and suppose that P is M-bounded. Then*

i) $\mathscr{B}(0,M)^k$ is weakly compact,

ii) $\mathbf{c} \mapsto P\gamma(\mathbf{c},.)$ is weakly lower semi-continuous.

Proof of Lemma 3.1. A more general statement of Lemma 3.1 can be found in Section 5.2 of [Fis10], for quantization with Bregman divergences. However, since the proof is much more simple in the special case of the squared-norm based quantization on a Hilbert space, it is briefly recalled here.

Since $\mathscr{H}$ is reflexive, according to Banach-Alaoglu-Bourbaki's Theorem (see, e.g., Theorem 3.16 in [Bre11]), combined with Tychonoff's Theorem (see, e.g., Theorem 2.2.8 in [Dud02]), $\mathscr{B}(0,M)^k$ is a compact subset of $\mathscr{H}^k$ for the weak topology. This proves *i)*.

Let x be a fixed element of $\mathscr{H}^k$. Since $\mathbf{c} \mapsto \|x - c_i\|^2$ is weakly lower semi-continuous (see, e.g., Proposition 3.13 in [Bre11]), $\mathbf{c} \mapsto \gamma(\mathbf{c},x)$ is weakly lower semi-continuous over $\mathscr{B}(0,M)^k$. Let $\mathbf{c}_n$ be a sequence of $\mathscr{B}(0,M)^k$ such that $\mathbf{c}_n \rightharpoonup_{n\to\infty} \mathbf{c}$, for the weak topology, for some $\mathbf{c} \in \mathscr{B}(0,M)^k$. Then

$$\gamma(\mathbf{c},x) \leq \liminf_{n\to\infty} \gamma(\mathbf{c}_n,x).$$

Applying Fatou's Lemma (see, e.g., Lemma 4.3.3 in [Dud02]) yields

$$R(\mathbf{c}) \leq \liminf_{n\to\infty} R(\mathbf{c}_n).$$

Hence *ii)* is proved. It is worth noting that this proves the existence of optimal codebooks for bounded distributions. □

Let $\mathbf{c}'_n$ be a sequence of optimal codebooks such that $\|c'_{1,n} - c'_{2,n}\| \to B$, as $n \to \infty$. Then, according to Lemma 3.1, there exists a subsequence $\mathbf{c}_n$ and an optimal codebook $\mathbf{c}^*$, such that $\mathbf{c}_n \rightharpoonup_{n\to\infty} \mathbf{c}^*$, for the weak topology. Then it is clear that $(c_{1,n} - c_{2,n}) \rightharpoonup_{n\to\infty} (c^*_1 - c^*_2)$.

Since $u \mapsto \|u\|$ is weakly lower semi-continuous on $\mathscr{H}$ (see, e.g., Proposition 3.13 in [Bre11]), it follows that

$$\|c^*_1 - c^*_2\| \leq \liminf_{n\to\infty} \|c_{1,n} - c_{2,n}\| = B.$$

Noting that $\mathbf{c}^*$ is an optimal codebook, and the support of P has more than k points, Proposition 1 of [GLP07] ensures that $\|c^*_1 - c^*_2\| > 0$.

The uniform lower bound on p_{min} follows from the argument that, since the support of P contains more than k points, then $R^*_k < R^*_{k-1}$, where R^*_j denotes the minimum distortion achievable for j-points quantizers (see, e.g., Proposition 1 in [GLP07]). Denote by α the quantity $R^*_{k-1} - R^*_k$, and suppose that $p_{min} < \frac{\alpha}{4M^2}$. Then there exists an optimal codebook of size k, $\mathbf{c}^{*,k} = (c^{*,k}_1, \ldots, c^{*,k}_k)$, such that $p_{min} = P(V_1(\mathbf{c}^{*,k}))$. Let $\mathbf{c}^{*,k-1}$ denote an optimal codebook of size $(k-1)$, and define the following k-points quantizer

$$\begin{cases} Q(x) = c^{*,k}_1 & \text{if} \quad x \in V_1(\mathbf{c}^{*,k}), \\ Q(x) = c^{*,k-1}_j & \text{if} \quad x \in V_j(\mathbf{c}^{*,k-1}) \cap \left(V_1(\mathbf{c}^{*,k})\right)^c. \end{cases}$$

Since $P(\partial V_1(\mathbf{c}^{*,k})) = P(\partial V_j(\mathbf{c}^{*,k-1})) = 0$, for $j = 1, \ldots, k-1$, Q is defined P almost surely. Then it is easy to see that

$$R(Q) \leq p_{min} 4M^2 + R^*_{k-1} < R^*_k.$$

Hence the contradiction. Therefore we have $p_{min} \geq \frac{\alpha}{4M^2}$.

3.4.2 Proof of Proposition 3.2

According to Lemma 3.1, $\mathscr{M}$ is weakly compact. Since, according to Proposition 3.13 in [Bre11], $\mathbf{c}^* \mapsto \|\mathbf{c} - \mathbf{c}^*\|$ is weakly lower semi continuous, its minimum is attained over $\mathscr{M}$. This proves i).

The proof of ii) is based on the following lemma.

Lemma 3.2. *Let* $\mathbf{c}$ *and* $\mathbf{c}^*$ *be in* $\mathscr{B}(0,M)^k$, *and* $x \in V_i(\mathbf{c}^*) \cap V_j(\mathbf{c}) \cap \mathscr{B}(0,M)$, *for* $i \neq j$. *Then*

$$\left| \left\langle x - \frac{c_i + c_j}{2}, c_i - c_j \right\rangle \right| \leq 4\sqrt{2}M \|\mathbf{c} - \mathbf{c}^*\|, \tag{3.11}$$

$$d(x, \partial V_i(\mathbf{c}^*)) \leq \frac{4\sqrt{2}M}{B} \|\mathbf{c} - \mathbf{c}^*\|. \tag{3.12}$$

The two statements of Lemma 3.2 emphasize the fact that, provided that $\mathbf{c}$ and $\mathbf{c}^*$ are quite similar, the areas on which the label may differ with respect to $\mathbf{c}$ and $\mathbf{c}^*$ should be close to the boundary of Voronoi diagrams. This idea is mentioned in the proof of Corollary 1 in [AGG05]. Nevertheless we provide here a simpler proof.

Proof of Lemma 3.2. Let x be in $V_i(\mathbf{c}^*) \cap V_j(\mathbf{c}) \cap \mathscr{B}(0,M)$, then $\|x - c_j\|^2 \leq \|x - c_i\|^2$, which leads to $\left\langle c_i - c_j, x - \frac{c_i + c_j}{2} \right\rangle \leq 0$. Since $\|x - c_i^*\| \leq \|x - c_j^*\|$, we may write

$$\|x - c_i\| \leq \|x - c_j\| + \|c_i - c_i^*\| + \|c_j - c_j^*\|.$$

Taking square on both sides leads to

$$\begin{aligned} \|x - c_i\|^2 - \|x - c_j\|^2 &\leq 2\|x - c_j\|(\|c_i - c_i^*\| + \|c_j - c_j^*\|) \\ &\quad + \left(\|c_i - c_i^*\| + \|c_j - c_j^*\|\right)^2 \\ &\leq 8M(\|c_i - c_i^*\| + \|c_j - c_j^*\|) \\ &\leq 8\sqrt{2}M\|\mathbf{c} - \mathbf{c}^*\|. \end{aligned}$$

Since $\|x - c_i\|^2 - \|x - c_j\|^2 = -2\left\langle x - \frac{c_i + c_j}{2}, c_i - c_j \right\rangle$, (3.11) is proved.

To prove (3.12), remark that, since $x \in V_i(\mathbf{c}^*)$, $d(x, \partial V_i(\mathbf{c}^*)) \leq d(x, h_{i,j}^*)$, where $h_{i,j}^*$ is the hyperplane defined by $\left\{x \in \mathscr{B}(0,M) | \|x - c_i^*\| = \|x - c_j^*\|\right\}$. Using quite simple geometric arguments, we deduce that

$$d(x, h_{i,j}^*) = \left| \left\langle x - \frac{c_i^* + c_j^*}{2}, \frac{c_i^* - c_j^*}{\left\|c_i^* - c_j^*\right\|} \right\rangle \right|.$$

The same arguments as in the proof of (3.11) guarantee that

$$\begin{aligned} \left| \left\langle x - \frac{c_i^* + c_j^*}{2}, \frac{c_i^* - c_j^*}{\left\|c_i^* - c_j^*\right\|} \right\rangle \right| &= \left\langle x - \frac{c_i^* + c_j^*}{2}, \frac{c_i^* - c_j^*}{\left\|c_i^* - c_j^*\right\|} \right\rangle \\ &\leq \frac{4\sqrt{2}M}{B} \|\mathbf{c} - \mathbf{c}^*\|. \end{aligned}$$

□

Equipped with Lemma 3.2, we are in a position to prove $ii)$. Let $\mathbf{c}$ be in $\mathscr{B}(0,M)^k$, and $(W_1(\mathbf{c}),\dots,W_k(\mathbf{c}))$ be a Voronoi partition associated with $\mathbf{c}$, as defined in Section 3.2. Let $\mathbf{c}^*$ be in $\mathscr{M}$, then $\ell(\mathbf{c},\mathbf{c}^*)$ can be decomposed as follows :

$$\begin{aligned} P\gamma(\mathbf{c},.) &= \sum_{i=1}^{k} P(\|x-c_i\|^2 \mathbb{1}_{W_i(\mathbf{c})}) \\ &= \sum_{i=1}^{k} P(\|x-c_i\|^2 \mathbb{1}_{V_i(\mathbf{c}^*)}) + \sum_{i=1}^{k} P(\|x-c_i\|^2 (\mathbb{1}_{W_i(\mathbf{c})} - \mathbb{1}_{V_i(\mathbf{c}^*)})). \end{aligned}$$

Since, for all $i=1,\dots k$, $P(x\mathbb{1}_{V_i(\mathbf{c}^*)}(x)) = P(V_i(\mathbf{c}^*))c_i^*$ (centroid condition), we may write

$$P(\|x-c_i\|^2 \mathbb{1}_{V_i(\mathbf{c}^*)}) = P(V_i(\mathbf{c}^*))\|c_i - c_i^*\|^2 + P(\|x-c_i^*\|^2 \mathbb{1}_{V_i(\mathbf{c}^*)}),$$

from which we deduce

$$P\gamma(\mathbf{c},.) = P\gamma(\mathbf{c}^*,.) + \sum_{i=1}^{k} P(V_i(\mathbf{c}^*))\|c_i - c_i^*\|^2 + \sum_{i=1}^{k} P(\|x-c_i\|^2 (\mathbb{1}_{W_i(\mathbf{c})} - \mathbb{1}_{V_i(\mathbf{c}^*)})),$$

which leads to

$$\ell(\mathbf{c},\mathbf{c}^*) \geq p_{min}\|\mathbf{c}-\mathbf{c}^*\|^2 + \sum_{i=1}^{k}\sum_{j\neq i} P\left((\|x-c_j\|^2 - \|x-c_i\|^2)\mathbb{1}_{V_i(\mathbf{c}^*)\cap W_j(\mathbf{c})}\right).$$

Since $x \in W_j(\mathbf{c}) \subset V_j(\mathbf{c})$, $\|x-c_j\|^2 - \|x-c_i\|^2 \leq 0$. Thus it remains to bound from above

$$\sum_{i=1}^{k}\sum_{j\neq i} P\left((\|x-c_i\|^2 - \|x-c_j\|^2)\mathbb{1}_{V_i(\mathbf{c}^*)\cap W_j(\mathbf{c})}\right).$$

Noticing that

$$\|x-c_i\|^2 - \|x-c_j\|^2 = 2\left\langle c_j - c_i, x - \frac{c_i+c_j}{2}\right\rangle,$$

and using Lemma 3.2, we get

$$\sum_{i=1}^{k} P(\|x-c_i\|^2(\mathbb{1}_{W_i(\mathbf{c})} - \mathbb{1}_{V_i(\mathbf{c}^*)})) \geq -8\sqrt{2}M\|\mathbf{c}-\mathbf{c}^*\| p\left(\frac{4\sqrt{2}M}{B}\|\mathbf{c}-\mathbf{c}^*\|\right).$$

Consequently, if P satisfies (3.3), then, if $\|\mathbf{c}-\mathbf{c}^*\| \leq \frac{Br_0}{4\sqrt{2}M}$,

$$\ell(\mathbf{c},\mathbf{c}^*) \geq \frac{p_{min}}{2}\|\mathbf{c}-\mathbf{c}^*\|^2. \tag{3.13}$$

Now turn to the case where $\|\mathbf{c}-\mathbf{c}^*(\mathbf{c})\| \geq \frac{Br_0}{4\sqrt{2}M}$. Let $\mathscr{B}^o(\mathbf{c}^*,r)$ denote the open ball with radius r and center $\mathbf{c}^*$. Since $\mathscr{B}(0,M)^k \cap \left(\bigcup_{\mathbf{c}^*\in\mathscr{M}} \mathscr{B}^o(\mathbf{c}^*, \frac{Br_0}{4\sqrt{2}M})\right)^c$ is weakly compact, according to Lemma 3.1, and $\mathbf{c} \longmapsto P\gamma(\mathbf{c},.)$ is weakly lower semi-continuous, its minimum over this set is attained. Such a minimum is a local minimum, or is at the boundary $\|\mathbf{c}-\mathbf{c}^*(\mathbf{c})\| = \frac{Br_0}{4\sqrt{2}M}$. Hence we deduce

$$\begin{aligned} \ell(\mathbf{c},\mathbf{c}^*) &\geq \varepsilon \wedge \frac{p_{min}B^2r_0^2}{64M^2} \\ &\geq \left(\varepsilon \wedge \frac{p_{min}B^2r_0^2}{64M^2}\right)\frac{\|\mathbf{c}-\mathbf{c}^*(\mathbf{c})\|^2}{4kM^2}. \end{aligned}$$

Note that, since $B \le 2M$ and $r_0 \le 2M$, $\left(\varepsilon \wedge \frac{p_{min} B^2 r_0^2}{64M^2}\right)/4kM^2 \le p_{min}/2$. This proves ii).

In the case where $\mathscr{H} = \mathbb{R}^d$, the weak topology coincides with the usual topology. Consequently $\mathscr{B}(0,M)^k$ is compact. Suppose that $\mathscr{M}$ is not finite. Then there exists a sequence $\mathbf{c}_n$ of optimal codebooks, and an optimal codebook $\mathbf{c}^*$, such that $\|\mathbf{c}_n - \mathbf{c}^*\| \to_{n\to\infty} 0$. For n large enough, we have

$$\|\mathbf{c}_n - \mathbf{c}^*\| \le \frac{Br_0}{4\sqrt{2}M},$$

and $\ell(\mathbf{c}_n, \mathbf{c}^*) = 0$. This contradicts (3.13).

3.4.3 Proof of Theorem 3.1

Throughout this Subsection P is assumed to satisfy a margin condition with radius r_0, and to be ε-separated. A non decreasing map $\Phi : \mathbb{R}^+ \to \mathbb{R}^+$ is called subroot if $x \mapsto \frac{\Phi(x)}{\sqrt{x}}$ is non increasing.

The following localization Theorem, derived from Theorem 6.1 in [BBM08], is the main argument of our proof.

Theorem 3.2. *Let $\mathscr{F}$ be a class of bounded measurable functions such that there exist $b > 0$ and $\omega : \mathscr{F} \longrightarrow \mathbb{R}^+$ satisfying*

(i) $\forall f \in \mathscr{F} \quad \|f\|_\infty \le b$,

(ii) $\forall f \in \mathscr{F} \quad \mathrm{Var}(f) \le \omega(f)$.

Let K be a positive constant, Φ a sub-root function. Then if r^* is the unique solution of the equation $\Phi(r) = r/24K$, the following holds. Assume that

$$\forall r \ge r^* \qquad \mathbb{E}\left(\sup_{\omega(f)\le r} |(P - P_n)f|\right) \le \Phi(r).$$

Then, for all $x > 0$, with probability larger than $1 - e^{-x}$,

$$\forall f \in \mathscr{F} \quad Pf - P_n f \le K^{-1}\left(\omega(f) + r^* + \frac{(9K^2 + 16Kb)x}{4n}\right).$$

A proof of Theorem 3.2 is given in Section 5.3 of [Lev13], which corresponds here to Section 2.5.4.

3.4.3.1 Proof of (3.7)

We begin with the finite dimensional case. The proof of (3.7) follows from the combination of Proposition 3.2 and a direct application of Theorem 3.2. To be more precise, let $\mathscr{F}$ denote the set

$$\mathscr{F} = \left\{\gamma(\mathbf{c},.) - \gamma(\mathbf{c}^*(\mathbf{c}),.)| \quad \mathbf{c} \in \mathscr{B}(0,M)^k\right\}.$$

Since, for all $i \in \{1,\dots,k\}$,

$$\left|\|x - c_i\|^2 - \|x - c_i^*(\mathbf{c})\|^2\right| \le 4M\|c_i - c_i^*(\mathbf{c})\|,$$

it follows that, for every $f \in \mathscr{F}$,

$$\begin{cases} \|f\|_\infty & \le & 8M^2, \\ \mathrm{Var}_P(f) & \le & 16M^2\|\mathbf{c} - \mathbf{c}^*(\mathbf{c})\|^2. \end{cases}$$

Define $\omega(f) = 16M^2\|\mathbf{c} - \mathbf{c}^*(\mathbf{c})\|^2$. It remains to bound from above the complexity term. This is done in the following Proposition, derived from the proof of Theorem 1 in [CL06].

Proposition 3.5. *One has*

$$\mathbb{E} \sup_{f \in \mathscr{F}, \omega(f) \leq \delta} |(P - P_n)f| \leq \frac{(2\sqrt{2} + 64)\sqrt{kd}}{\sqrt{n}} \left(\sqrt{\log(4|\bar{\mathscr{M}}|\sqrt{kd})} + 1 \right) \sqrt{\delta}. \tag{3.14}$$

The proof of Proposition 3.5 derives from classical chaining arguments, and can be found in Section 3.5.1. Let Φ be defined as the right-hand side of (3.14). Observing that $\Phi(\delta)$ takes the form $\Phi(\delta) = \Xi\sqrt{\delta/n}$, the solution δ^* of the equation $\Phi(\delta) = \delta/24K$ may be written, for any $K > 0$,

$$\delta^* = \frac{K^2\Xi^2}{n}.$$

Applying Theorem 3.2 to $\mathscr{F}$ leads to, with probability larger than $1 - e^{-x}$,

$$(P - P_n)(\gamma(\mathbf{c}, .) - \gamma(\mathbf{c}^*(\mathbf{c}), .)) \leq K^{-1}16M^2\|\mathbf{c} - \mathbf{c}^*(\mathbf{c})\|^2 + \frac{K\Xi^2}{n} + \frac{9K + 128M^2}{4n}x.$$

Introducing the inequality $\kappa_0 \ell(\mathbf{c}, \mathbf{c}^*) \geq \|\mathbf{c} - \mathbf{c}^*(\mathbf{c})\|^2$ provided by Proposition 3.2, choosing $K = 32M^2\kappa_0$ leads to (3.7).

3.4.3.2 Proof of (3.6)

Similarly to the proof of (3.7), the proof of (3.6) is based on an application of Theorem 3.2 to the set $\mathscr{F}$, defined in the above Subsection. However, the technique used to bound the complexity term is slightly different, and leads to the following result.

Proposition 3.6. *One has*

$$\mathbb{E} \sup_{f \in \mathscr{F}, \omega(f) \leq \delta} |(P - P_n)f| \leq \left|\bar{\mathscr{M}}\right| \frac{4\sqrt{\pi k}}{\sqrt{n}} \sqrt{\delta}. \tag{3.15}$$

This proof relies on the use of Gaussian complexities combined with Slepian's Lemma (see, e.g., Theorem 3.14 in [Mas07]), as done in [CPR12]. We postpone it to the following Subsection. Let Φ' be defined as the right-hand side of (3.15), and let δ' denote the solution of the equation $\Phi'(\delta) = \delta/24K$, for some positive $K > 0$. Then δ' can be expressed as

$$\delta' = C\frac{\left|\bar{\mathscr{M}}\right|^2 K^2 k}{n},$$

where $C = 9216\pi$. As in the proof of (3.7), choosing $K = 32\kappa_0 M^2$, applying Theorem 3.2 and combining it with Proposition 3.2 leads to the result.

3.4.3.3 Proof of Proposition 3.6

As mentioned above, this proof relies on the use of Gaussian complexities (see, e.g., [BM02]). As will be shown below, avoiding Dudley's entropy argument by introducing some Gaussian random vectors allows us to take advantage of the underlying Hilbert space structure. The first step is to decompose the complexity term according to optimal codebooks, in the following way

$$\mathbb{E} \sup_{\|\mathbf{c}-\mathbf{c}^*(\mathbf{c})\|^2 \leq \delta/16M^2} \left|(P-P_n)(\gamma(\mathbf{c},.)-\gamma(\mathbf{c}^*(\mathbf{c}),.))\right| \leq \sum_{\mathbf{c}^* \in \bar{\mathcal{M}}} \mathbb{E} \sup_{\|\mathbf{c}-\mathbf{c}^*\|^2 \leq \delta/16M^2} \left|(P-P_n)(\gamma(\mathbf{c},.)-\gamma(\mathbf{c}^*,.))\right|.$$

Next we bound from above every term of the right-hand side. Let $\mathbf{c}^*$ be fixed, and let $\sigma_1,\ldots,\sigma_n$ denote some independent Rademacher variables. According to the symmetrization principle (see, e.g., Section 2.2 of [Kol06]),

$$\mathbb{E} \sup_{\|\mathbf{c}-\mathbf{c}^*\|^2 \leq \delta/16M^2} \left|(P-P_n)(\gamma(\mathbf{c},.)-\gamma(\mathbf{c}^*,.))\right| \leq 2\mathbb{E}_{X,\sigma} \sup_{\|\mathbf{c}-\mathbf{c}^*\|^2 \leq \delta/16M^2} \frac{1}{n}\sum_{i=1}^n \sigma_i(\gamma(\mathbf{c},X_i)-\gamma(\mathbf{c}^*,X_i)),$$

where $\mathbb{E}_Y$ denotes integration with respect to the distribution of Y. Let $g_1,\ldots,g_n$ denote some independent standard Gaussian variables. Applying Lemma 4.5 in [LT91] leads to

$$\mathbb{E}_{X,\sigma} \sup_{\|\mathbf{c}-\mathbf{c}^*\|^2 \leq \delta/16M^2} \frac{1}{n}\sum_{i=1}^n \sigma_i(\gamma(\mathbf{c},X_i)-\gamma(\mathbf{c}^*,X_i)) \leq \sqrt{\frac{\pi}{2}}\mathbb{E}_{X,g} \sup_{\|\mathbf{c}-\mathbf{c}^*\|^2 \leq \delta/16M^2} \frac{1}{n}\sum_{i=1}^n g_i(\gamma(\mathbf{c},X_i)-\gamma(\mathbf{c}^*,X_i)).$$

To derive bounds on the Gaussian complexity defined above, the following comparison result between Gaussian processes is needed.

Theorem 3.3 (Slepian's Lemma)**.** *Let X_t and Z_t, t in $\mathcal{V}$, be some centered real Gaussian processes. Assume that*

$$\forall s,t \in \mathcal{V} \quad \mathrm{Var}(Z_s - Z_t) \leq \mathrm{Var}(X_s - X_t),$$

then

$$\mathbb{E}\sup_{t\in\mathcal{V}} Z_t \leq 2\mathbb{E}\sup_{t\in\mathcal{V}} X_t.$$

A proof of Theorem 3.3 can be found in Theorem 3.14 of [Mas07]. For a fixed sample $X_1,\ldots,X_n$, define the Gaussian process $Z_\mathbf{c}$ by

$$Z_\mathbf{c} = \sum_{i=1}^n g_i(\gamma(\mathbf{c},X_i)-\gamma(\mathbf{c}^*,X_i)),$$

over the set $\mathcal{V}(\delta) = \mathcal{B}(\mathbf{c}^*, \frac{\sqrt{\delta}}{4M})$, where $\mathbf{c}^*$ is a fixed optimal codebook. For $i = 1,\ldots,n$,

$\mathbf{c}, \mathbf{c}' \in \mathcal{V}(\delta)$, we have

$$\begin{aligned}\big(\gamma(\mathbf{c},X_i)-\gamma(\mathbf{c}',X_i)\big)^2 &\le \sup_{j=1,\dots,k}\Big(\|X_i-c_j\|^2-\|X_i-c'_j\|^2\Big)^2\\ &\le \sup_{j=1,\dots,k}\Big(-2\Big\langle c_j-c'_j,X_i\Big\rangle+\|c_j\|^2-\|c'_j\|^2\Big)^2\\ &\le \sup_{j=1,\dots,k}\Big(8\Big\langle c_j-c'_j,X_i\Big\rangle^2+2(\|c_j\|^2-\|c'_j\|^2)^2\Big).\end{aligned}$$

Define now the Gaussian process $X_{\mathbf{c}}$ by

$$X_{\mathbf{c}}=2\sqrt{2}\sum_{i=1}^{n}\sum_{j=1}^{k}\Big\langle c_j-c_j^*,X_i\Big\rangle\xi_{i,j}+\sqrt{2n}\sum_{j=1}^{k}(\|c_j\|^2-\|c_j^*\|^2)\xi'_j,$$

where the ξ 's and ξ' 's are independent standard Gaussian variables. It is straightforward that $\mathrm{Var}(Z_{\mathbf{c}}-Z_{\mathbf{c}'})\le \mathrm{Var}(X_{\mathbf{c}}-X_{\mathbf{c}'})$. Therefore, applying Theorem 3.3 leads to

$$\begin{aligned}\mathbb{E}_g\sup_{\mathbf{c}\in\mathcal{V}(\delta)}Z_{\mathbf{c}} &\le 2\mathbb{E}_\xi\sup_{\mathbf{c}\in\mathcal{V}(\delta)}X_{\mathbf{c}}\\ &\le 4\sqrt{2}\mathbb{E}_\xi\sup_{\mathbf{c}\in\mathcal{V}(\delta)}\sum_{i=1}^{n}\sum_{j=1}^{k}\Big\langle c_j-c_j^*,X_i\Big\rangle\xi_{i,j}\\ &\quad+2\sqrt{2n}\mathbb{E}_{\xi'}\sup_{\mathbf{c}\in\mathcal{V}(\delta)}\sum_{j=1}^{k}(\|c_j\|^2-\|c_j^*\|^2)\xi'_j.\end{aligned}\tag{3.16}$$

Using almost the same technique as in the proof of Theorem 2.1 in [BDL08], the first term of the right-hand side of (3.16) can be bounded as follows :

$$\begin{aligned}\mathbb{E}_\xi\sup_{\mathbf{c}\in\mathcal{V}(\delta)}\sum_{i=1}^{n}\sum_{j=1}^{k}\Big\langle c_j-c_j^*,X_i\Big\rangle\xi_{i,j} &= \mathbb{E}_\xi\sup_{\mathbf{c}\in\mathcal{V}(\delta)}\sum_{j=1}^{k}\left\langle c_j-c_j^*,\left(\sum_{i=1}^{n}\xi_{i,j}X_i\right)\right\rangle\\ &\le \mathbb{E}_\xi\sup_{\mathbf{c}\in\mathcal{V}(\delta)}\|\mathbf{c}-\mathbf{c}^*\|\sqrt{\sum_{j=1}^{k}\left\|\sum_{i=1}^{n}\xi_{i,j}X_i\right\|^2}\\ &\le \frac{\sqrt{\delta}}{4M}\sqrt{\sum_{j=1}^{k}\mathbb{E}_\xi\left\|\sum_{i=1}^{n}\xi_{i,j}X_i\right\|^2}\\ &\le \frac{\sqrt{k\delta}}{4M}\sqrt{\sum_{i=1}^{n}\|X_i\|^2}.\end{aligned}$$

Then, applying Jensen's inequality ensures that

$$\mathbb{E}_X\sqrt{\sum_{i=1}^{n}\|X_i\|^2}\le\sqrt{n}M.$$

Similarly, the second term of the right-hand side of (3.16) can be bounded from above by

$$\begin{aligned}\mathbb{E}_{\xi'}\sup_{\mathbf{c}\in\mathcal{V}(\delta)}\sum_{j=1}^{k}(\|c_j\|^2-\|c_j^*\|^2)\xi'_j &\le \mathbb{E}_{\xi'}\sup_{\mathbf{c}\in\mathcal{V}(\delta)}\sqrt{\sum_{j=1}^{k}\Big(\|c_j\|^2-\|c_j^*\|^2\Big)^2}\sqrt{\sum_{j=1}^{k}{\xi'_j}^2}\\ &\le \frac{\sqrt{k\delta}}{2}.\end{aligned}$$

Combining these two bounds ensures that, for a fixed $\mathbf{c}^*$,

$$\mathbb{E}_{X,g} \sup_{\|\mathbf{c}-\mathbf{c}^*\|^2 \leq \delta/16M^2} Z_{\mathbf{c}} \leq 2\sqrt{2kn}\sqrt{\delta},$$

which leads to the desired result.

3.4.3.4 Proof of (3.8)

The proof of (3.8) also relies on an application of Theorem 3.2. Let the loss of $\hat{\mathbf{c}}_n$ be decomposed as follows,

$$\begin{aligned} P(\gamma(\hat{\mathbf{c}}_n) - \gamma(\mathbf{c}^*)) &\leq (P - P_n)(\gamma(\hat{\mathbf{c}}_n) - \gamma(\mathbf{c}^*)) \\ &\leq (P - P_n)\left\langle \hat{\mathbf{c}}_n - \mathbf{c}^*(\hat{\mathbf{c}}_n), \Delta(\mathbf{c}^*(\hat{\mathbf{c}}_n), .)\right\rangle \\ &\quad + (P - P_n)\|\hat{\mathbf{c}}_n - \mathbf{c}^*(\hat{c}_n)\| R(\hat{\mathbf{c}}_n, \mathbf{c}^*(\hat{c}_n), .), \end{aligned} \tag{3.17}$$

where

$$\Delta(\mathbf{c}^*, x) = -2((x - c_1^*)\mathbb{1}_{V_1(\mathbf{c}^*)}, \ldots, (x - c_k^*)\mathbb{1}_{V_k(\mathbf{c}^*)}),$$

and

$$R(\mathbf{c}, \mathbf{c}^*, x) = \sum_{i,j=1,\ldots,k} \mathbb{1}_{V_i(\mathbf{c}^*) \cap W_j(\mathbf{c})} \|\mathbf{c} - \mathbf{c}^*\|^{-1} \left[\|c_i - c_i^*\|^2 + 2\left\langle x - \frac{c_i + c_j}{2}, c_i - c_j \right\rangle \right],$$

where we recall that $W_i(\mathbf{c})$ denotes an element of a Voronoi partition, such that $\bar{W}_i(\mathbf{c}) \subset V_i(\mathbf{c})$. The proof of (3.6) consists in applying Theorem 3.2 to the two terms in the right-hand side of (3.17).

The first term on the right-hand side of (3.17) may be thought of as the dominant term in the decomposition of the loss. Define

$$\mathscr{F}_2 = \left\{ \left\langle \mathbf{c} - \mathbf{c}^*(\mathbf{c}), \Delta(\mathbf{c}^*(\mathbf{c}), .)\right\rangle) | \quad \mathbf{c} \in \mathscr{B}(0, M) \right\}.$$

In order to apply Theorem 3.2, the following lemmas are needed.

Lemma 3.3. *Let $f \in \mathscr{F}_2$, then*

$$\begin{cases} \|f\|_\infty & \leq & 8M, \\ \mathrm{Var}_P(f) & \leq & 4\|\mathbf{c} - \mathbf{c}^*(\mathbf{c})\|^2 R(\mathbf{c}^*). \end{cases}$$

Proof of Lemma 3.3. Elementary calculation shows that

$$\begin{aligned} \mathrm{Var}(\left\langle \mathbf{c} - \mathbf{c}^*, \Delta(\mathbf{c}^*, .)\right\rangle) =& P(\left\langle \mathbf{c} - \mathbf{c}^*, \Delta(\mathbf{c}^*, .)\right\rangle)^2 - (P(\left\langle \mathbf{c} - \mathbf{c}^*, \Delta(\mathbf{c}^*, .)\right\rangle))^2 \\ =& \sum_{i=1}^k P\left[\left\langle c_i - c_i^*, -2(x - c_i^*)\right\rangle^2 \mathbb{1}_{V_i(\mathbf{c}^*)}(x) \right] \\ \leq& 4\|\mathbf{c} - \mathbf{c}^*\|^2 R(\mathbf{c}^*). \end{aligned}$$

□

Let $\omega_2(f)$ be defined as $4\|\mathbf{c} - \mathbf{c}^*(\mathbf{c})\|^2 R(\mathbf{c}^*)$. It remains to bound from above the expectation of the maximum deviation between P and P_n over the set $\mathscr{F}_2$.

Lemma 3.4. *One has*

$$\mathbb{E} \sup_{f\in\mathscr{F}_2,\omega_2(f)\le\delta} |(P-P_n)f| \le \frac{2|\mathscr{M}|}{\sqrt{n}}\sqrt{\delta}. \tag{3.18}$$

Proof of Lemma 3.4. This proof is inspired from the proof of Lemma 4.3 in [BDL08]. The first step is the following

$$\mathbb{E} \sup_{f\in\mathscr{F}_2,\omega_2(f)\le\delta} |(P-P_n)f| \le \mathbb{E} \sup_{\mathbf{c}^*\in\mathscr{M},\|\mathbf{c}-\mathbf{c}^*\|\le\sqrt{\delta/4R(\mathbf{c}^*)}} \left|(P-P_n)\left\langle \mathbf{c}-\mathbf{c}^*,\Delta(\mathbf{c}^*,.)\right\rangle\right|.$$

For a general function $h(Z)$ depending on a random map Z, we denote by $\mathbb{E}_Z h$ the expectation of h taken with respect to Z. Introducing some Rademacher independent random variables σ_i and using a symmetrization inequality such as in Section 2.2 of [Kol06] leads to

$$\begin{aligned}
&\mathbb{E} \sup_{\|\mathbf{c}-\mathbf{c}^*\|\le\sqrt{\delta/4R(\mathbf{c}^*)},\mathbf{c}^*\in\mathscr{M}} \left|(P-P_n)\left\langle \mathbf{c}-\mathbf{c}^*,\Delta(\mathbf{c}^*,.)\right\rangle\right| \\
&\qquad \le 2\mathbb{E}_X\mathbb{E}_\sigma \sup_{\|\mathbf{c}-\mathbf{c}^*\|\le\sqrt{\delta/4R(\mathbf{c}^*)},\mathbf{c}^*\in\mathscr{M}} \left\langle \mathbf{c}-\mathbf{c}^*, \frac{1}{n}\sum_{i=1}^n \sigma_i\Delta(\mathbf{c}^*,X_i)\right\rangle \\
&\qquad \le \sqrt{\delta/4R(\mathbf{c}^*)}\,2\mathbb{E}_X\mathbb{E}_\sigma \sup_{\mathbf{c}^*\in\mathscr{M}} \left\|\frac{1}{n}\sum_{i=1}^n \sigma_i\Delta(\mathbf{c}^*,X_i)\right\|,
\end{aligned}$$

using Cauchy-Schwarz inequality. Eventually,

$$\begin{aligned}
\mathbb{E}_X\mathbb{E}_\sigma \sup_{\mathbf{c}^*\in\mathscr{M}} \left\|\frac{1}{n}\sum_{i=1}^n \sigma_i\Delta(\mathbf{c}^*,X_i)\right\| &\le \sum_{\mathbf{c}^*\in\mathscr{M}} \mathbb{E}_X\mathbb{E}_\sigma \left\|\frac{1}{n}\sum_{i=1}^n \sigma_i\Delta(\mathbf{c}^*,X_i)\right\| \\
&\le \sum_{\mathbf{c}^*\in\mathscr{M}} \sqrt{\mathbb{E}_X\mathbb{E}_\sigma \left\|\frac{1}{n}\sum_{i=1}^n \sigma_i\Delta(\mathbf{c}^*,X_i)\right\|^2} \\
&\le \sum_{\mathbf{c}^*\in\mathscr{M}} \frac{1}{\sqrt{n}}\sqrt{\mathbb{E}_X\|\Delta(\mathbf{c}^*,X)\|^2} \\
&\le \frac{2|\mathscr{M}|\sqrt{R(\mathbf{c}^*)}}{\sqrt{n}},
\end{aligned}$$

where Jensen's inequality has been used to obtain the second line. This gives the desired result. □

The contribution of the first term in the right-hand side of (3.17) is described by the following Proposition.

Proposition 3.7. *Let K_2 be a positive constant and $x>0$. Then, with probability larger than $1-e^{-x}$,*

$$(P-P_n)\left\langle \hat{\mathbf{c}}_n-\mathbf{c}^*(\hat{\mathbf{c}}_n),\Delta(\mathbf{c}^*(\hat{\mathbf{c}}_n),.)\right\rangle \le K_2^{-1}\left[4R(\mathbf{c}^*)\|\hat{\mathbf{c}}_n-\mathbf{c}^*(\hat{\mathbf{c}}_n)\|^2 + \frac{48^2|\mathscr{M}|^2K_2^2}{n} + \frac{9K_2^2+128M^2\sqrt{k}K_2}{4n}x\right].$$

The proof follows from a direct application of Theorem 3.2 to the set $\mathscr{F}_2$, replacing the value $C(1/2)$ with 4 to ease the calculation.

The second term in the right-hand side of (3.17) may be thought of as a residual term. Deriving sharper bounds on this term requires more accurate chaining techniques, as exposed below. Define

$$\mathscr{F}_3 = \left\{ \|\mathbf{c} - \mathbf{c}^*(\mathbf{c})\| R(\mathbf{c}, \mathbf{c}^*(\mathbf{c}), .) | \mathbf{c} \in \mathscr{B}(0, M) \right\}.$$

In order to apply Theorem 3.2, the following intermediate results are needed.

Lemma 3.5. *Let $f \in \mathscr{F}_3$, then*

$$\begin{cases} \|f\|_\infty & \leq & 2M\sqrt{k}C_\infty, \\ \mathrm{Var}_P(f) & \leq & C_\infty^2 \|\mathbf{c} - \mathbf{c}^*(\mathbf{c})\|^2, \end{cases}$$

with

$$C_\infty = (2\sqrt{k} + 8\sqrt{2})M.$$

Proof of Lemma 3.5. The proof of Lemma 3.5 follows from a bound on $R(\mathbf{c}, \mathbf{c}^*(\mathbf{c}), x)$, namely

$$\begin{aligned} |R(\mathbf{c}, \mathbf{c}^*, x)| &= \|\mathbf{c} - \mathbf{c}^*\|^{-1} \left| \sum_{i,j} \mathbb{1}_{V_i(\mathbf{c}^*) \cap W_j(\mathbf{c})} \left(\|c_i - c_i^*\|^2 \right.\right. \\ &\qquad \left.\left. + 2 \left\langle x - \frac{c_i + c_j}{2}, c_i - c_j \right\rangle \right) \right| \\ &\leq \|\mathbf{c} - \mathbf{c}^*\|^{-1} \left[\sum_i \|c_i - c_i^*\|^2 \mathbb{1}_{V_i(\mathbf{c}^*)} \right. \\ &\qquad \left. + \sum_{i \neq j} 2 \left| \left\langle x - \frac{c_i + c_j}{2}, c_i - c_j \right\rangle \right| \mathbb{1}_{V_i(\mathbf{c}^*) \cap W_j(\mathbf{c})} \right]. \end{aligned}$$

Since, for all j in $\{1, \ldots, k\}$, $W_j(\mathbf{c}) \subset V_j(\mathbf{c})$, applying Lemma 3.2 leads to

$$\begin{aligned} |R(\mathbf{c}, \mathbf{c}^*, x)| &\leq \|\mathbf{c} - \mathbf{c}^*\|^{-1} \left[\|\mathbf{c} - \mathbf{c}^*\|^2 + 8\sqrt{2}M \|\mathbf{c} - \mathbf{c}^*\| \mathbb{1}_{N^*(\frac{4\sqrt{2}M}{B}\|\mathbf{c} - \mathbf{c}^*\|)} \right] \\ &\leq \|\mathbf{c} - \mathbf{c}^*\| + 8\sqrt{2}M \mathbb{1}_{N^*(\frac{4\sqrt{2}M}{B}\|\mathbf{c} - \mathbf{c}^*\|)} := F_{\|\mathbf{c} - \mathbf{c}^*\|}(x). \end{aligned} \tag{3.19}$$

Elementary calculations show that, for any $\delta > 0$,

$$\|F_\delta\|_\infty \leq (2\sqrt{k} + 8\sqrt{2})M = C_\infty,$$

from which we deduce the desired upper bounds on $\mathrm{Var}_P(f)$ and $\|f\|_\infty$, for f in $\mathscr{F}_3$. □

Let $\omega_3(f)$ be defined as $C_\infty^2 \|\mathbf{c} - \mathbf{c}^*(\mathbf{c})\|^2$. The complexity term associated with the class of functions $\mathscr{F}_3$ can be bounded as follows.

Proposition 3.8.

$$\mathbb{E} \sup_{f \in \mathscr{F}_3, \omega_3(f) \leq \delta} |(P - P_n)f| \leq \frac{8Q(k,d)}{\sqrt{C_\infty n}} \sqrt{C_2(\sqrt{\delta}/C_\infty)\sqrt{\delta}},$$

where

$$\begin{cases} C_2(r) & = & r + 8\sqrt{2}Mp\left(\frac{4\sqrt{2}M}{B}r\right), \\ Q(k,d) & = & 8\sqrt{K_0 P(k,d) \log(k^2(4k-2))}, \\ P(k,d) & = & k^2(2(k-1)(d+1) + 24(3d+4)), \end{cases}$$

and K_0 is an absolute constant.

The proof of Proposition 3.8 is based on Theorem 1 in [MV03] and its application to a more accurate version of Dudley's integral. For clarity, the proof is postponed to Section 3.5. Since P satisfies a margin condition with parameters (r_0, κ), with $\kappa \leq \frac{Bp_{min}}{128M^2}$, considering the two cases $4\sqrt{2}Mr/B \leq r_0$ and $4\sqrt{2}Mr/B \geq r_0$ yields

$$8\sqrt{2}Mp(\frac{4\sqrt{2}M}{B}r) \leq \frac{64M^2}{Br_0}r,$$

for $r \geq 0$. Using this inequality to bound C_2 from above in Proposition 3.8 leads to the following complexity result

$$\mathbb{E} \sup_{f \in \mathscr{F}_3, \omega_3(f) \leq \delta} |(P - P_n)f| \leq \frac{\Xi_3}{\sqrt{n}}\delta^{\frac{3}{4}}, \tag{3.20}$$

where

$$\Xi_3 = \frac{8MQ(k,d)}{C_\infty \sqrt{Br_0}}.$$

Let Φ_3 be defined as $\frac{\Xi_3}{\sqrt{n}}\delta^{\frac{3}{4}}$. Remark that Φ_3 is a sub-3/4 function. Consequently, for any $D > 0$, the solution of the equation $\Phi_3(\delta) = \delta/D$ is

$$\delta_3^* = \frac{(D\Xi_3)^4}{n^2}.$$

Choosing $K_3 > 0$, and $D = 6K_3C(3/4)$ in Theorem 3.2 and taking into account that $C(3/4) \leq 10$ leads to the following Proposition.

Proposition 3.9. *Let $K_3 > 0$. Then, with probability larger than $1 - e^{-x}$,*

$$(P - P_n)\left(\|\hat{\mathbf{c}}_n - \mathbf{c}^*(\hat{\mathbf{c}}_n)\| R(\hat{\mathbf{c}}_n, \mathbf{c}^*(\hat{\mathbf{c}}_n), .)\right) \leq K_3^{-1}\left[C_\infty^2 \|\hat{\mathbf{c}}_n - \mathbf{c}^*(\hat{\mathbf{c}}_n)\|^2 + \frac{60^4 K_3^4 \Xi_3^4}{n^2} + \frac{9K_3^2 + 32M\sqrt{k}C_\infty K_3}{4n}x\right],$$

with $\Xi_3 = \frac{8MQ(k,d)}{C_\infty\sqrt{Br_0}}$, and Q is a function composed of products of square roots of polynomial functions in k, d, and $\log(k)$.

We are now in position to prove (3.8). Proposition 3.2 provides κ_0 such that

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

Choosing $K_2 = 8R(\mathbf{c}^*)\kappa_0$ in Proposition 3.7, $K_3 = 2C_\infty^2\kappa_0$ in Proposition 3.9 and summing the two resulting inequalities leads to (3.8), valid on a set which has probability larger than $1 - 2e^{-x}$.

3.4.4 Proof of Proposition 3.3

Throughout this Subsection, $\mathscr{H} = \mathbb{R}^d$, and, for a codebook $\mathbf{c}$, let Q denote the associated nearest-neighbor quantizer. In the general case, such an association depends on how the boundaries are allocated. However, since the distributions involved in the minimax result have densities, how boundaries are allocated will not matter.

Let $k \geq 3$ be an integer. For convenience k is assumed to be divisible by 3. Let $m = 2k/3$. Let $z_1, \ldots, z_m$ denote a 6Δ-net in $\mathscr{B}(0, M-\rho)$, where $\Delta > 0$, and $w_1, \ldots, w_m$ a sequence of vectors such that $\|w_i\| = \Delta$. Finally, denote by U_i the ball $\mathscr{B}(z_i, \rho)$ and by U_i' the ball $\mathscr{B}(z_i + w_i, \rho)$. Slightly anticipating, define $\rho = \frac{\Delta}{16}$.

To get the largest Δ such that for all $i = 1, \ldots, k$, U_i and U_i' are included in $\mathscr{B}(0, M)$, it suffices to get the largest Δ such that there exists a 6Δ-net in $\mathscr{B}(0, M - \Delta/16)$. Since the cardinal of a 6Δ-net is larger than the largest number of balls of radius 6Δ which can be packed into $\mathscr{B}(0, M - \Delta/16)$, a sufficient condition on Δ to guarantee that a 6Δ-net can be found is given by

$$m \leq \left(\frac{M - \Delta/16}{6\Delta} \right)^d.$$

Since $\Delta \leq M$, Δ can be chosen as

$$\Delta = \frac{5M}{32 m^{1/d}}.$$

For such a Δ, ρ takes the value $\rho = \frac{\Delta}{16} = \frac{5M}{512 m^{1/d}}$. Therefore, ρ only depends on k, d, and M.

Let $z = (z_i)_{i=1,\ldots,m}$ and $w = (w_i)_{i=1,\ldots,m}$ be sequences as described above, such that, for $i = 1, \ldots, k$, U_i and U_i' are included in $\mathscr{B}(0, M)$. For a fixed $\sigma \in \{-1, +1\}^m$ such that $\sum_{i=1}^m \sigma_i = 0$, let P_σ be defined as

$$\begin{cases} P_\sigma(U_i) & = & \frac{1+\sigma_i \delta}{2m}, \\ P_\sigma(U_i') & = & \frac{1+\sigma_i \delta}{2m}, \\ P_\sigma & \underset{U_i}{\sim} & (\rho - \|x - z_i\|) \mathbb{1}_{\|x - z_i\| \leq \rho} d\lambda(x), \\ P_\sigma & \underset{U_i'}{\sim} & (\rho - \|x - z_i - w_i\|) \mathbb{1}_{\|x - z_i - w_i\| \leq \rho} d\lambda(x), \end{cases}$$

where λ denotes the Lebesgue measure. These cone-shaped distributions has been designed to have a continuous density, as in Theorem 4 in [Ant05]. To be more precise, for τ in $\{-1, +1\}^{\frac{m}{2}}$, $\sigma(\tau)$ is defined as the sequence in $\{-1, +1\}^m$ such that

$$\begin{cases} \sigma_i(\tau) & = & \tau_i, \\ \sigma_{i + \frac{m}{2}}(\tau) & = & -\sigma_i(\tau), \end{cases}$$

for $i = 1, \ldots, \frac{m}{2}$. Finally, for a quantizer Q let $R(Q, P_\sigma)$ denote the distortion of Q in the case where the source distribution is P_σ.

Similarly, for σ in $\{-1, +1\}^m$ satisfying $\sum_{i=1}^m \sigma_i = 0$, let Q_σ denote the quantizer defined by $Q_\sigma(U_i) = Q_\sigma(U_i') = z_i + \omega_i/2$ if $\sigma_i = -1$, $Q_\sigma(U_i) = z_i$ and $Q_\sigma(U_i') = z_i + \omega_i$ if $\sigma_i = +1$. Let $\mathscr{Q}$ denote the set of such quantizers. It can be proved that only quantizers in $\mathscr{Q}$ have to be considered.

Proposition 3.10. *Assume that $\delta \leq 1/3$, $\Delta > 0$, and $\rho \leq \frac{\Delta}{16}$. Then, for every quantizer Q there exists a quantizer Q_σ in $\mathscr{Q}$ such that*

$$\forall P_{\sigma'} \quad R(Q_\sigma, P_{\sigma'}) \leq R(Q, P_{\sigma'}).$$

The proof of Proposition 3.10 follows the proof of Step 3 of Theorem 1 in [BLL98], replacing distributions supported on a finite set with distributions supported on small balls. Provided that the radius of these balls are small enough, the results

are nearly the same in the two cases. The proof of Proposition 3.10 can be found in Section 3.5.4.

Since, for $\sigma \neq \sigma'$, $R(Q'_\sigma, P_\sigma) > R(Q_\sigma, P_\sigma)$, Proposition 3.10 ensures that the P_σ's have a unique optimal codebook, up to relabeling.

For any σ and σ' in $\{-1,+1\}^m$, denote by $\rho(\sigma,\sigma')$ the quantity $\sum_{i=1}^m |\sigma_i - \sigma'_i|$, and by $H(P_\sigma, P_{\sigma'})$ the Hellinger distance between P_σ and $P_{\sigma'}$. To apply Assouad's Lemma to the set $\{P_{\sigma(\tau)}\}_{\tau\in\{-1,+1\}^{\frac{m}{2}}}$, the following lemma is needed :

Lemma 3.6. *Let τ and τ' denote two sequences in $\{-1,+1\}^{\frac{m}{2}}$ such that $\rho(\tau,\tau') = 2$, then*

$$H(P^{\otimes n}_{\sigma(\tau)}, P^{\otimes n}_{\sigma(\tau')}) \leq \frac{4n\delta^2}{m},$$

where $P^{\otimes n}$ denotes the product law of a n-sample drawn from P.

Furthermore, for any σ and σ' in $\{-1,+1\}^m$,

$$R(Q_{\sigma'}, P_\sigma) = R(Q_\sigma, P_\sigma) + \frac{\Delta^2\delta}{8m}\rho(\sigma,\sigma').$$

A proof of Lemma 3.6 is given in Section 3.5.5. Equipped with Lemma 3.6, a direct application of Assouad's Lemma as in Theorem 2.12 of [Tsy09] yields, provided that $\delta = \frac{\sqrt{m}}{2\sqrt{n}}$,

$$\sup_{\tau\in\{-1,+1\}^{\frac{m}{2}}} \mathbb{E}\left(R(\hat{Q}_n, P_{\sigma(\tau)}) - R(Q_{\sigma(\tau)}, P_{\sigma(\tau)})\right) \geq c_0 M^2 \sqrt{\frac{k^{1-\frac{4}{d}}}{n}},$$

for any empirically designed quantizer $\hat{Q}_n$, where c_0 is an explicit constant.

Finally, it may be noticed that, for every $\delta \leq \frac{1}{3}$ and σ, P_σ satisfies a margin condition as in (3.9), and is ε-separated, with

$$\varepsilon = \frac{\Delta^2\delta}{2m}.$$

This concludes the proof of Proposition 3.3.

3.4.5 Proof of Proposition 3.4

As mentioned below Proposition 3.4, the inequality

$$\frac{\theta_{min}}{\theta_{max}} \geq \frac{2048k\sigma^2}{(1-\varepsilon)\tilde{B}^2(1-e^{-\tilde{B}^2/2048\sigma^2})},$$

ensures that, for every j in $\{1,\dots,k\}$, there exists i in $\{1,\dots,k\}$ such that $\left\|c_i^* - m_j\right\| \leq \tilde{B}/16$. To be more precise, let $\mathbf{m}$ denote the vector of means $(m_1,\dots,m_k)$, then

$$\begin{aligned}
R(\mathbf{m}) &\leq \sum_{i=1}^k \frac{\theta_i}{2\pi\sigma^2 N_i} \int_{V_i(\mathbf{m})} \|x - m_i\|^2 e^{-\frac{\|x-m_j\|^2}{2\sigma^2}} dx \\
&\leq \frac{p_{max}}{2(1-\varepsilon)\pi\sigma^2} \sum_{i=1}^k \int_{\mathbb{R}^2} \|x - m_i\|^2 e^{-\frac{\|x-m_j\|^2}{2\sigma^2}} dx \\
&\leq \frac{2kp_{max}\sigma^2}{1-\varepsilon}.
\end{aligned}$$

Assume that there exists i in $\{1,\dots,k\}$ such that, for all j, $\|c_j^* - m_i\| \geq \tilde{B}/16$. Then

$$\begin{aligned} R(\mathbf{c}) &\geq \frac{\theta_i}{2\pi\sigma^2}\int_{\mathscr{B}(m_i,\tilde{B}/32)} \frac{\tilde{B}^2}{1024} e^{-\frac{\|x-m_i\|^2}{2\sigma^2}} \\ &\geq \frac{\tilde{B}^2\theta_{min}}{2048\pi\sigma^2}\int_{\mathscr{B}(m_i,\tilde{B}/32)} e^{-\frac{\|x-m_i\|^2}{2\sigma^2}} \\ &> \frac{\tilde{B}^2\theta_{min}}{1024}\left(1-e^{-\frac{\tilde{B}^2}{2048\sigma^2}}\right) \\ &> R(\mathbf{m}). \end{aligned}$$

Hence the contradiction. Up to relabeling, it is now assumed that for $i = 1,\dots,k$, $\|m_i - c_i^*\| \leq \tilde{B}/16$. Take y in $N^*(x)$, for $x \leq \frac{\tilde{B}}{8}$, then, for every i in $\{1,\dots,k\}$,

$$\|y - m_i\| \geq \frac{\tilde{B}}{4},$$

which leads to

$$\sum_{i=1}^{k} \frac{\theta_i}{2\pi\sigma^2 N_i}\|y-m_i\|^2 e^{-\frac{\|y-m_i\|^2}{2\sigma^2}} \leq \frac{k\theta_{max}}{(1-\varepsilon)2\pi\sigma^2} e^{-\frac{\tilde{B}^2}{32\sigma^2}}.$$

Since the Lebesgue measure of $N^*(x)$ is smaller than $4k\pi Mx$, it follows that

$$P(N^*(x)) \leq \frac{2k^2 M\theta_{max}}{(1-\varepsilon)\sigma^2} e^{-\frac{\tilde{B}^2}{32\sigma^2}} x.$$

On the other hand, $\|m_i - c_i^*\| \leq \tilde{B}/16$ yields

$$\mathscr{B}(m_i, 3\tilde{B}/8) \subset V_i(\mathbf{c}^*).$$

Therefore,

$$\begin{aligned} P(V_i(\mathbf{c}^*)) &\geq \frac{\theta_i}{2\pi\sigma^2 N_i}\int_{\mathscr{B}(m_i,3\tilde{B}/8)} e^{-\frac{\|x-m_i\|^2}{2\sigma^2}} dx \\ &\geq \theta_i\left(1-e^{-\frac{9\tilde{B}^2}{128\sigma^2}}\right), \end{aligned}$$

hence $p_{min} \geq \theta_{min}\left(1-e^{-\frac{9\tilde{B}^2}{128\sigma^2}}\right)$. Consequently, provided that

$$\frac{\theta_{min}}{\theta_{max}} \geq \frac{2048k^2M^3}{(1-\varepsilon)7\sigma^2\tilde{B}(e^{\tilde{B}^2/32\sigma^2}-1)},$$

direct calculation shows that

$$P(N^*(x)) \leq \frac{Bp_{min}}{128M^2}x.$$

This ensures that P satisfies (3.3). According to Proposition 3.2, since $\mathscr{H} = \mathbb{R}^2$, $\mathscr{M}$ is finite.

3.5 Technical results

3.5.1 Proof of Proposition 3.5

The proof of Proposition 3.5 is derived from the proof of Lemma 3 in [CL06]. Let $\mathbf{c}^*$ be an optimal codebook, and $\mathbf{c}$ be a codebook. We denote by $f_{\mathbf{c}^*,\mathbf{c}}$ the function $\gamma(\mathbf{c},.)-\gamma(\mathbf{c}^*,.)$, so that

$$\mathscr{F}_1 = \left\{ f_{\mathbf{c}^*(\mathbf{c}),\mathbf{c}} | \quad \mathbf{c} \in \mathscr{B}(0,M)^k \right\}.$$

Let $\Psi_1(r)$ denote the function

$$\Psi_1(r) = \mathbb{E} \sup_{\mathbf{c}^* \in \mathscr{M}, \|\mathbf{c}-\mathbf{c}^*(\mathbf{c})\| \leq r} \left| (P-P_n) f_{\mathbf{c}^*(\mathbf{c}),\mathbf{c}} \right|.$$

Since

$$\left\{ (\mathbf{c}^*(\mathbf{c}), \mathbf{c}^*) | \quad \|\mathbf{c}-\mathbf{c}^*(\mathbf{c})\| \leq r \right\} \subset \left\{ (\mathbf{c}^*, \mathbf{c}) | \quad \mathbf{c}^* \in \mathscr{M}, \|\mathbf{c}-\mathbf{c}^*\| \leq r \right\},$$

it is easy to see that

$$\Psi_1(r) \leq \mathbb{E} \sup_{\mathbf{c}^* \in \mathscr{M}, \|\mathbf{c}-\mathbf{c}^*\| \leq r} \left| (P-P_n) f_{\mathbf{c}^*,\mathbf{c}} \right|.$$

The rest of the proof is derived from a chaining technique, used in the proof of Proposition 2.4 in Chapter 2 or Lemma 3 in [CL06]. Set $\varepsilon_j = 2^{-j} r$, for $j \geq 0$, and for every $\mathbf{c}^*$ in $\mathscr{M}$, denote by $N_j(\mathbf{c}^*)$ an ε_j net of $\mathscr{B}(\mathbf{c}^*, r)$, such that for every $\mathbf{c}$ in $\mathscr{B}(\mathbf{c}^*, r)$ there exists $\mathbf{c}_j$ in $N_j(\mathbf{c}^*)$ such that $\|\mathbf{c}_j - \mathbf{c}\| \leq \varepsilon_j$. According to the proof of Theorem 2 in [AGG05] or Lemma 3 in [CL06], such an $N_j(\mathbf{c}^*)$ can be defined, with

$$|N_j(\mathbf{c}^*)| \leq \left(\frac{2r\sqrt{kd}}{\varepsilon_j} \right) := n(\varepsilon_j).$$

By a dominated convergence Theorem, for any fixed $\mathbf{c}^*$ in $\mathscr{M}$ and $\mathbf{c}$,

$$f_{\mathbf{c}^*,\mathbf{c}_j} \overset{L_1(P),a.s.}{\underset{j\to\infty}{\longrightarrow}} f_{\mathbf{c}^*,\mathbf{c}}.$$

This allows us to decompose the expression of Ψ_1 as follows.

$$\begin{aligned}
\Psi_1(r) &\leq \mathbb{E} \sup_{\mathbf{c}^* \in \mathscr{M}, \|\mathbf{c}-\mathbf{c}^*\| \leq r} \left| (P-P_n) f_{\mathbf{c}^*,\mathbf{c}} \right| \\
&\leq \mathbb{E} \sup_{\mathbf{c}^* \in \mathscr{M}, \mathbf{c}_0 \in N_0(\mathbf{c}^*)} \left| (P-P_n) f_{\mathbf{c}^*,\mathbf{c}_0} \right| \\
&\quad + \sum_{j>1} \mathbb{E} \sup_{\mathbf{c}^* \in \mathscr{M}, \mathbf{c}_j \in N_j(\mathbf{c}^*), \mathbf{c}_{j-1} \in N_{j-1}(\mathbf{c}^*)} \left| (P-P_n)(f_{\mathbf{c}^*,\mathbf{c}_j} - f_{\mathbf{c}^*,\mathbf{c}_{j-1}}) \right|, \\
&:= A_1 + A_2.
\end{aligned}$$

It remains to bound from above these two terms.

Bound on A_1

Introducing some Rademacher random variables σ_i, $i = 1,\ldots,n$ and using the symmetrization principle as in [Kol06] leads to

$$\begin{aligned}
&\mathbb{E} \sup_{\mathbf{c}^* \in \mathscr{M}, \mathbf{c}_0 \in N_0(\mathbf{c}^*)} \left| (P-P_n) f_{\mathbf{c}^*,\mathbf{c}_0} \right| \\
&\qquad \leq 2\mathbb{E}_X \mathbb{E}_\sigma \sup_{\lambda=\pm 1, \mathbf{c}^* \in \mathscr{M}, \mathbf{c}_0 \in N_0(\mathbf{c}^*)} \frac{1}{n} \sum_{i=1}^n \sigma_i \lambda f_{\mathbf{c}^*,\mathbf{c}_0}(X_i).
\end{aligned}$$

Let introduce here a maximal inequality derived from Lemma 2.3 in [Mas07].

Lemma 3.7. *Let $x_1,\dots,x_n$ denote a sequence of points in $\mathcal{X}$, and let $\sigma_1,\dots,\sigma_n$ denote a sequence of independent Rademacher random variables. Let $\mathcal{F}$ be a set of real valued functions over $\mathcal{X}$ such that $|\mathcal{F}|<\infty$, and*

$$\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^{n}f^2(x_i)\le v.$$

Then

$$\mathbb{E}_\sigma\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i f(x_i)\le\sqrt{2v\log(|\mathcal{F}|)}.$$

In our case, for all $\mathbf{c}^*$ in $\mathcal{M}$ and $\mathbf{c}_0$ in $N_0(\mathbf{c}^*)$,

$$\frac{1}{n}\sum_{i=1}^{n}f^2_{\mathbf{c}^*,\mathbf{c}_0}(X_i)\le\frac{16M^2r^2}{n},$$

and

$$\left|\{\lambda=\pm1,\mathbf{c}^*\in\mathcal{M},\mathbf{c}_0\in N_0(\mathbf{c}^*)\}\right|\le|\mathcal{M}|(4\sqrt{kd})^{kd}.$$

Therefore, a direct application of Lemma 3.7 yields

$$A_1\le\frac{8\sqrt{2}M}{\sqrt{n}}\sqrt{kd\log(4|\mathcal{M}|\sqrt{kd})}.$$

Bound on A_2

Let $j>1$. Using the same symmetrization argument as above leads to

$$\begin{aligned}A_{2,j}&:=\mathbb{E}\sup_{\mathbf{c}^*\in\mathcal{M},\mathbf{c}_j\in N_j(\mathbf{c}^*),\mathbf{c}_{j-1}\in N_{j-1}(\mathbf{c}^*)}\left|(P-P_n)(f_{\mathbf{c}^*,\mathbf{c}_j}-f_{\mathbf{c}^*,\mathbf{c}_{j-1}})\right|\\&\le\mathbb{E}_X\mathbb{E}_\sigma\sup_{\substack{\lambda=\pm1,\mathbf{c}^*\in\mathcal{M},\\ \mathbf{c}_j\in N_j(\mathbf{c}^*),\mathbf{c}_{j-1}\in N_{j-1}(\mathbf{c}^*)}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i\lambda(f_{\mathbf{c}^*,\mathbf{c}_j}(X_i)-f_{\mathbf{c}^*,\mathbf{c}_{j-1}}(X_i)).\end{aligned}$$

Since

$$\left\|f_{\mathbf{c}^*,\mathbf{c}_j}(.)-f_{\mathbf{c}^*,\mathbf{c}_{j-1}}(.)\right\|_\infty\le8Mr2^{-(j-1)},$$

and

$$\left|\{\lambda=\pm1,\mathbf{c}^*\in\mathcal{M},\mathbf{c}_j\in N_j(\mathbf{c}^*),\mathbf{c}_{j-1}\in N_{j-1}(\mathbf{c}^*)\}\right|\le2|\mathcal{M}|n(\varepsilon_j)^2,$$

a direct application of Lemma 3.7 leads to

$$A_{2,j}\le64Mr\sqrt{kd\log(|\mathcal{M}|\sqrt{kd}2^{j+2})}2^{-(j-1)}.$$

Comparing a sum with an integral, and observing that

$$\int_0^1\sqrt{\log(-x)}dx\le1$$

ensures that

$$A_2=\sum_{j>1}A_{2,j}\le\frac{256Mr}{\sqrt{n}}\left(\sqrt{\log(|\mathcal{M}|\sqrt{kd})}+1\right).$$

Combining the two bounds and remarking that

$$\mathbb{E}\sup_{f\in\mathcal{F}_1,\omega_1(f)\le\delta}|(P-P_n)f|\le\Psi_1\left(\frac{\sqrt{\delta}}{4M}\right)$$

gives the result of Proposition 3.5.

3.5.2 Proof of Proposition 3.8

The proof of Proposition 3.8 is based on a sharper chaining technique than the one used in Proposition 2.4 in the previous chapter. We intend to bound from above the complexity term

$$\mathbb{E} \sup_{\omega_3(f)\leq\delta, f\in\mathscr{F}_3} |(P-P_n)f|.$$

To this aim, define

$$\Psi_3(r) = \mathbb{E} \sup_{\|\mathbf{c}-\mathbf{c}^*(\mathbf{c})\|\leq r} \left|(P-P_n)R(\mathbf{c},\mathbf{c}^*,.)\right|,$$

where we recall that

$$R(\mathbf{c},\mathbf{c}^*,x) = \sum_{i,j=1,\dots,k} \mathbb{1}_{V_i(\mathbf{c}^*)\cap W_j(\mathbf{c})} \|\mathbf{c}-\mathbf{c}^*\|^{-1} \left[\|c_i - c_i^*\|^2 + 2\left\langle x - \frac{c_i+c_j}{2}, c_i - c_j \right\rangle \right],$$

where $(W_1(\mathbf{c}),\dots,W_k(\mathbf{c}))$ is a Voronoi partition, defined in Section 3.2. For technical reasons, this Voronoi partition must be specified. Denote by $\mathscr{C}(p)$ the set of subsets of $\mathbb{R}^d$ made of intersections of at most p half spaces (closed or open).

Since $R(\mathbf{c},\mathbf{c}^*,.)$ does not depend on how ties are broken, or, in other words, $R(\mathbf{c},\mathbf{c}^*,.)$ does not depend on the choice of $W_j(\mathbf{c})$'s among the partition cells satisfying

$$\overset{o}{V}_j(\mathbf{c}) \subset W_j(\mathbf{c}) \subset V_j(\mathbf{c}),$$

we choose a Voronoi partition such that every $W_j \in \mathscr{C}(k-1)$. For instance, if $H_{i,j}$ denotes the closed half-space $\{\|x-c_i\| \leq \|x-c_j\|\}$ and $\overset{o}{H}_{i,j}$ the open half space $\{\|x-c_i\| < \|x-c_j\|\}$. It is possible to build a Voronoi partition such that every cell is in $\mathscr{C}(k-1)$, choosing

$$W_j(\mathbf{c}) = \bigcap_{i<j} \overset{o}{H}_{i,j} \cap \bigcap_{i>j} H_{i,j}.$$

In short, this convention consists in allocating points on boundaries between V_j's to the smallest possible index. As a consequence, it is immediate that

$$\Psi_3(r) = \mathbb{E} \sup_{\|\mathbf{c}-\mathbf{c}^*(\mathbf{c})\|\leq r, W_j(\mathbf{c})\in\mathscr{C}(k-1)} \left|(P-P_n)R(\mathbf{c},\mathbf{c}^*,.)\right|.$$

The following set of function of interest is then introduced.

$$\mathscr{G}(r) = \{0\} \cup \frac{1}{F_r} \left\{ \lambda \|\mathbf{c}-\mathbf{c}^*\|^{-1} \sum_{i,j} \mathbb{1}_{V_i(\mathbf{c}^*)\cap W_j(\mathbf{c})\cap\mathscr{B}(0,M)} \left(\|c_i - c_i^*\|^2 + 2\left\langle x - \frac{c_i+c_j}{2}, c_i - c_j \right\rangle \right) \right| \quad \lambda = \pm 1, \mathbf{c}\in\mathscr{B}(0,M)^k, \mathbf{c}^*\in\mathscr{M}, W_j\in\mathscr{C}(k-1) \right\},$$

where F_r is defined in (3.19) as an envelope of $\mathscr{G}(r)$.

Let $\sigma_1,\dots,\sigma_n$ denote a sequence of independent Rademacher variables. As developed in the proof of Proposition 3.5, the first step is a symmetrization inequality

$$\begin{aligned}\Psi_3(r) &\le \mathbb{E}\sup_{\mathbf{c}^*\in\mathscr{M},\|\mathbf{c}-\mathbf{c}^*\|\le r}\left|(P-P_n)R(\mathbf{c},\mathbf{c}^*,.)\right| \\ &\le 2\mathbb{E}_X\mathbb{E}_\sigma\sup_{\lambda=\pm1,\mathbf{c}^*\in\mathscr{M},\|\mathbf{c}-\mathbf{c}^*\|\le r}\frac{1}{n}\sum_{i=1}^n\lambda\sigma_i R(\mathbf{c},\mathbf{c}^*,X_i) \\ &\le 2\mathbb{E}_X\mathbb{E}_\sigma\sup_{g\in\mathscr{G}(r)}\frac{1}{n}\sum_{i=1}^n\sigma_i F_r(X_i)g(X_i) \\ &:= 2\mathbb{E}_X\mathscr{R}_n.\end{aligned}$$

The next step is to chain the set $\mathscr{G}(r)$. To this aim, define for any set of real valued function $\mathscr{F}$, norm $\|\quad\|$ on $\mathscr{F}$ and $\varepsilon>0$, the covering number $\mathscr{N}(\mathscr{F},\|\quad\|,\varepsilon)$ as the cardinal of the smallest covering of $\mathscr{F}$ with balls of radius ε for the norm $\|\quad\|$.

To be more precise, for any g in $\mathscr{G}(r)$, and any finite subset $S\subset\mathbb{R}$, we define

$$\|g\|_{L_2(S)}=\sqrt{\frac{1}{|S|}\sum_{s\in S}g^2(s)},$$

and, with a slight abuse of notation, $\|g\|_{L_2(P_n)}=\sqrt{1/n\sum_{i=1}^n g^2(X_i)}$. The technical result concerning the covering numbers of $\mathscr{G}(r)$ is the following.

Proposition 3.11. *Let S be a finite set, and $0<\varepsilon<1$. There exists some constant $K>0$, not depending on S, such that*

$$\mathscr{N}(\mathscr{G}(r),\varepsilon,L_2(S)))\le\left(\frac{k^2(4k-2)}{\varepsilon}\right)^{KP(k,d)},$$

with $P(k,d)=k^2(2(k-1)(d+1)+24(3d+4))$.

For clarity, the proof of Proposition 3.11 is postponed to the following Subsection. An immediate consequence of Proposition 3.11 is that

$$\mathscr{N}(\mathscr{G}(r),\varepsilon,L_2(P_n))\le\left(\frac{k^2(4k-2)}{\varepsilon}\right)^{KP(k,d)}:=n(\varepsilon),$$

for any n-sample $X_1,\dots,X_n$. Consequently, let $X_1,\dots,X_n$ be fixed, and set $\varepsilon_0=1$, $\varepsilon_j=2^{-2j}\varepsilon_0$, for $j>1$.

For $j=0$, since F_r is an envelope of $\mathscr{G}(r)$, a 1 covering of $\mathscr{G}(r)$ for the $L_2(P_n)$ norm is the ball of center $g_0=0$ and radius 1.

For $j>1$, Proposition 3.11 provides a ε_j covering $\mathscr{G}_j(r)$ of $\mathscr{G}(r)$ for the $L_2(P_n)$ norm with cardinality at most $n(\varepsilon_j)$. For any g in $\mathscr{G}(r)$, denote by g_j the projection of g onto this covering, so that $\|g-g_j\|_{L_2(P_n)}\le\varepsilon_j$. For short we will write $n_j=n(\varepsilon_j)$.

It is easy to see that for every i in $\{1,\dots,n\}$,

$$g_j(X_i)\underset{j\to\infty}{\longrightarrow}g(X_i).$$

Then $\mathscr{R}_n$ may be decomposed as follows :

$$\begin{aligned}\mathscr{R}_n &= \mathbb{E}_\sigma\sup_{\lambda=\pm1,\mathbf{c}^*\in\mathscr{M},\|\mathbf{c}-\mathbf{c}^*\|\le r}\frac{1}{n}\sum_{i=1}^n\lambda\sigma_i R(\mathbf{c},\mathbf{c}^*,X_i) \\ &\le \sum_{j>1}\mathbb{E}_\sigma\sup_{g\in\mathscr{G}}\frac{1}{n}\sum_{i=1}^n\sigma_i F_r(X_i)(g_j(X_i)-g_{j-1}(X_i)) \\ &:= \sum_{j>1}b_j.\end{aligned}$$

A direct application of Lemma 3.7 for every b_j yields

$$\begin{aligned} b_j &\le \frac{1}{\sqrt{n}} \sqrt{2 \sup_{g \in \mathscr{G}(r)} \left\| (g_j - g_{j-1}) F_r \right\|_{L_2(P_n)}^2 \log(n_j n_{j-1})} \\ &\le \frac{1}{\sqrt{n}} \sqrt{2 \log(n_j n_{j-1}) \sup_{g \in \mathscr{G}(r)} 2C_\infty \|F_r\|_{L_2(P_n)} \|g_j - g_{j-1}\|_{L_2(P_n)}} \\ &\le \frac{4\sqrt{C_\infty}}{\sqrt{n}} \sqrt{\|F_r\|_{L_2(P_n)}\|} \sqrt{\log(n(\varepsilon_j))} \sqrt{\varepsilon_{j-1}}. \end{aligned}$$

Denote by ε'_j the quantity $\sqrt{\varepsilon_j} = 2^{-j}$. Since $x \mapsto \sqrt{\log(n(x^2))}$ is non-increasing, it is quite easy to see that

$$\frac{\sqrt{\log(n(\varepsilon'^2_j))}\varepsilon'_{j-1}}{4} = \sqrt{\log(n(\varepsilon'^2_j))}\varepsilon'_{j+1} \le \int_{\varepsilon'_{j+1}}^{\varepsilon'_j} \sqrt{\log(n(x^2))} dx.$$

From this we deduce that

$$\begin{aligned} \sum_{j>1} \sqrt{\varepsilon_{j-1} \log(n(\varepsilon_j))} &\le 4 \int_0^{1/2} \sqrt{\log(n(x^2))} dx \\ &\le \int_0^{1/2} \sqrt{KP(k,d) \log\left(\frac{k^2(4k-2)}{x^2}\right)} dx \\ &\le 2\sqrt{KP(k,d)\log(k^2(4k-2))} \\ &\quad + 4\sqrt{KP(k,d)} \int_0^{1/2} \sqrt{\log(1/x^2)}. \end{aligned}$$

Since $\int_0^{1/2} \sqrt{\log(1/x^2)} \le 1$, we get

$$\sum_{j>1} \sqrt{\varepsilon_{j-1} \log(n(\varepsilon_j))} \le 8\sqrt{KP(k,d)\log(k^2(4k-2))} := Q(k,d).$$

Thus

$$\mathscr{R}_n \le \frac{4Q(k,d)\sqrt{C_\infty}}{\sqrt{n}} \sqrt{\|F_r\|_{L_2(P_n)}\|}.$$

It remains now to take expectations with respect to the n-sample $X_1, \dots, X_n$. Since $x \mapsto \sqrt{x}$ is a concave map,

$$\mathbb{E}_X\left(\sqrt{\|F_r\|_{L_2(P_n)}}\right) \le \sqrt{\|F_r\|_{L_2(P)}} = \sqrt{C_2(r)}.$$

Gathering all terms leads to

$$\Psi_3(r) \le \frac{8Q(k,d)\sqrt{C_\infty C_2(r)}}{\sqrt{n}}.$$

Replacing $\sqrt{\delta}/C_\infty$ with r gives the result of Proposition 3.8.

3.5.3 Proof of Proposition 3.11

Let S be a finite subset of $\mathbb{R}^d$, and denote by $\mathscr{C}(p)$ the set of subsets of $\mathbb{R}^d$ which are intersections of at most p half spaces (closed or open). For short, $\mathscr{N}(\mathscr{F},\varepsilon)$ will denote $\mathscr{N}(\mathscr{F},L_2(S),\varepsilon)$.

The proof of Proposition 3.11 is based on the following result, Theorem 1 in [MV03].

Theorem 3.4. *Let P denote a measure on Ω. Let $\mathscr{F}$ be a set of maps from Ω into $[-1,1]$. Then, for every $0<t<1$,*

$$\mathscr{N}(\mathscr{F},t,L_2(P)) \leq \left(\frac{2}{t}\right)^{Kvc(\mathscr{F},ct)},$$

where K and c are constants, and $vc(\mathscr{F},ct)$ denotes the t-shattering dimension of $\mathscr{F}$, as defined in [MV03].

Note that, for every $t>0$, $vc(\mathscr{F},ct) \leq d_p(\mathscr{F})$, where $d_p(\mathscr{F})$ denotes the pseudo-dimension of $\mathscr{F}$, that is the largest integer p such that there exists $x_1,\ldots,x_p \in \Omega$, and $t_1,\ldots,t_p$ real numbers, satisfying the following property : for every $\sigma \in \{-1,1\}^p$ there exists $f_\sigma \in \mathscr{F}$ such that, for $i=1,\ldots,p$, $\sigma_i(f_\sigma(x_i)-t_i)>0$. As a consequence, the quantity of interest is $d_p(\mathscr{G}(r))$.

Recalling that every g in $\mathscr{G}$ can be written

$$R(\mathbf{c},\mathbf{c}^*,x) = \|\mathbf{c}-\mathbf{c}^*\|^{-1} \sum_{i,j} \mathbb{1}_{V_i(\mathbf{c}^*)\cap W_j(\mathbf{c})\cap \mathscr{B}(0,M)} \Big(\|c_i - c_i^*\|^2 + 2\left\langle x - \frac{c_i+c_j}{2}, c_i - c_j \right\rangle \Big),$$

where, for every j in $\{1,\ldots,k\}$, $W_j(\mathbf{c})$ is in $\mathscr{C}(k-1)$, it may be noticed that g takes the form of a sum of k^2 maps of the type $\ell \mathbb{1}_C \mathbb{1}_{\mathscr{B}(0,1)}$, where ℓ denotes an affine map, and C is an element of $\mathscr{C}(2(k-1))$. Let $\mathscr{A}ff(\mathbb{R}^d,\mathbb{R})$ denote the space of affine maps between $\mathbb{R}^d$ and $\mathbb{R}$.

It is worth pointing out that every map $\ell \mathbb{1}_C \mathbb{1}_{\mathscr{B}(0,1)}$ involved in the above decomposition of $R(\mathbf{c},\mathbf{c}^*,.)$ admits F_r as an envelop.

Denote by

$$\mathscr{G}'(r) = \left\{ \frac{\ell \mathbb{1}_C \mathbb{1}_{\mathscr{B}}(0,M)}{F_r} ; \ell \in \mathscr{A}ff(\mathbb{R}^d,\mathbb{R}), C \in \mathscr{C}(2(k-1)) \right\}.$$

We immediately deduce that

$$\mathscr{N}(\mathscr{G}(r),\varepsilon) \leq \left(\mathscr{N}(\mathscr{G}'(r),\varepsilon/k^2)\right)^{k^2}.$$

Consider now the set of functions $\mathscr{N}(\mathscr{G}'(r),\varepsilon)$. The following lemma offers a decomposition of $\mathscr{N}(\mathscr{G}'(r),\varepsilon)$.

Lemma 3.8. *Denote by $\mathscr{F}_s$ $s=1,\ldots,p$ a collection of set of functions taking values in $[-1,1]$. Then*

$$\mathscr{N}\left(\prod_{s=1}^p \mathscr{F}_s, \varepsilon\right) \leq \prod_{s=1}^p \mathscr{N}(\mathscr{F}_s,\varepsilon/p).$$

In order to apply Lemma 3.8, a crucial point is to only deal with maps taking values in $[-1,1]$. To this aim, we define the set

$$\mathscr{H} = \left\{ \frac{f \mathbb{1}_{\{|f| \leq F_r\}}}{F_r} \right\},$$

where f is in $\mathscr{A}ff(\mathbb{R}^d, \mathbb{R})$, and the set

$$\mathscr{P}_d = \left\{ \left\{ \mathbb{1}_{\{\ell \leq a\}} | \ell \in \mathscr{L}(\mathbb{R}^d, \mathbb{R}), a \in \mathbb{R} \right\} \cup \left\{ \mathbb{1}_{\{\ell < a\}} | \ell \in \mathscr{L}(\mathbb{R}^d, \mathbb{R}), a \in \mathbb{R} \right\} \right\},$$

where $\mathscr{L}(\mathbb{R}^d, \mathbb{R})$ denotes the set of linear maps from $\mathbb{R}^d$ to $\mathbb{R}$. We may write

$$\mathscr{G}'(r) \subset \mathscr{H} \times \prod_{i=1}^{2(k-1)} \mathscr{P}_d \times \mathbb{1}_{\mathscr{B}(0,M)},$$

It is well known that

$$d_p(\mathscr{P}_d) = d.$$

Since every set of functions in this decomposition is composed of functions taking values in $[-1,1]$, we intend to apply Theorem 3.4 to every set. Consequently it remains to bound from above the pseudo-dimensions of these sets of functions.

First we deal with $\mathscr{H}$:

Lemma 3.9. *One has*

$$d_p(\mathscr{H}) = d_p\left(\left\{ f \mathbb{1}_{\{|f| \leq F_r\}} | f \in \mathscr{A}ff(\mathbb{R}^d, \mathbb{R}) \right\} \right) \leq 24(3d+4).$$

Proof of Lemma 3.9. The first equality is obvious, so we only have to deal with the inequality. We recall that the pseudo-dimension of the set of functions

$$\left\{ f \mathbb{1}_{\{|f| \leq F_r\}} | f \in \mathscr{A}ff(\mathbb{R}^d, \mathbb{R}) \right\}$$

is the Vapnik dimension of the set of functions

$$\mathscr{H}' = \left\{ \mathbb{1}_{\{f \mathbb{1}_{\{|f| \leq F_r\}} - t \leq 0\}} | f \in \mathscr{A}ff(\mathbb{R}^d, \mathbb{R}), t \in \mathbb{R} \right\}.$$

Let $x_1, \ldots, x_{2m}$ denote $2m$ points in $\mathbb{R}^d$. Since $F_r(x) = c_1 + c_2 \mathbb{1}_{N^*(x)}$, where c_1 et c_2 are constants, at least m points fall in an area on which F_r takes the form $F_r(x) = c$, for some constant c. Without loss of generality, we assume that $x_1, \ldots, x_m$ fall in such an area. Consequently, we have to bound from above the quantity $\left| \left\{ \mathbb{1}_{\{f \mathbb{1}_{\{|f| \leq c\}} - t \leq 0\}} \right\} (x_1, \ldots, x_m) \right|$. Observing that

$$\left\{ \mathbb{1}_{\{|f| \leq c\}} \right\} = \left\{ \mathbb{1}_{\{f \leq c\}} \times \mathbb{1}_{\{f \geq -c\}} \right\},$$

we deduce that

$$\begin{aligned} &\left| \left\{ \mathbb{1}_{\{f \mathbb{1}_{\{|f| \leq c\}} - t \leq 0\}} \right\} (x_1, \ldots, x_m) \right| \\ &\qquad \leq \left| \left\{ \mathbb{1}_{\{f \mathbb{1}_{\{f \leq c\}} - t \leq 0\}} \right\} (x_1, \ldots, x_m) \right| \times \left| \left\{ \mathbb{1}_{\{f \mathbb{1}_{\{f \geq -c\}} - t \leq 0\}} \right\} (x_1, \ldots, x_m) \right|. \end{aligned}$$

Noticing that $d_{VC}\left(\left\{ \mathbb{1}_{\{f \leq c\}} | f \in \mathscr{A}ff(\mathbb{R}^d, \mathbb{R}) \right\} \right) = d+1$ and making use of Sauer's lemma leads to, provided that $m \geq d+1$,

$$\left| \left\{ \mathbb{1}_{\{f \leq c\}} \right\} (x_1, \ldots, x_m) \right| = \left| \left\{ \mathbb{1}_{\{f \geq -c\}} \right\} (x_1, \ldots, x_m) \right| \leq \left(\frac{em}{d+1} \right)^{(d+1)},$$

which ensures that

$$\left|\left\{\mathbb{1}_{\{|f|\leq c\}}\right\}(x_1,\dots,x_m)\right| \leq \left(\frac{em}{d+1}\right)^{2(d+1)}.$$

Choose a configuration of $\left\{\mathbb{1}_{|f(x_1)|\leq c},\dots,\mathbb{1}_{|f(x_m)|\leq c}\right\}$, for instance by indexing the x_i's so that $|f(x_1)|>c,\dots,|f(x_r)|>c$ and $|f(x_{r+1})|\leq c,\dots,|f(x_m)|\leq c$. For the r first x_i's, only two configurations remains for $\mathscr{H}'(x_1,\dots,x_r)$, the configuration $(0,\dots,0)$ and the configuration $(1,\dots,1)$. Concerning the $m-r+1$ last x_i's, we may write $\left|\mathscr{H}'(x_{m-r+1},\dots,x_m)\right| \leq \left|\left\{\mathbb{1}_{\{f-t\leq 0\}}\right\}(x_{m-r+1},\dots,m)\right|$. Next, $\left|\left\{\mathbb{1}_{\{f-t\leq 0\}}\right\}(x_{m-r+1},\dots,m)\right| \leq \left|\left\{\mathbb{1}_{\{f-t\leq 0\}}\right\}(x_1,\dots,m)\right|$. Eventually, a direct application of Sauer's lemma guarantees that $\left|\left\{\mathbb{1}_{\{f-t\leq 0\}}\right\}(x_1,\dots,m)\right| \leq \left(\frac{em}{d+2}\right)^{(d+2)}$, provided that $m\geq d+2$. Consequently, we get, for $m\geq d+2$,

$$\begin{aligned}
\left|\mathscr{H}'(x_1,\dots,x_{2m})\right| &= \left|\left\{\mathbb{1}_{\{f\mathbb{1}_{\{|f|\leq F_r\}}-t\leq 0\}}\right\}(x_1,\dots,x_{2m})\right| \\
&\leq 2^m \times \left|\left\{\mathbb{1}_{\{f\mathbb{1}_{\{|f|\leq c\}}-t\leq 0\}}\right\}(x_1,\dots,x_m)\right| \\
&\leq 2^m \times \left|\left\{\mathbb{1}_{\{|f|\leq c\}}\right\}(x_1,\dots,x_m)\right| \times 2\left(\frac{em}{d+2}\right)^{d+2} \\
&\leq 2^m \times 2\left(\frac{em}{d+1}\right)^{2(d+1)}\left(\frac{em}{d+2}\right)^{d+2}.
\end{aligned}$$

To give an upper bound on $d_p(\mathscr{H})$, we have to find $m\geq d+2$ such that

$$2\left(\frac{em}{d+1}\right)^{2(d+1)}\left(\frac{em}{d+2}\right)^{d+2} < 2^m.$$

Noticing that $x\mapsto \log_2(x)$ is a strictly concave map, we deduce that

$$2(d+1)\log_2\left(\frac{em}{d+1}\right)+(d+2)\log_2\left(\frac{em}{d+2}\right) < (3d+4)\log_2\left(\frac{3em}{3d+4}\right).$$

Consequently, a sufficient condition on m is given by

$$\left(\frac{3em}{3d+4}\right)^{3d+4} \leq 2^{m-1}.$$

Using the same method as in [BH89], the choice $m=\left\lceil 3(3d+4)\log_2(3e)\right\rceil$ turns out to be adequate. At last, noticing that $2m\leq 24(3d+4)$, we immediately deduce that $d_p(\mathscr{H})\leq 2m\leq 24(3d+4)$. □

Applying Lemma 3.8 and Theorem 3.4 yields

$$\begin{aligned}
\mathscr{N}\left(\mathscr{G}'(r),\varepsilon\right) &\leq \mathscr{N}\left(\mathscr{H},\varepsilon/(2k-1)\right)\times \mathscr{N}\left(\mathscr{P}_d,\varepsilon/(2k-1)\right)^{2k-2} \\
&\leq \left(\frac{4k-2}{\varepsilon}\right)^{K[24(3d+4)+(d+1)(2k-2)]}.
\end{aligned}$$

At last, the result of Proposition 3.11 is given by

$$\begin{aligned}
\mathscr{N}\left(\mathscr{G}(r),\varepsilon\right) &\leq \mathscr{N}\left(\mathscr{G}'(r),\varepsilon/k^2\right)^{k^2} \\
&\leq \left(\frac{2(2k-1)k^2}{\varepsilon}\right)^{Kk^2[24(3d+4)+2(d+1)(k-1)]}.
\end{aligned}$$

3.5.4 Proof of Proposition 3.10

The proof of Proposition 3.10 is based on elementary properties of distributions with finite support, which are extended to the case where the source distribution is supported on small balls. Throughout this Subsection, a source distribution $P_{\sigma'}$ is fixed, so that $R(Q, P_{\sigma'})$ may be denoted by $R(Q)$.

Lemma 3.10. *Let z_1 and z_2 be points in $\mathbb{R}^d$, denote by R the quantity $\|z_1 - z_2\|$, by U_i the ball $\mathcal{B}(z_i, \rho)$. At last, let P denote the cone-shaped distribution with density*

$$\frac{2(d+1)}{V}(\mathbb{1}_{\|x-z_i\|\leq\rho}(\rho - \|x - z_i\|),$$

over each ball U_i, where V denote the volume of the unit ball. Then, if

$$(R/2 - 3\rho)^2 \geq \rho^2 \frac{2d(d+1)}{(d+2)(d+3)} \quad \text{and} \quad \rho \leq \frac{R}{2},$$

then the best 2-quantizer Q_2^* is such that $Q_2^*(U_i) = z_i$ for $i = 1, 2$. Furthermore, the best 1-quantizer Q_1^* is such that $Q_1^*(U_1 \cup U_2) = (z_1 + z_2)/2$.

Proof of Lemma 3.10. Let V_i denote the Voronoi cell associated with z_i in the Voronoi diagram generated by (z_1, z_2). Denote by Q_2^* the quantizer satisfying $Q_2^*(U_i) = z_i$ for $i = 1, 2$.

For any quantizer Q denote by $R_i(Q) = \int_{V_i} \|x - Q(x)\|^2 dx$ the contribution of the cell i to the distortion of Q. Denote by V the volume of the unit ball, and by S its surface. Recalling that $S = d \times V$, an elementary calculation shows that

$$\begin{aligned} R_i(Q_2^*) &= \frac{1}{2}\frac{d+1}{\rho^{d+1}V}\int_0^\rho S(\rho r^{d+1} - r^{d+2})dr \\ &= \rho^2 \frac{d(d+1)}{2(d+2)(d+3)}. \end{aligned}$$

Let $m_i^{in} = |Q(U_i) \cap V_i|$ and $m_i^{out} = |Q(U_i)) \cap V_i^c|$ denote the number of images of U_i sent inside and outside V_i. For a given i, there are three situations of interest, which are described below.

1. $m_i^{out} = 0$ and $m_i^{in} = 1$, then it is clear that $R_i(Q_2^*) \leq R_i(Q)$.
2. $m_i^{out} = 0$ and $m_i^{in} = 2$, then $R_i(Q) \geq 0 = R_i(Q^*) - \rho^2 \frac{d(d+1)}{2(d+2)(d+3)}$.
3. $m_i^{out} \geq 1$, then there exists $z \in U_i$ such that $Q(z) \notin V_i$. Consequently, $\|z - Q(z)\| \geq \frac{R}{2} - \rho$. Let $z' \in \mathcal{B}(z_i, \rho)$, then

$$\|z' - Q(z')\| \geq \|z - Q(z')\| - 2\rho \geq \|z - Q(z)\| - 2\rho \geq \frac{R}{2} - 3\rho.$$

Hence we deduce

$$R_i(Q) \geq 1/2(\frac{R}{2} - 3\rho)^2 = R_i(Q_2^*) + 1/2\left((\frac{R}{2} - 3\rho)^2 - \rho^2 \frac{d(d+1)}{(d+2)(d+3)}\right).$$

Since Q is a 2-quantizer, it is easy to see that

$$|\left\{i; m_i^{in} \geq 2\right\}| \leq |\left\{i; m_i^{out} \geq 1\right\}|.$$

From this we deduce that

$$R(Q) = \sum_{\{i;m_i^{in}\geq 2, m_i^{out}=0\}} R_i(Q) + \sum_{\{i;m_i^{out}\geq 1\}} R_i(Q) + \sum_{\{i;m_i^{in}=1, m_i^{out}=0\}} R_i(Q)$$
$$\geq R(Q_2^*) + \sum_{\{i;m_i^{in}\geq 2, m_i^{out}=0\}} \frac{1}{2}\left(\left(\frac{R}{2}-3\rho\right)^2 - \rho^2\frac{d(d+1)}{(d+2)(d+3)}\right).$$

Taking $(R/2-3\rho)^2 \geq \rho^2\frac{2d(d+1)}{(d+2)(d+3)}$ ensures that $R(Q) \geq R(Q_2^*)$. □

Considering the distributions P_σ, σ in $\{-1,+1\}^m$, taking $\rho \leq \frac{\Delta}{16}$ ensures that the conditions of Lemma 3.10 are satisfied when considering $P_{\sigma|U_i\cup U_i'}$. We turn now to the proof of Proposition 3.10.

Let Q be a k-quantizer. The following construction provides $Q_\sigma \in \mathscr{Q}$ such that $R(Q_\sigma) \leq R(Q)$. Let V_i denote the union of the Voronoi cells associated with z_i and $z_i+\omega_i$, in the Voronoi diagram generated by the sequences z and ω. We adopt the following notation

$$\begin{cases} n_i(Q) &= |Q(\mathscr{B}(0,M)) \cap V_i|, \\ n_i^{out}(Q) &= \left|Q(V_i)\cap V_i^c\right|, \\ I_j(Q) &= \{i; n_i(Q)=j\}, \\ i_j(Q) &= \left|I_j(Q)\right|, \\ i_{\geq j}(Q) &= \sum_{i\geq j} i_j(Q). \end{cases}$$

The first step is to add code points to empty cells. From the k-quantizer Q, a quantizer Q_1 is built as follows.

- If $n_i(Q) \geq 1$, then we take $Q_{1|V_i} \equiv Q_{|V_i}$.
- If $n_i(Q) = 0$, then we set $Q_1(U_i) = Q_1(U_i') = z_i + \frac{w_i}{2}$.

Notice that Q_1 is a $(k+i_0(Q))$-quantizer. Denote $k_1 = k+i_0$ and $p_\pm = \frac{1\pm\delta}{2m}$, then $R(Q_1)$ can be bounded as follows.

Let i be an integer between 1 and m. We denote by $R_i(Q)$ the contribution of V_i to the risk $R(Q)$. If $i \in I_{\geq 1}$, then $R_i(Q) = R_i(Q_1)$. Otherwise, if $i \in I_0(Q)$,

$$R_i(Q_1) = 2p_\pm\rho^2\frac{d(d+1)}{(d+2)(d+3)} + p_\pm\frac{\Delta^2}{2}.$$

Furthermore, if $i \in I_0$, then $n_i^{out}(Q) \geq 1$, which ensures that, as in the proof of Lemma 3.10,

$$R_i(Q) \geq p_\pm\left(\frac{(A-2)\Delta}{2} - 2\rho\right)^2.$$

Since $A \geq 6$ and $\rho \leq \frac{\Delta}{16}$, we may write

$$\begin{aligned} R_i(Q) - R_i(Q_1) &\geq p_\pm\left[(2\Delta-2\rho)^2 - 4\rho^2 - \frac{\Delta^2}{2}\right] \\ &\geq p_\pm\left[2\Delta(\frac{3\Delta}{4}) - \frac{\Delta^2}{2}\right] \\ &\geq p_-\frac{3\Delta^2}{2}. \end{aligned}$$

Summing all the contributions of V_i's leads to

$$R(Q_1) \leq R(Q) - i_0(Q)p_-\frac{3\Delta^2}{2}.$$

Next, we build the quantizer Q_2 according to the following rule.

– If $n_i(Q_1) \geq 2$, then $Q_2(U_i) = z_i$ and $Q_2(U_i') = z_i + w_i$.
– If $n_i(Q_1) = 1$, then $Q_2(U_i) = Q_2(U_i') = z_i + \frac{w_i}{2}$.

Since for $i = 1, \ldots, k$, $n_i(Q_1) \geq 1$, Q_2 is defined on every V_i. Notice that, since $I_j(Q_1) = I_j(Q)$ for $j \geq 2$, Q_2 has $k_2 = k + i_0(Q) - \sum_{p \geq 3}(p-2)i_p(Q)$ code points. The following lemma offers a relation between $R(Q_2)$ and $R(Q_1)$.

Lemma 3.11. *One has*

$$R(Q_2) \leq R(Q_1) + i_{\geq 3}(Q)\frac{p_+\Delta^2}{128}.$$

Proof of Lemma 3.11. Let i be an integer between 1 and m. Several cases may occur, as described below.

– Assume that $n_i(Q_1) = 1$.
  – If $n_i^{out}(Q_1) = 0$, then $R_i(Q_1) \geq R_i(Q_2)$, according to Lemma 3.10.
  – If $n_i^{out}(Q_1) \geq 1$, then, using the same technique as mentioned to bound $R(Q_1)$ from above, $R_i(Q_1) - R_i(Q_2) \geq p_\pm \frac{3\Delta^2}{2}$, which leads to $R_i(Q_1) \geq R_i(Q_2)$.
– Assume that $n_i(Q_1) = 2$.
  – If $n_i^{out}(Q_1) = 0$, then $R_i(Q_1) \geq R_i(Q_2)$, according to Lemma 3.10.
  – If $n_i^{out}(Q_1) \geq 1$, then, since $R_i(Q_2) = 2p_\pm \frac{\rho^2 d}{d+2} \leq p_+ \frac{\Delta^2}{128}$, $R_i(Q_1) - R_i(Q_2) \geq \Delta^2 \geq 0$.
– At last, assume that $n_i(Q_i) \geq 3$. If $n_i^{out}(Q_1) \geq 1$, then $R_i(Q_1) \geq R_i(Q_2)$. If $n_i^{out}(Q_1) = 0$, then $R_i(Q_1) \geq 0 = R_i(Q_1) - 2p_\pm \frac{\Delta^2}{128}$. In both cases $R(Q_2) \leq R(Q_1) + p_+ \frac{\Delta^2}{128}$.

Noticing that $I_{\geq 3}(Q_1) = I_{\geq 3}(Q)$, and summing the contributions $R_i(Q_2)$ leads to the desired result. □

The last step is to build a quantizer Q_σ from Q_2 with exactly k code points.
– If $k_2 = k$, set $Q_\sigma = Q_2$.
– If $k_2 < k$, choose $(k - k_2)$ V_i such that $n_i(Q_2) = 1$ (elementary calculation shows that there exist at least $k - k_2$ such V_i's). For every such V_i, set $Q_\sigma(U_i) = z_i$ and $Q_\sigma(U_i') = z_i + \omega_i$. Then

$$R(Q_\sigma) \leq R(Q_2) - (k - k_2)p_-\frac{\Delta^2}{2}.$$

– If $k_2 > k$, choose $(k_2 - k)$ cells V_i such that $n_i(Q_2) = 2$. For every such V_i, define $Q_\sigma(U_i) = Q_\sigma(U_i') = z_i + \frac{\omega_i}{2}$. Then

$$R(Q_\sigma) \leq R(Q_2) + (k_2 - k)p_+\frac{\Delta^2}{2}.$$

By construction, Q_σ has exactly k code points, and is an element of $\mathcal{Q}$. Finally, a result on the risk of Q_σ is given by the following Proposition.

Proposition 3.12. *Let Q be a quantizer and Q_σ built as mentioned above. Then,*

$$R(Q_\sigma) \leq R(Q).$$

Proof of Proposition 3.12. Since $\delta \leq \frac{1}{3}$, easy calculation ensures that $1 - \frac{p_-}{p_+} \leq \frac{1}{2}$.

Suppose that $k_2 \le k$. Comparing the risk of Q to the risks of Q_1, Q_2 and Q_σ leads to

$$R(Q_\sigma) \le R(Q) - i_0 p_- \frac{3\Delta^2}{2} + (i_0 + 2i_{\ge 3} - \sum_{p\ge 3} p i_p) p_- \frac{\Delta^2}{2} + i_{\ge 3} p_+ \frac{\Delta^2}{128}.$$

Since $\sum_{p\ge 3} p i_p \ge 3 i_{\ge 3}$, it is clear that

$$\begin{aligned} R(Q_\sigma) &\le R(Q) - p_- i_0 \frac{\Delta^2}{2} + \Delta^2 i_{\ge 3} (\frac{p_+}{128} - \frac{p_-}{2}) \\ &\le R(Q). \end{aligned}$$

Next, suppose that $k_2 > k$. Then

$$\begin{aligned} R(Q_\sigma) &\le R(Q) + \left(i_0 + 2i_{\ge 3} - \sum_{p\ge 3} p i_p \right) p_+ \frac{\Delta^2}{2} + i_{\ge 3} p_+ \frac{\Delta^2}{128} - i_0 p_- \frac{3\Delta^2}{2} \\ &\le R(Q) + i_0 \frac{\Delta^2}{2} (p_+ - 3p_-) + p_+ i_{\ge 3} \Delta^2 (\frac{1}{128} - \frac{1}{2}), \end{aligned}$$

which yields $R(Q_\sigma) \le R(Q)$. □

3.5.5 Proof of Lemma 3.6

Let introduce, for distributions P and Q with densities f and g the affinity

$$\alpha(P,Q) = \int \sqrt{fg},$$

so that $H^2(P,Q) = 2(1 - \alpha(P,Q))$. Elementary calculation shows that, if $\rho(\sigma,\sigma') = 4$, then

$$\alpha(P_\sigma, P_{\sigma'}) = 1 + \frac{2}{m}\left(\sqrt{1-\delta^2} - 1\right) \ge 1 - \frac{2\delta^2}{m}.$$

Hence we deduce

$$\begin{aligned} H^2(P_\sigma^{\otimes n}, P_{\sigma'}^{\otimes n}) &= 2(1 - \alpha(P_\sigma^{\otimes n}, P_{\sigma'}^{\otimes n})) \\ &= 2(1 - \alpha^n(P_\sigma, P_{\sigma'})) \\ &\le \frac{4n\delta^2}{m}. \end{aligned}$$

Finally, since $\rho(\tau,\tau') = 2$ implies $\rho(\sigma(\tau),\sigma(\tau')) = 4$, for τ, τ' in $\{-1,+1\}^{\frac{m}{2}}$, the first part of Lemma 3.6 is proved.

Next, for simplicity assume that σ is such that $\sigma_1 = \ldots \sigma_{\frac{m}{2}} = +1$ and $\sigma_{\frac{m}{2}+1} = \ldots = \sigma_m = -1$. Let $\mathscr{S}^-$ and $\mathscr{S}^+$denote the set of mistakes of σ', that is

$$\begin{cases} \mathscr{S}^- &= \{i = 1,\ldots,\frac{m}{2} | \quad \sigma'_i = -1\}, \\ \mathscr{S}^+ &= \{i = \frac{m}{2}+1,\ldots,m | \quad \sigma'_i = +1\}. \end{cases}$$

Finally let s^+ and s^- respectively denote $|\mathscr{S}^+|$ and $|\mathscr{S}^-|$. Since $\sum_{i=1}^m \sigma'_i = 0$, it is clear that $s^+ = s^- := s$.

As in Subsection 3.5.4, let $R_i(Q_{\sigma'})$ denote the contribution of $U_i \cup U'_i$ to the distortion. Then, for i in $\mathscr{S}^-$, elementary calculation shows that

$$R_i(Q_{\sigma'}) = R_i(Q_\sigma) + \frac{(1+\delta)\Delta^2}{4m}.$$

Symmetrically, for i in $\mathscr{S}^+$,

$$R_i(Q_{\sigma'}) = R_i(Q_\sigma) - \frac{(1-\delta)\Delta^2}{4m}.$$

Summing with respect to i and taking into account that $s^+ = s^- = s$ leads to

$$R(Q_{\sigma'}) = R(Q_\sigma) + s\frac{\Delta^2\delta}{2m}.$$

Remarking that $s = \frac{\rho(\sigma,\sigma')}{4}$ concludes the proof of Lemma 3.6.

Chapitre 4

Variable selection for k-means quantization

Les résultats obtenus dans le chapitre précédent laissent penser que la quantification vectorielle par la méthode de minimisation du risque empirique serait adaptée aux problèmes de grande dimension. Pour ce type de problèmes, une procédure de sélection de variables est habituellement appliquée avant toute tentative de classification. Nous nous sommes intéressés à une procédure de type Lasso k-means, combinant classification et sélection de variables. Bien que de tels algorithmes aient déjà été implémentés (voir par exemple [SWF12]), peu de résultats théoriques sur cette méthode ont été prouvés. Le Chapitre 4 montre que, sous la condition de marge définie dans le Chapitre 3, cette procédure de Lasso k means fournit des dictionnaires parcimonieux, et converge vers des approximations parcimonieuses des dictionnaires optimaux. De plus, des bornes supérieures sur la perte des dictionnaires Lasso k means sont données, du même ordre de grandeur que celles obtenues dans [vdG08] pour les modèles linéaires généralisés.

Sommaire

Recent results in quantization theory provide theoretical bounds on the distortion of squared-norm based quantizers (see, e.g., [BDL08] or Theorem 3.1 in Chapter 3). These bounds are valid whenever the source distribution has a bounded support, regardless of the dimension of the underlying Hilbertian space.

However, it remains of interest to select relevant variable for quantization. This task is usually performed through a coordinate energy-ratio thresholding (see, e.g., [ABCP13] or [SB08]), or maximizing a constrained empirical Between Cluster Sum of Squares criterion (see, e.g., [CWLX14] or [WT10]). This paper offers a Lasso type procedure to select the relevant variables for k-means clustering, as exposed in [SWF12]. Moreover, some non-asymptotic convergence results on the distortion are derived for this procedure, along with consistency results toward sparse codebooks.

4.1 Introduction

Let P be a distribution over $\mathbb{R}^d$. Quantization is the issue of replacing P with a finite set of points, without loosing too much information. To be more precise, if k denotes an integer, a k points quantizer Q is defined as a map from $\mathbb{R}^d$ into a finite subset of $\mathbb{R}^d$ with cardinality k. In other words, a k-quantizer divide $\mathbb{R}^d$ into k groups, and assigns each group a representative.

The quantization theory was originally developed as a way to answer signal compression issues in the late 40's (see, e.g., [GG91]). However, unsupervised classification is also in the scope of its application. Isolating meaningful groups from a cloud of data is a topic of interest in many fields, from social science to biology.

Assume that P has a finite second moment, and let Q be a k points quantizer. The performance of Q in representing P is measured by the distortion

$$R(Q) = P\|x - Q(x)\|^2,$$

where Pf means integration of f with respect to P. It is worth pointing out that many other distortion functions can be defined, using $\|x - Q(x)\|^r$ or more general distance functions (see, e.g., [Fis10] or [GL00]). However, the choice of the Euclidean squared norm is convenient, since it allows to fully take advantage of the Euclidean structure of $\mathbb{R}^d$, as described in Chapter 3. Moreover, from a practical point of view, the k-means algorithm (see [Llo82]) is designed to minimize this squared-norm distortion and can be easily implemented.

Since the distortion is based on the Euclidean distance between a point and its image, it is well known that only nearest-neighbor quantizers are to be considered (see, e.g., [GL00] or [Pol82b]). These quantizers are quantizers of the type $x \mapsto \arg\min_{j=1,\dots,k} \|x - c_j\|$, where the c_i's are elements of $\mathbb{R}^d$ and are called code points. A vector of code points $(c_1, \dots, c_k)$ is called a codebook, so that the distortion takes the form

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

It has been proved in [Pol81] that, whenever $P\|x\|^2 < \infty$, there exists optimal codebooks, denoted by $\mathbf{c}^*$.

Let $X_1, \dots, X_n$ denote an independent and identically distributed sample drawn from P, and denote by P_n the associated empirical distribution, namely $P_n(A) = 1/n\,|\{i|X_i \in A\}|$, for every measurable subset A. The aim is to design a codebook from this n-sample, whose distortion is as close as possible to the optimum $R(\mathbf{c}^*)$. The k-means algorithm provides the empirical codebook $\hat{\mathbf{c}}_n$, defined by

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

It is worth pointing out that, if $P^{(p)} \neq \delta_0$, where $P^{(p)}$ denotes the marginal distribution of P on the p-th coordinate and δ_0 denotes the Dirac distribution at point 0, then $\hat{\mathbf{c}}_n^{(p)} = (\hat{c}_1^{(p)}, \dots, \hat{c}_k^{(p)}) \neq 0$. This shows that the k-means algorithm does not provide sparse solutions, even if $P^{(p)}$ is a noise distribution.

Consequently, when d is large, a variable selection procedure is usually performed preliminary to the k-means algorithm. The variable selection can be achieved using penalized BCCS strategies, as exposed in [CWLX14] or [WT10]. Though these procedures offer good performance in classifying the sample $X_1, \dots, X_n$, under the assumption that the marginal distributions $P^{(j)}$ are independent, no theoretical result on the prediction performance has been given. An other way to perform variable selection can be to select coordinates whose empirical variances are larger than a determined ratio of the global variance, following the idea of [SB08]. This algorithm has shown good results on practical examples, such as curve clustering (see, e.g., [ABCP13]). However, there is no theoretical result on the prediction performance of the selected coordinates. At last, it may be interesting to consider the variable selection issue as a subspace selection issue. In this framework, procedures based on PCA combined with k-means, such as Reduced K-means and Factorial K-means (see, e.g., [DeS] and [VK01]), offer good performance in practice. In addition, some theoretical results in prediction are available (see, e.g., [Ter12] and [Ter13]), ensuring that the empirically designed codebooks converge amost surely to optimal low-dimensional codebooks. However, no convergence rate has been derived so far for these algorithms, and the choice of the dimension of candidate subspaces remains a major issue.

This paper exposes a theoretical study of a Lasso type procedure combined with the k-means procedure, as suggested in [SWF12]. Some results on the prediction performance and on the consistency to a sparse codebook are derived for this procedure, in the spirit of [vdG13]. Some sparsity results on the empirical codebook are also given. It is worth pointing out that these results are valid when P satisfies a margin condition, as defined in Definition 3.1 in Chapter 3, extending the scope of the asymptotic results proposed in [SWF12].

The paper is organized as follows. Some notation are introduced in Section 4.2, along with the Lasso k-means procedure and the different assumptions. The consistency and prediction results are gathered in Section 4.3, and the proof of these results are exposed in Section 4.5. At last, technical proofs are to be found in Section 4.6.

4.2 Notation

Let x be in $\mathbb{R}^d$, then the p-th coordinate of x will be denoted by $x^{(p)}$. Throughout this paper, it is assumed that, for every $p = 1,\dots,d$, there exist a sequence M_p, such that $|x^{(p)}| \leq M_p$, P-almost surely. In other words P is assumed to have bounded marginal distributions $P^{(p)}$. To shorten notation, the Euclidean coordinate-wise product $\prod_{p=1}^{d}[-M_p, M_p]$ will be denoted by C. To frame quantization as a contrast minimization issue, let us introduce the following contrast function

$$\gamma : \begin{cases} \left(\mathbb{R}^d\right)^k \times \mathbb{R}^d & \longrightarrow & \mathbb{R} \\ (\mathbf{c}, x) & \longmapsto & \min\limits_{j=1,\dots,k} \left\|x - c_j\right\|^2, \end{cases}$$

where $\mathbf{c} = (c_1,\dots,c_k)$ denotes a codebook, that is a kd-dimensional vector. The risk $R(\mathbf{c})$ then takes the form $R(\mathbf{c}) = R(Q) = P\gamma(\mathbf{c},.)$, where we recall that Pf denotes the integration of the function f with respect to P. Similarly, the empirical risk $\hat{R}_n(\mathbf{c})$ can be defined as $\hat{R}_n(\mathbf{c}) = P_n\gamma(\mathbf{c},.)$, where P_n is the empirical distribution associated with $X_1,\dots,X_n$, in other words $P_n(A) = 1/n\,|\{i|X_i \in A\}|$, for every measurable subset $A \subset \mathbb{R}^d$.

It is worth pointing out that, if $P\|x\|^2 < \infty$, then there exist such minimizers $\hat{\mathbf{c}}_n$ and $\mathbf{c}^*$ (see, e.g., Theorem 4.12 in [GL00]). Throughout this paper it is assumed that there exists a unique optimal quantizer $\mathbf{c}^*$, up to relabeling code points.

To size the influence of the different coordinates on the quantization error, the following coordinate-wise quantization error and variance are introduced. Let $S \subset \{1,\dots,d\}$ denote a subset of coordinates, and $P^{(S)}$ denote the marginal distribution of P over the set $\mathbb{R}^{|S|}$. We may define

$$\begin{cases} \sigma_S^2 &= & P^S\|x\|^2, \\ \hat{\sigma}_S^2 &= & P_n^S\|x\|^2, \\ R_S^* &= & \min_{\mathbf{c}\in C^S} P^S\gamma(\mathbf{c},.), \\ \hat{R}_S^* &= & \min_{\mathbf{c}\in C^S} P_n^S\gamma(\mathbf{c},.), \end{cases}$$

where the vector x is element of $\mathbb{R}^{|S|}$. Elementary properties of the distortion show that, if $S = S_1 \cup S_2$, with empty intersection, then

$$\begin{cases} \sigma_S^2 &= & \sigma_{S_1}^2 + \sigma_{S_2}^2, \\ \hat{\sigma}_S^2 &= & \hat{\sigma}_{S_1}^2 + \hat{\sigma}_{S_2}^2, \\ R_S^* &\geq & R_{S_1}^* + R_{S_2}^*, \\ \hat{R}_S^* &\geq & \hat{R}_{S_1}^* + \hat{R}_{S_2}^*. \end{cases} \tag{4.1}$$

These elementary properties will be of importance when choosing which coordinate to select.

The following technical inequality is needed, in order to connect the loss $\ell(\mathbf{c}, \mathbf{c}^*)$ to the distance between codebooks.

Definition 4.1. *Assume that there exists a unique optimal quantizer $\mathbf{c}^*$. Then P satisfies a margin condition if there exists $\kappa_0 > 0$ such that*

$$\forall \mathbf{c} \in C^k \quad \ell(\mathbf{c}, \mathbf{c}^*) \geq \kappa_0 \|\mathbf{c} - \mathbf{c}^*\|^2. \tag{4.2}$$

As exposed in Chapter 3, Definition 4.1 may be thought of as a margin condition in the framework of squared distance based quantization. Some examples of distributions satisfying (4.2) are given in Section 3.2 of Chapter 3. Roughly, if P is well concentrated around k poles, then (4.2) will hold. It is also worth mentioning that the condition required in [SWF12] is much stronger than the condition required in Definition 4.1, since it requires P to be a mixture of components centered on the different optimal code points, and that the Hessian matrix of the risk function located at the optimal codebook is positive definite. As exposed in Chapter 2, the condition mentioned above implies Definition 4.1.

The Lasso k-means procedure, introduced in [SWF12], is defined as follows.

$$\hat{\mathbf{c}}_{n,\lambda} \in \arg\min_{\mathbf{c}\in C^k} P_n\gamma(\mathbf{c},.) + \lambda I(\mathbf{c}), \tag{4.3}$$

where $I(\mathbf{c})$ denotes a possibly weighted penalty function of the codebook $\mathbf{c}$. This paper provides results for two types of penalties $I(\mathbf{c})$: a Lasso type penalty where the weights are chosen to be 1, and a Weighted Lasso type penalty with adaptive weights.

4.2.0.1 Lasso type penalty

In this case the penalty function is chosen as

$$I(\mathbf{c}) = I_L(\mathbf{c}) = \sum_{p=1}^{d} \sqrt{c_1^{(p)2} + \ldots + c_k^{(p)2}}. \tag{4.4}$$

This L_1 type penalty is shaped to drive the irrelevant (p)-th coordinates $c_1^{(p)},\ldots,c_k^{(p)}$ together to zero, as exposed in [Bac08]. The following Proposition gives a theoretical guarantee on the coordinates which are not driven to zero.

Proposition 4.1. *Let p be in $\{1,\ldots,d\}$. If*

$$\sqrt{\hat{\sigma}_p^2 - \hat{R}_p^*} < \frac{\lambda}{2},$$

then

$$\hat{\mathbf{c}}_{n,\lambda}^{(p)} = (\hat{c}_{n,\lambda,1}^{(p)},\ldots,\hat{c}_{n,\lambda,k}^{(p)}) = (0,\ldots,0).$$

Roughly, Proposition 4.1 ensures that the Lasso k-means procedure selects only variables whose empirical quantization error is small compared to its empirical variance. These variables may be interpreted as relevant variables for the empirical k-quantization error. However, when M_p is small, the choice of the penalty $I_L(\mathbf{c})$ will drive the (p)-th coordinates to 0, even if $P^{(p)}$ is supported on k points. This scaling issue can be addressed using a Weighted Lasso penalty, as done in [SWF12].

4.2.0.2 Weighted Lasso type penalty

The original procedure of Lasso k-means exposed in [SWF12] is indeed a Weighted Lasso type procedure. However, different weights are proposed here. For these weights theoretical guarantees are provided on the convergence of the Lasso k-means estimator to a sparse codebook. The proposed penalty function is the following

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

where the empirical coordinate-wise variances are defined above. The following Proposition gives a necessary condition for the p-th coordinate not to be driven to 0.

Proposition 4.2. *Let p be in $\{1,\dots,d\}$. If*

$$\sqrt{1-\frac{\hat{R}_p^*}{\hat{\sigma}_p^2}} < \frac{\lambda}{2},$$

then

$$\hat{\mathbf{c}}_{n,\lambda}^{(p)} = (\hat{c}_{n,\lambda,1}^{(p)},\dots,\hat{c}_{n,\lambda,k}^{(p)}) = (0,\dots,0).$$

The scaling issue mentioned above turns out to be addressed, since only the ratios between empirical variances and empirical k-quantization error are to be considered to determinate relevant variables. As in the Lasso penalty case, coordinates with large ratios between empirical k-quantization error over empirical variance will be driven to zero.

It is worth mentioning that in these two cases non-zero coordinates are only empirically characterized. The following Section provides convergence results to sparse codebooks, along with prediction results.

4.3 Results

4.3.1 Lasso k-means distortion and consistency

Throughout this Subsection the penalty function $I(\mathbf{c})$ is chosen as $I_L(\mathbf{c})$. It is well known that Lasso type procedures may be thought of as model selection procedures over L_1 balls (see, e.g., [MM11]). This leads to the following result.

Theorem 4.1. *Let M_∞ denote $\max_{p=1,\dots,d} M_p$. Choose*

$$\lambda \geq \frac{6kM_\infty\sqrt{2\log(d)}}{\sqrt{n}}\left(1+\frac{1}{2k}\sqrt{\frac{x}{\log(d)}}\right),$$

for some $x>0$. Then, for every $\varepsilon>0$, with probability larger than $1-\left(\frac{\sqrt{k}dM_\infty}{\varepsilon}+1\right)e^{-x}$, we have

$$\ell(\hat{\mathbf{c}}_{n,\lambda},\mathbf{c}^*) \leq \inf_{r>0}\inf_{I_L(\mathbf{c})\leq r}\left(\ell(\mathbf{c},\mathbf{c}^*)+\lambda(2r+3)\varepsilon\right).$$

For any codebook $\mathbf{c}$, let $\|\mathbf{c}\|_0$ be defined by $|\{p|\mathbf{c}^{(p)}\neq(0,\dots,0)\}|$. Furthermore, assume that P satisfies (4.2). Then the best sparse approximation of $\mathbf{c}^*$ at order λ is defined by

$$\mathbf{c}_\lambda^* \in \arg\min_{\mathbf{c}\in C^k} 3\ell(\mathbf{c},\mathbf{c}^*)+\frac{8\lambda^2}{\kappa_0}\|\mathbf{c}\|_0,$$

where κ_0 denotes the constant in (4.2). As in the empirical case of Proposition 4.1, the non-zero coordinates of $\mathbf{c}_\lambda^*$ may be characterized in the following way.

Proposition 4.3. *Let p be in $\{1,\dots,d\}$. If*

$$\sigma_p^2 - R_p^* < \frac{8\lambda^2}{3\kappa_0},$$

then

$$\mathbf{c}_\lambda^{*(p)} = (0,\dots,0).$$

The proof of Proposition 4.3 is given in Section 4.5. Equipped with this Proposition, we are now in position to state convergence results.

Theorem 4.2. *Denote by* $M_\infty = \max_{p=1,\dots,d}$. *If*

$$\lambda \geq 32e\sqrt{2\pi}M_\infty \frac{\sqrt{k\log(kd)}}{\sqrt{n}}\left(1+\frac{\sqrt{\log(d\sqrt{n})+x}}{\sqrt{k\log(kd)}}\right),$$

then, with probability larger than $1-e^{-x}$,

$$\lambda I(\hat{\mathbf{c}}_{n,\lambda}-\mathbf{c}^*_\lambda) \leq \left(3\ell(\mathbf{c}^*_\lambda,\mathbf{c}^*)+\frac{8\lambda^2}{3\kappa_0}\|\mathbf{c}^*_\lambda\|_0\right)\vee\lambda^2. \tag{4.5}$$

Moreover, on the same event, the following prediction result holds

$$\ell(\hat{\mathbf{c}}_{n,\lambda},\mathbf{c}^*) \leq \frac{4\lambda^2\|\mathbf{c}^*\|_0}{\kappa_0}\vee(2\lambda^2). \tag{4.6}$$

Theorem 4.2 can be considered as an application of Theorem 2.1 in [vdG13] to the framework of vector quantization. It is worth noticing that the constant factor $32e\sqrt{2\pi}$ follows from factors in symmetrization arguments and other comparison Theorems. It may probably be improved, optimizing the constants in the derivations. This consistency result shows that, provided that λ is chosen large enough, $\hat{\mathbf{c}}_{n,\lambda}$ converges toward the sparse approximation $\mathbf{c}^*_\lambda$ at a rate smaller than $d\lambda$. This $d\lambda$ rate corresponds to the case where $\mathbf{c}^*_\lambda = \mathbf{c}^*$, and is clearly suboptimal. Consequently much smaller rates are expected. The prediction result provides a distortion rate smaller than $d\lambda^2$. When d is large, this rate is of little interest. However, if a standard k means algorithm is performed on the set S of variable selected by the Lasso k-means procedure, in the spirit of [MM12], then hopefully a distortion rate of $k|S|M_\infty^2/n$ could be attained, compared to the best codebook based on this subset (see, e.g., Theorem 3.1 in Chapter 3)). As announced in Section 4.2, when $X^{(p)}$ has a small range, then the p-th coordinate will be driven to 0 by the Lasso k-means procedure, regardless of its separation capacity. To address this scaling issue, some results are given for a Weighted Lasso k-means procedure in the following Subsection.

4.3.2 Weighted Lasso k-means distortion and consistency

In this Section the penalty function is $\hat{I}_{WL}(\mathbf{c}) = \sum_{p=1}^d \hat{\sigma}_p\sqrt{c_1^{(p)2}+\ldots+c_k^{(p)2}}$. The fact that the weights $\hat{\sigma}_p$ depends on the sample makes the proofs more intricate than in the previous case. To address this issue, this penalty function is connected to a deterministic penalty function, namely

$$I_{WL}(\mathbf{c}) = \sum_{p=1}^d \sigma_p\sqrt{c_1^{(p)2}+\ldots+c_k^{(p)2}}.$$

Denote by T the quantity $\max_{p=1,\dots,d}\frac{M_p}{\sigma_p}$. The following Proposition relates $\hat{I}_{WL}$ to I_{WL}.

Proposition 4.4. *Suppose that* $2n > T^2\sqrt{\log(d)}$. *For every* $y < \log(d)\left(\frac{2n}{T^2\sqrt{\log(d)}} - 1\right)^2$, *we have, with probability larger than* $1 - e^{-y}$, *for all* $\mathbf{c}$ *in* C^k,

$$\sqrt{1-\alpha(y)}I_{WL}(\mathbf{c}) \leq \hat{I}_{WL}(\mathbf{c}) \leq \sqrt{1+\alpha(y)}I_{WL}(\mathbf{c}), \tag{4.7}$$

where $\alpha(y) = \frac{T^2\sqrt{\log(d)}}{\sqrt{2n}}\left(1 + \sqrt{\frac{y}{\log(d)}}\right)$.

The proof of Proposition 4.4 is given in Section 4.5. Proposition 4.4 ensures that, provided that enough sample points are at disposal to correctly estimates the coordinate-wise variances, the data-driven penalty function $\hat{I}_{WL}(\mathbf{c})$ should be close to the deterministic penalty function $I_{WL}(\mathbf{c})$. Equipped with this Proposition, some results can be derived for the k-means procedure with penalty $I_{WL}(\mathbf{c})$ which can be related to results for the Weighted Lasso k-means procedure we propose. This is the idea motivating the following results.

Theorem 4.3. *Let* T *denote* $\max_{p=1,\dots,d}\frac{M_p}{\sigma_p}$. *Suppose that* $2n > T^2\sqrt{\log(d)}$, *and let* $x > 0$. *Choose*

$$y < \log(d)\left(\frac{2n}{T^2\sqrt{\log(d)}} - 1\right)^2.$$

Suppose that

$$\lambda \geq \frac{1}{\sqrt{1-\alpha(y)}}\frac{6kM_\infty\sqrt{2\log(d)}}{\sqrt{n}}\left(1 + \frac{1}{2k}\sqrt{\frac{x}{\log(d)}}\right),$$

where $\alpha(y)$ *is defined in Proposition 4.4. Then, for every* $\varepsilon > 0$, *with probability larger than* $1 - e^{-y} - \left(\frac{\sqrt{k}\sigma^2 T}{\varepsilon} + 1\right)e^{-x}$, *we have*

$$\ell(\hat{\mathbf{c}}_{n,\lambda}, \mathbf{c}^*) \leq \inf_{r>0}\inf_{I_{WL}(\mathbf{c})\leq r}\left(\ell(\mathbf{c}, \mathbf{c}^*) + \lambda(2r+4)\varepsilon\right).$$

As for the Lasso k-means case, Proposition 4.3 proves that the Weighted Lasso k-means codebook performs well in distortion compared to optimal codebooks over L_1-balls. As in Proposition 4.1, it is worth mentioning that Proposition 4.3 is valid even when P does not satisfy (4.2). The proof of Proposition 4.3 is postponed to Section 4.5.

For any codebook $\mathbf{c}$, let $S(\mathbf{c})$ be defined as the set of coordinates p such that $(c_1^{(p)}, \dots, c_k^{(p)}) \neq (0, \dots, 0)$. As done in the previous Section, let $\mathbf{c}_\lambda^*$ be defined as the sparse approximation of $\mathbf{c}^*$ at order λ, by

$$\mathbf{c}_\lambda^* = \arg\min_{\mathbf{c}\in C^k} 3\ell(\mathbf{c}, \mathbf{c}^*) + \frac{8(1+\alpha)\lambda^2\sigma^2_{S(\mathbf{c})}}{\kappa_0},$$

where α is a parameter which will be chosen as $\alpha(y)$, for some $y > 0$. The non-zero coordinates of $\mathbf{c}_\lambda^*$ may be characterized in the following way.

Proposition 4.5. *Let* p *be in* $\{1, \dots, d\}$. *If*

$$1 - \frac{R_p^*}{\sigma_p^2} < \frac{8(1+\alpha)\lambda^2}{3\kappa_0},$$

then

$$\mathbf{c}_\lambda^{*(p)} = (0, \dots, 0).$$

It is worth mentioning that the thresholds takes into account only ratios of the type k-quantization error over variances, avoiding scaling issues. Equipped with this sparse approximation of $\mathbf{c}^*$, we are now in position to state the consistency and prediction results for the Weighted Lasso k-means procedure.

Theorem 4.4. *Suppose that* $2n > T^2\sqrt{\log(d)}$. *Choose*

$$y < \log(d)\left(\frac{2n}{T^2\sqrt{\log(d)}} - 1\right)^2,$$

and $x > 0$. *If*

$$\lambda \geq \frac{1}{\sqrt{1-\alpha(y)}} 32e\sqrt{2\pi}T\sqrt{\frac{k\log(kd)}{n}}\left(1+\sqrt{\frac{\log(\sigma^2\sqrt{n})+x}{k\log(kd)}}\right),$$

where $\alpha(y)$ *is defined in Proposition 4.4, then, with probability larger than* $1-e^{-x}-e^{-y}$, *we have*

$$\sqrt{1-\alpha(y)}\lambda I_{WL}(\hat{\mathbf{c}}_{n,\lambda}-\mathbf{c}^*_\lambda) \leq \left[3\ell(\mathbf{c}^*_\lambda,\mathbf{c}^*)+\frac{8(1+\alpha(y))\lambda^2\sigma^2_{S(\mathbf{c}^*_\lambda)}}{\kappa_0}\right] \vee \left[(1-\alpha(y))\lambda^2\right]. \tag{4.8}$$

Furthermore, on the same event, the following prediction result holds

$$\ell(\hat{\mathbf{c}}_{n,\lambda},\mathbf{c}^*) \leq \left[\frac{8(1+\alpha(y))\sigma^2_{S(\mathbf{c}^*)}\lambda^2}{\kappa_0}\right] \vee \left[\sqrt{1-\alpha(y)}\lambda^2\right]. \tag{4.9}$$

As for the Lasso k-means case, Theorem 4.4 ensures that $\hat{\mathbf{c}}_{n,\lambda}$ is close to its sparse approximation, in the sense of I_{WL}, with a rate possibly much smaller than $\lambda\sigma^2$. This rate corresponds to the case where the sparse approximation of $\mathbf{c}^*$ is $\mathbf{c}^*$. This leads to expect much smaller rates for the deviation between $\hat{\mathbf{c}}_{n,\lambda}$ and $\mathbf{c}^*_\lambda$. However, the prediction result is much more interesting, since it guarantees a distortion rate of $\sigma^2\lambda^2$ for the Weighted Lasso k-means procedure. As mentioned below Theorem 4.2, it is likely that this distortion rate could be improved by performing a standard k-means procedure on the set S of selected variables, possibly leading to a distortion rate of $k\sigma^2_S T^2_S/n$ (see, e.g., Theorem 3.1 in Chapter 3), compared to the optimal codebook with support S.

4.4 Simulations

This Section illustrates the theoretical results obtained in Section 4.3 with numerical experiments. For this purpose, the algorithm proposed in [SWF12] has been implemented with the R software, and applied on simulated data drawn from Gaussian mixture distributions with varying dimensions, up to 300. Special attention has been paid to the set of selected variables for both Lasso k-kmeans and Weighted Lasso k-means procedures, introduced in Section 4.2.

4.4.1 Algorithm

The algorithm implemented to perform the Lasso k-means procedure is inspired from the *Regularized k-means clustering*, introduced in [SWF12], and is described below.

Definition 4.2 (Lasso k-means algorithm)**.**

Initialization *Choose* $(c_1,\dots,c_k)$ *randomly among the data set (or as the result of a standard* k*-means procedure).*
Allocation Step *Build the allocation matrix* $\hat{L}$ *defined by* $\hat{L}_{i,j} = \mathbb{1}_{X_i \in V_j(\mathbf{c})}$.
Minimization Step *For every coordinate* $p = 1,\dots,d$*, use a Newton algorithm to mimimize*

$$\frac{1}{n}\left\|X^{(p)} - \hat{L}\mathbf{c}^{(p)}\right\|^2 + \lambda w(p)\sqrt{c_1^{(p)2} + \dots + c_k^{(p)2}},$$

where $w(p) = 1$ *or* $w(p) = \hat{\sigma}_p$*, and refresh* $\mathbf{c}$.
→ ***Repeat*** *the Allocation / Minimization steps until stability and store the resulting codebook.*
→ ***Repeat*** *the latter subprocedure with varying initialization steps, and pick up the codebook with the smallest penalized risk* $P_n\gamma(\mathbf{c},.) + \lambda I(\mathbf{c})$.

When the dimension d is large, it turns out that the number of initialization steps is a crucial parameter, to avoid local minimums of the penalized empirical risk. Based on our limited numerical experience, when $d = 3000$, no less than 50 initialization steps are needed to obtain a meaningful codebook. It may also be noted that the Newton minimization step slows down the algorithm, compared to the original Lloyd's algorithm (see, e.g., [Llo82]). This issue can partially be addressed by choosing as initialization the result of a standard k-means algorithm, when the penalization parameter λ is small.

4.4.2 Model and theoretical predictions

In this Subsection the two models of Gaussian mixtures from which the simulated data are generated is introduced. For both of these models the size k of codebooks will match the number of components of the mixture, namely $k = 4$.

4.4.2.1 Model 1

Let $d \geq 30$. It is assumed that P is a Gaussian mixture distribution, with means

$$\begin{cases} \mu_1 &= (3, 2.9, 2.8, \dots, 0.1, \underbrace{0,\dots,0}_{d-30}), \\ \mu_2 &= 0, \\ \mu_3 &= -\mu_1, \\ \mu_4 &= (3, -2.9, 2.8, \dots, -0.1, \underbrace{0,\dots,0}_{d-30}), \end{cases}$$

and with weights defined by $\theta_1 = 0.3$, $\theta_2 = 0.2$, $\theta_3 = 0.2$ and $\theta_4 = 0.3$. Furthermore it is assumed that every component of the mixture has the same diagonal covariance matrix, namely $\Sigma_1 = I_d$. It may also be noted that, for $p = 1,\dots,30$,

$$|\mu_1^{(p)}| = |\mu_3^{(p)}| = |\mu_4^{(p)}| := m_p.$$

Throughout this Chapter, the variables p such that $1 \leq p \leq 30$ will be named *active* variables, whereas variables such that $p \geq 31$ will be named *non active* variables. Simulations have been carried out with dimensions $d = 30$ and $d = 300$. For this model both Lasso k-means and Weighted Lasso k-means have been implemented. Some theoretical guarantees on the support of the sparse approximation of $\mathbf{c}^*$ at order λ for the Lasso k-means, as defined in Section 4.3, are given below.

Proposition 4.6. *For the Lasso k-means procedure, for any $\lambda \geq 0$,*

– *If $1 \leq p \leq 30$, and*

$$m_p^2 < \frac{10\lambda^2}{3\kappa_0} + 1 - \frac{1}{k^2},$$

then $\mathbf{c}_\lambda^{*(p)} = 0$.

– *If $p \geq 31$, and*

$$\frac{8\lambda^2}{3\kappa_0} > 1 - \frac{1}{k^2},$$

then $\mathbf{c}_\lambda^{*(p)} = 0$.

Proposition 4.6 is an application of Proposition 4.3 to Model 1. A short proof is to be found in Section 4.5.9. Assuming that the conditions exposed in Proposition 4.6 are also necessary conditions, the support of the empirical codebook $\hat{\mathbf{c}}_{n,\lambda}$ is expected to behave as follows. When $\lambda = 0$, all the coordinates are selected. Then, all the non active variables $p \geq 31$ will be driven to 0 when λ reaches a threshold, and, at last, knowing that m_p decreases linearly with p, the number of selected active variables $p \leq 30$ will decrease linearly as λ grows. For the Weighted Lasso k-means procedure, the set of variables driven to 0 may be characterized by the following Proposition 4.7.

Proposition 4.7. *For the Weighted Lasso k-means procedure, for any $\lambda \geq 0$ and α defined as in Theorem 4.3,*

– *If $1 \leq p \leq 30$, and*

$$\frac{8(1+\alpha)\lambda^2}{3\kappa_0} > 1 - \frac{1}{k^2(1+0.8m_p^2)},$$

then $\mathbf{c}_\lambda^{*(p)} = 0$.

– *If $p \geq 31$, and*

$$\frac{8(1+\alpha)\lambda^2}{3\kappa_0} > 1 - \frac{1}{k^2},$$

then $\mathbf{c}_\lambda^{*(p)} = 0$.

A short proof of Proposition 4.7 can be found in Section 4.5.9. Assuming that the conditions mentioned in Proposition 4.7 are also necessary, the set of selected variables should behave as follows. When λ reaches a certain threshold, all the non active variables ($d \geq 31$) should be driven to 0. Next the number of selected active variables ($d \leq 30$) should reach 0 quicker than for the Lasso k-means procedure.

4.4.2.2 Model 2

In Model 2 it is assumed that P is a Gaussian mixture distribution, with the same means and weights as in Model 1, and dimension $d \geq 30$. However, the variable-wise variances are now assumed to be scaled with the variable means. In other words, every component has the same diagonal covariance matrix $\Sigma_{2,p}$, with diagonal coefficients s_p^2 defined by

$$\begin{cases} s_p &= 1 - \frac{p-1}{30} & \text{if} \quad 1 \leq p \leq 30, \\ s_p &= 1 & \text{if} \quad p \geq 31. \end{cases}$$

Hence the marginal distribution $P^{(p)}$ of the p-th variable may be thought of as a contraction of the marginal distribution of the first variable $P^{(1)}$, for $p = 1, \ldots, 30$. For this model, simulations of both Lasso k-means and Weighted Lasso k-means have been carried out, with dimension $d = 30$ and $d = 300$. The theoretical results on the support of the best sparse approximation for the Lasso k-means strategy are gathered in the following Proposition.

Proposition 4.8. *For the Lasso k-means procedure, for any $\lambda \geq 0$,*

- *If $1 \leq p \leq 30$, and*

$$m_p^2 < \frac{120\lambda^2}{\kappa_0(41 - 5/k^2)},$$

  then $\mathbf{c}_\lambda^{*(p)} = 0$.

- *If $p \geq 31$, and*

$$\frac{8\lambda^2}{3\kappa_0} > 1 - \frac{1}{k^2},$$

  then $\mathbf{c}_\lambda^{*(p)} = 0$.

Proposition 4.8 follows from Proposition 4.3, adapted to Gaussian mixture models. A proof is given in Section 4.5.9. It is interesting to note that, assuming Proposition 4.8 is also a necessary condition for a variable to be driven to 0, the threshold for inactive variables ($p \geq 31$) may be larger than the threshold for active variables, for m_p small enough. To be more precise, let T_i denote the quantity in Proposition 4.8 such that $\lambda > T_i$ implies $\mathbf{c}_\lambda^{*(p)} = 0$ for $p \geq 31$, and $T_a(p)$ such that $\lambda > T_a(p)$ implies $\mathbf{c}_\lambda^{*(p)} = 0$, for $p \leq 30$. Then easy calculation shows that

$$\forall 25 \leq p \leq 30 \quad T_a(p) < T_i.$$

From this it is easy to deduce that the Lasso k-means strategy may fail to recover the "true" support $\{1, \ldots, 30\}$, since the small range active variables corresponding to $25 \leq p \leq 30$ should be driven out of the selected support before the inactive variables $p \geq 31$. This issue can theoretically be addressed considering the Weighted Lasso k-means strategy, as mentioned in the following Proposition.

Proposition 4.9. *For the Weighted Lasso k-means procedure, for any $\lambda \geq 0$ and α defined as in Theorem 4.3,*

- *If $1 \leq p \leq 30$, and*

$$\frac{8(1+\alpha)\lambda^2}{3\kappa_0} > 1 - \frac{5}{14k^2},$$

  then $\mathbf{c}_\lambda^{*(p)} = 0$.

– *If* $p \geq 31$, *and*

$$\frac{8(1+\alpha)\lambda^2}{3\kappa_0} > 1 - \frac{1}{k^2},$$

then $\mathbf{c}_\lambda^{*(p)} = 0$.

A proof of Proposition 4.9 can be found in Section 4.5.9. Assuming that the conditions given in Proposition 4.9 are also necessary conditions ensures that the true support may be recovered by the Weighted Lasso k-means strategy. In fact, choosing mean squares based weights may address the scaling issue, by providing a common threshold for the active variables ($p \leq 30$), larger than the common threshold for the inactive variables ($p \geq 31$). As exposed in the following Subsection, most of these theoretical conjectures are confirmed by numerical simulations.

4.4.3 Numerical experiments

In this Subsection the results of the implementation of both Lasso k-means and Weigthed Lasso k-means are exposed, for dimension $d = 30$ and $d = 300$. Furthermore, every simulation is repeated 10 times, except for Figure 4.1, Figure 4.4 and Figure 4.8. At last, the Lasso coefficient λ is chosen as $c \times \log(d)/n$, where c is varying from 0 to at most 15.

4.4.3.1 Model 1 simulations

We begin by illustrating the dependance of the selected variables on the parameter λ, for one iteration of the Lasso k-means algorithm.

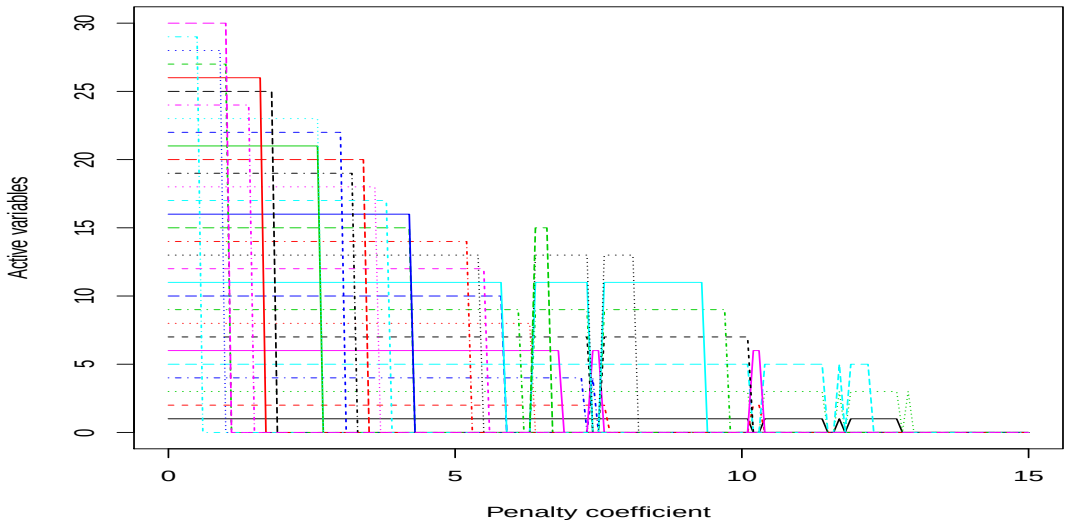

FIGURE 4.1 – For $d = 30$, collection of selected variables for the Lasso k-means algorithm.

Figure 4.1 shows that the variable-wise disappearance thresholds roughly behave as expected by Proposition 4.6 for active variables ($1 \leq p \leq 30$), that is proportional to the penalty coefficient c, at least for $c \leq 5$. For $c \geq 5$, the behavior of the selected variables seems more chaotic. In fact, based on our experience, the Lasso

k-means algorithm is more likely to be trapped into local minima when c belongs to a certain interval away from 0. This leads to much more fluctuations of the selected set of variables, as illustrated by the following Figure 4.2.

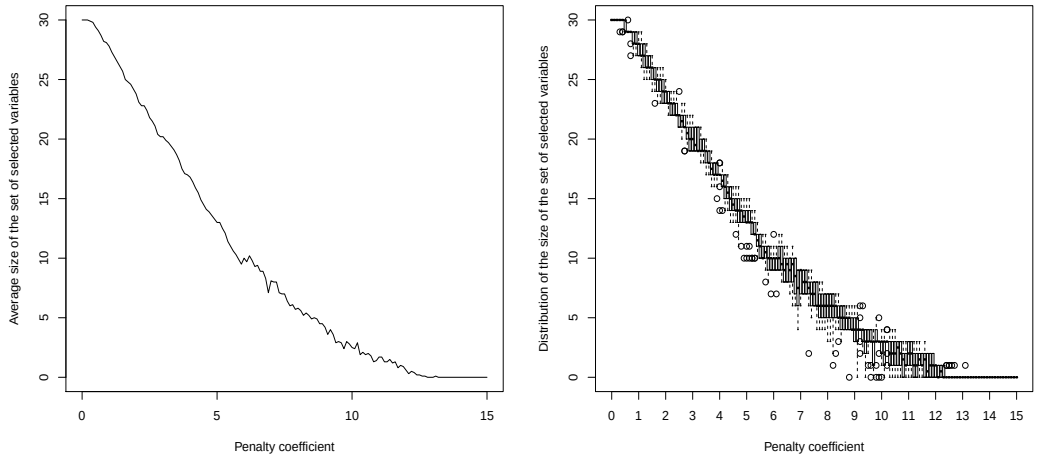

FIGURE 4.2 – For $d = 30$, size of the set of selected variables for the Lasso k-means algorithm, with 10 iterations.

The fact that the selected set of variable shows more variations for intermediate c's may be explained as follows. For small c's, the Lasso k-means empirical codebook is not too far of the regular k-means empirical codebook, which is a starting point of our algorithm. For large c's, only few variable remains active, which render the number of possible local minima small.

In the case where $d \geq 31$, some non active variables are present in Model 1. The following Figure 4.3 shows that the Lasso k-means procedure quickly drives these variables to 0.

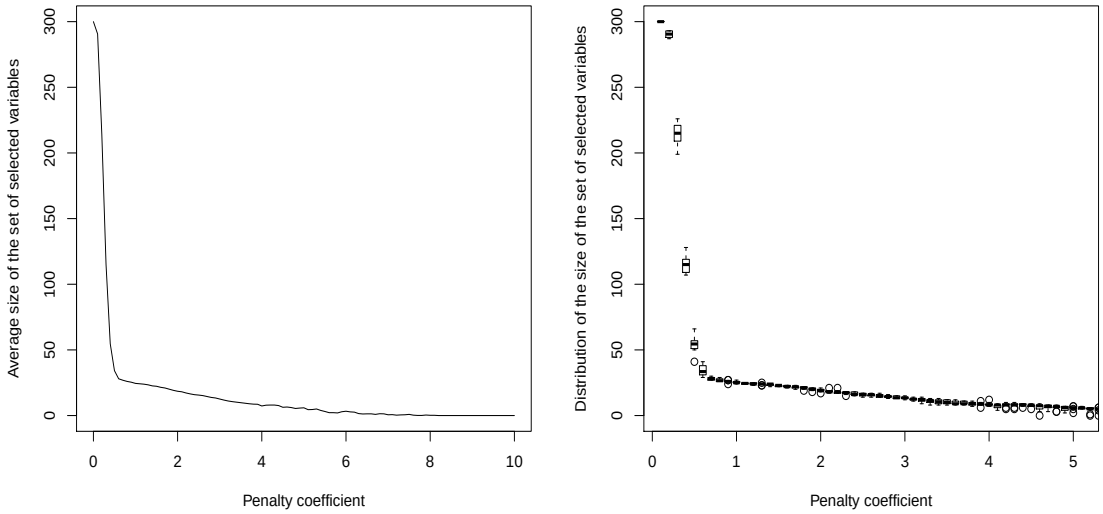

FIGURE 4.3 – For $d = 300$, size of the set of selected variables for the Lasso k-means algorithm, with 10 iterations.

Figure 4.3 confirms the prediction of Proposition 4.6, that is non active variables are driven to 0 almost together, whereas the set of active variables decreases linearly with the penalty coefficient.

For the Weighted Lasso k-means procedure, Figure 4.4 below shows that the set of selected active variables is likely to have more fluctuations than for the Lasso k-means case.

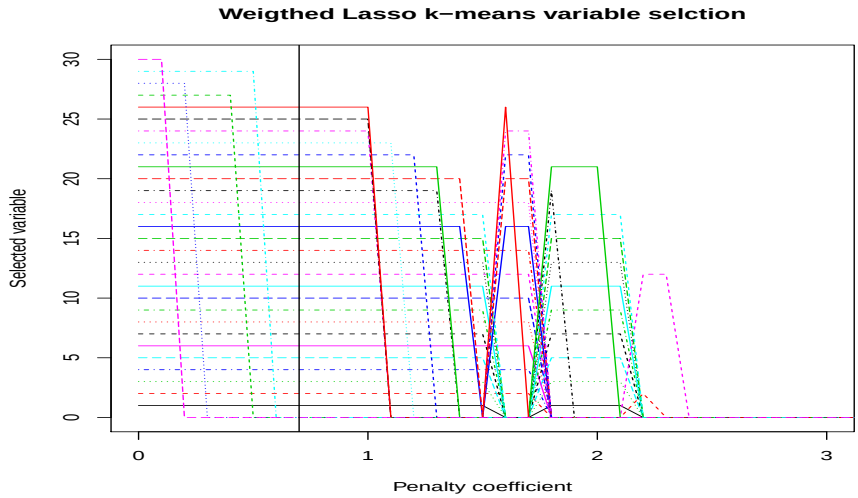

FIGURE 4.4 – For $d = 30$, collection of selected variables by the Weighted Lasso k-means algorithm.

However, this set of selected active variables roughly behaves as expected by Proposition 4.7 : the first variables are the last to disappear from the support, and the set of selected variables decreases faster than in the Lasso k-means case, as illustrated by Figure 4.5.

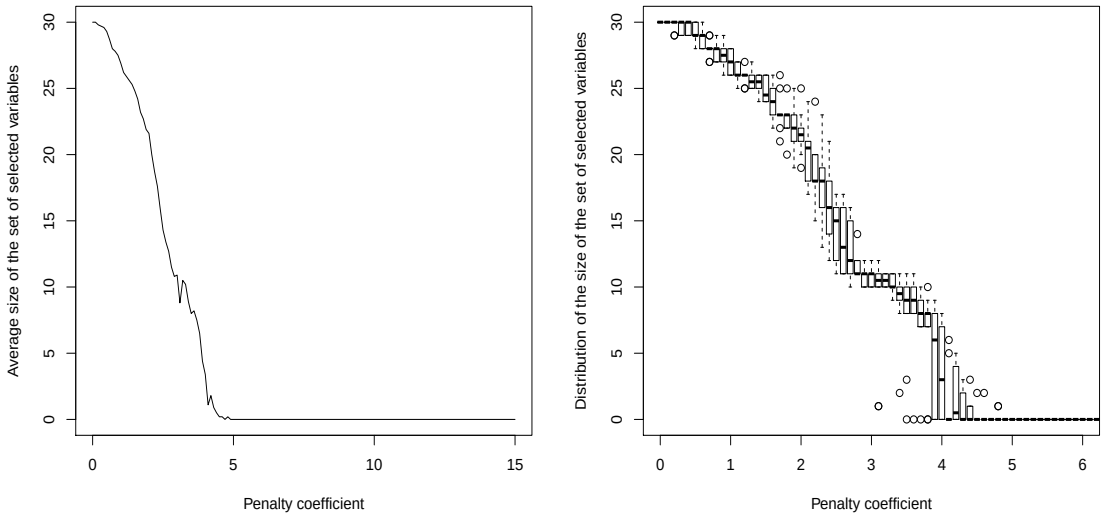

FIGURE 4.5 – For $d = 30$, size of the set of selected variables for the Weighted Lasso k-means algorithm, with 10 iterations.

4.4.3.2 Model 2 simulations

For Model 2, simulations have been carried out with $d = 300$ for the Lasso k-means and Weighted Lasso k-means procedures. Figures 4.6 and 4.7 below describe the mean size of the set of selected variables over 10 iterations.

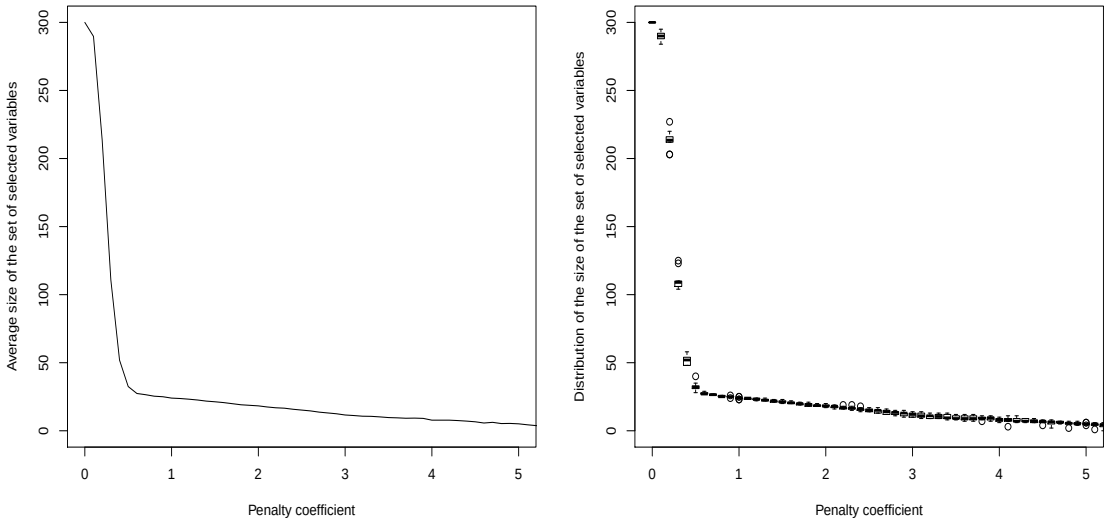

FIGURE 4.6 – For $d = 300$, mean size of the set of selected variables for the Lasso k-means algorithm, over 10 iterations.

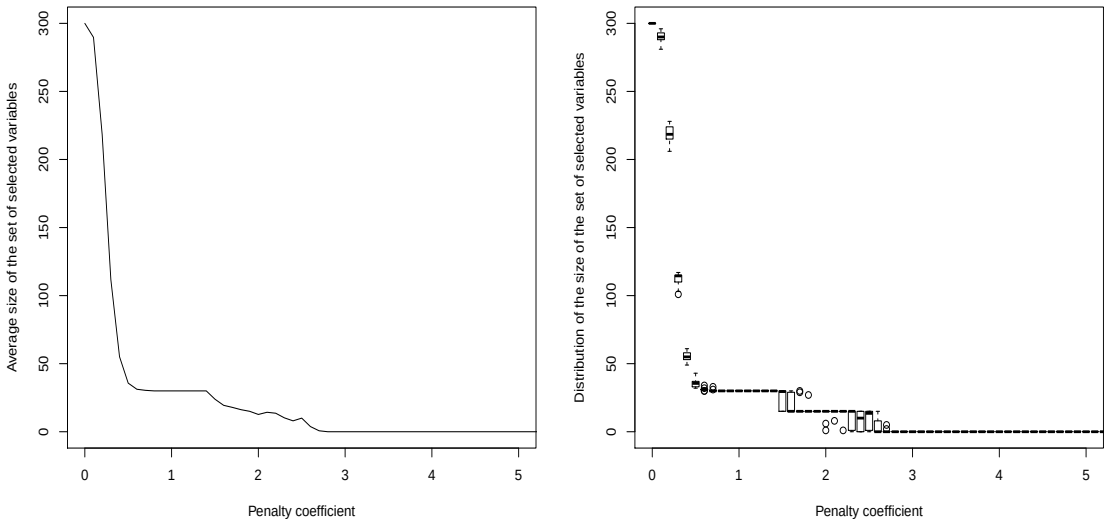

FIGURE 4.7 – For $d = 300$, mean size of the set of selected variables for the Lasso k-means algorithm, over 10 iterations.

These figures overall confirm the theoretical predictions of Proposition 4.8 and 4.9. For the Lasso k-means strategy the inactive variables are driven out of the support once c reaches a certain threshold, and the active variables are driven to 0 linearly with respect to c. For the Weighted Lasso k-means strategy, Figure 4.7

shows that the inactive variables are not selected when c reaches a threshold, and that for a range of intermediate c's the set of selected variables remains constant (in fact the set of selected variables is exactly the set of active variables in this case, as shown below). However, it is difficult to deduce that there exists a common threshold for active variable selection from Figure 4.7. This may follow from the fluctuations of the support size when c is located in this area, as shown by the boxplot diagram. These fluctuations are likely to be caused in part by the small number of performed initializations.

According to Proposition 4.8 and Proposition 4.9, it may be interesting to focus on the 6 last active variables, that is $25 \leq p \leq 30$.

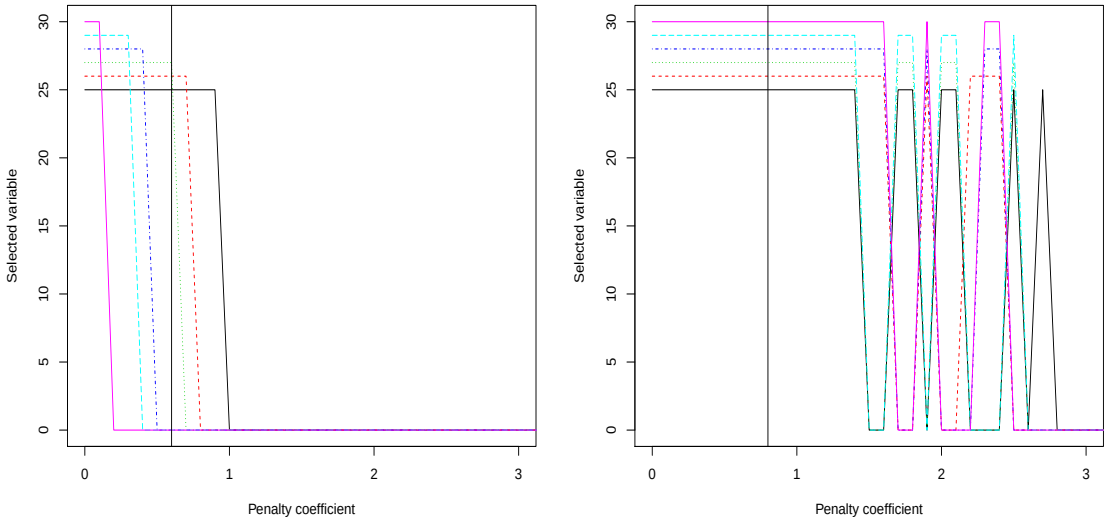

FIGURE 4.8 – For $d = 300$, the left-hand Figure and right-hand Figure respectively represent the occurence of the 6 last active variables in the support for the Lasso k-means and Weighted Lasso k-means procedures. The vertical line represents the threshold beyond which no inactive variable is selected.

Figure 4.8 partially confirms the theoretical predictions, but for the last four active variables only. For the Lasso k-means strategy, variables 27, 28, 29 and 30 are not selected without inactive variables being selected too. This shows that the "true" support is never selected in Model 2 by the Lasso k-means strategy. On the opposite, the Weigthed Lasso k-means procedure succeed in recovering the true support, for a range of penalty coefficients.

4.5 Proofs

In this Section the results are derived for a general penalty function

$$I_w(\mathbf{c}) = \sum_{p=1}^{d} w_p \sqrt{c_1^{(p)2} + \ldots + c_k^{(p)2}},$$

for any positive sequence $(w_p)_{p=1,\ldots,d}$. In the Lasso case, $w_p = 1$, whereas in the Weighted Lasso case $w_p = \hat{\sigma}_p$.

4.5.1 Proof of Proposition 4.1 and Proposition 4.2

Let $V_1, \dots, V_k$ be a Voronoi partition associated with $\hat{\mathbf{c}}_{n,\lambda}$, and let $\hat{L}$ be the matrix of assignments, defined by

$$\hat{L}_{i,j} = \mathbb{1}_{X_i \in V_j}.$$

Suppose that $\hat{\mathbf{c}}_{n,\lambda}^{(p)} \neq 0$, where through this Subsection $\hat{\mathbf{c}}_{n,\lambda}^{(p)}$ will denote the column vector $(\hat{c}_{n,\lambda,1}^{(p)}, \dots, \hat{c}_{n,\lambda,k}^{(p)})^t$, and denote by $\hat{X}^{(p)}$ the column vector $(X_1^{(p)}, \dots, X_k^{(p)})^t$. Then the Karush-Kuhn-Tucker condition, for the empirical risk strategy penalized with $I_w(\hat{\mathbf{c}}_{n,\lambda})$, implies that (see, e.g., the proof of Theorem 2 in [SWF12])

$$\frac{-2}{\sqrt{n}} \hat{L}^t \left(\hat{X}^{(p)} - \hat{L} \hat{\mathbf{c}}_{n,\lambda}^{(p)} \right) + \sqrt{n} \lambda \frac{w_p \hat{\mathbf{c}}_{n,\lambda}^{(p)}}{\| \hat{\mathbf{c}}_{n,\lambda}^{(p)} \|} = 0. \tag{4.10}$$

Since $\hat{L}^t \left(\hat{X}^{(p)} - \hat{L} \hat{\mathbf{c}}_{n,\lambda}^{(p)} \right)$ is the following vector of size k

$$\left(\sum_{X_i \in V_1} (X_i^{(p)} - \hat{c}_{n,\lambda,1}^{(p)}), \dots, \sum_{X_i \in V_k} (X_i^{(p)} - \hat{c}_{n,\lambda,k}^{(p)}) \right),$$

it may be noted that

$$\left\| \hat{L}^t \left(\hat{X}^{(p)} - \hat{L} \hat{\mathbf{c}}_{n,\lambda}^{(p)} \right) \right\|^2 = \sum_{j=1}^{k} n_j^2 (\bar{c}_j^{(p)} - \hat{c}_{n,\lambda,j}^{(p)})^2,$$

where n_j denote the number of sample vector X_i's in V_j, and $\bar{c}_j$ denote the empirical mean of the sample over the set V_j, that is $\bar{c}_j = \frac{1}{n_j} \sum_{X_i \in V_j} X_i$. Denote by $\hat{p}_j$ the empirical weight of V_j, that is $\hat{p}_j = n_j / n$, then

$$\frac{1}{n^2} \left\| \hat{L}^t \left(\hat{X}^{(p)} - \hat{L} \hat{\mathbf{c}}_{n,\lambda}^{(p)} \right) \right\|^2 \leq \sum_{j=1}^{k} \hat{p}_j (\bar{c}_j^{(p)} - \hat{c}_{n,\lambda,j}^{(p)})^2,$$

where $\hat{p}_j \leq 1$ has been used. Let Q_1 be the quantizer which maps V_j to $\bar{c}_j$, then it is easy to see that

$$\sum_{j=1}^{k} \hat{p}_j (\bar{c}_j^{(p)} - \hat{c}_{n,\lambda,j}^{(p)})^2 = \hat{R}_p(\hat{\mathbf{c}}_{n,\lambda}) - \hat{R}_p(Q_1).$$

Since $\hat{R}_p(\hat{\mathbf{c}}_{n,\lambda}) - \hat{R}_p(Q_1) \leq \hat{\sigma}_p^2 - \hat{R}_p$, (4.10) ensures that

$$\frac{\lambda w_p}{2} \leq \sqrt{\hat{\sigma}_p^2 - \hat{R}_p}.$$

Taking $w_p = 1$ gives the result of Proposition 4.1 and $w_p = \hat{\sigma}_p$ gives the result of Proposition 4.2.

4.5.2 Proof of Proposition 4.4

Hoeffding's inequality ensures that, for every $p = 1, \dots, d$, $\frac{\hat{\sigma}_p^2}{\sigma_p^2} - 1$ is a subgaussian random variable with variance bounded by $\frac{T^4}{4n}$. For a comprehensive introduction to subgaussian random variables and its application to empirical processes theory,

the interested reader is referred to [Mas07]. Applying Theorem 3.12 in [Mas07] and a bounded difference concentration inequality (see, e.g., Theorem 5.1 in [Mas07]) yields, with probability larger than $1-e^{-y}$,

$$\max_{p=1,\dots,d}\left|\frac{\hat{\sigma}_p^2}{\sigma_p^2}-1\right| \leq \frac{T^2\sqrt{\log(d)}}{\sqrt{2n}}\left(1+\sqrt{\frac{y}{\log(d)}}\right).$$

Taking into account that $2n > T^2\sqrt{\log(d)}$ and $y < \log(d)\left(\frac{2n}{T^2\sqrt{\log(d)}}-1\right)^2$ leads to the result.

4.5.3 Proof of Proposition 4.3

Let S be a subset of $\{1,\dots,d\}$, and let p be in S such that

$$\sigma_p^2 - R_p^* < \frac{8\lambda^2}{3\kappa_0}.$$

Denote by $\mathbf{c}_S^*$ an optimal codebook with support S, that is

$$\mathbf{c}_S^* = \arg\min_{S(\mathbf{c})=S} R(\mathbf{c}).$$

Then, according to (4.1), we may write

$$\begin{aligned} R(\mathbf{c}_{S\setminus\{p\}}^*) - R(\mathbf{c}_S^*) &\leq R_{S\setminus\{p\}}^* + \sigma_{(S\setminus\{p\})^c}^2 - (R_{S\setminus\{p\}}^* + R_p^*) - \sigma_{S^c}^2 \\ &\leq \sigma_p^2 - R_p^*. \end{aligned}$$

Therefore

$$3R(\mathbf{c}_{S\setminus\{p\}}^*) + \frac{8\lambda^2}{\kappa_0}\|\mathbf{c}_{S\setminus\{p\}}^*\|_0 < 3R(\mathbf{c}_S^*) + \frac{8\lambda^2}{\kappa_0}\|\mathbf{c}_S^*\|_0.$$

4.5.4 Proof of Proposition 4.5

Adopting the notation of the previous Subsection, let p be in S such that $1-\frac{R_p^*}{\sigma_p^2} < \frac{8(1+\alpha)\lambda^2}{\kappa_0}$. Then, it can be derived the same way as in the previous Subsection that

$$R(\mathbf{c}_{S\setminus\{p\}}^*) - R(\mathbf{c}_S^*) \leq \sigma_p^2 - R_p^*.$$

This leads to

$$3R(\mathbf{c}_{S\setminus\{p\}}^*) + \frac{8\lambda^2(1+\alpha)}{\kappa_0}\sigma_{S\setminus\{p\}}^2 < 3R(\mathbf{c}_S^*) + \frac{8\lambda^2(1+\alpha)}{\kappa_0}\sigma_S^2.$$

4.5.5 Proof of Theorem 4.1

As in the proof of Proposition 4.1, throughout this Subsection, the penalty function is chosen as $I_w(\mathbf{c})$, for a sequence of weights w. Let $T(w)$ denote the quantity

$$T(w) = \max_{p=1,\dots,d}\frac{M_p}{w_p}.$$

Then $T(w) = M_\infty$ in the Lasso case and $T(w) = T$ in the Weighted Lasso case. Let also $\bar{M}(w)$ be defined as $\sqrt{k}\|w\|^2 T(w)$. It is immediate that, for every $\mathbf{c}$ in C^k, $I_w(\mathbf{c}) \leq \bar{M}(w)$.

Let $\bar{\gamma}$ be defined by

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

for every $\mathbf{c}$ in C^k and x in $\mathbb{R}^d$. The following Proposition, inspired from Theorem 2.1 in [BDL08], offers an upper bound on the deviations between P_n and P on the set of possible $\bar{\gamma}$ constrained by $I_w(\mathbf{c})$.

Proposition 4.10. *Suppose that w is deterministic. Let $x > 0$. Then, with probability larger than $1 - e^{-x}$, we have*

$$\sup_{I_w(\mathbf{c}) \leq r} (P - P_n)\bar{\gamma}(\mathbf{c}, .) \leq r \frac{6kT(w)\sqrt{2\log(d)}}{\sqrt{n}} \left(1 + \frac{1}{2k}\sqrt{\frac{x}{\log(d)}}\right).$$

It is worth mentioning that the requirements that w is deterministic prevents from directly choosing $w_p = \hat{\sigma}_p$. This issue will be addressed in the following Subsection. Now choose $\lambda \geq \frac{6kT(w)\sqrt{2\log(d)}}{\sqrt{n}} \left(1 + \frac{1}{2k}\sqrt{\frac{x}{\log(d)}}\right)$, and let $\varepsilon > 0$. Define $K(\varepsilon) = \left\lceil \frac{\bar{M}(w)}{\varepsilon} \right\rceil$, that is the smallest integer larger than $\frac{\bar{M}(w)}{\varepsilon}$, and $\hat{m} = \left\lceil \frac{I_w(\hat{\mathbf{c}}_{n,\lambda})}{\varepsilon} \right\rceil$. Then, applying a union bound to Proposition 4.10, it follows that, with probability larger than $1 - K(\varepsilon)e^{-x}$, for all $m = 1, \dots, K(\varepsilon)$,

$$\sup_{I_w(\mathbf{c}) \leq m\varepsilon} (P - P_n)\bar{\gamma}(\mathbf{c}, .) \leq m\varepsilon \frac{6kT(w)\sqrt{2\log(d)}}{\sqrt{n}} \left(1 + \frac{1}{2k}\sqrt{\frac{x}{\log(d)}}\right).$$

On this event, we have

$$\begin{aligned} P_n\bar{\gamma}(\hat{\mathbf{c}}_{n,\lambda}, .) + \lambda I_w(\hat{\mathbf{c}}_{n,\lambda}) &\leq \inf_{r>0} \inf_{I_w(\mathbf{c}) \leq r} \left(P_n\bar{\gamma}(\mathbf{c}, .) + \lambda r\right) \\ &\leq \inf_{m=1,\dots,K(\varepsilon)} \inf_{I_w(\mathbf{c}) \leq m\varepsilon} \left(P_n\bar{\gamma}(\mathbf{c}, .) + \lambda m\varepsilon\right). \end{aligned}$$

It follows that

$$\begin{aligned} P\bar{\gamma}(\hat{\mathbf{c}}_{n,\lambda}, .) &\leq P_n\bar{\gamma}(\hat{\mathbf{c}}_{n,\lambda}, .) + \lambda\hat{m}\varepsilon \\ &\leq P_n\bar{\gamma}(\hat{\mathbf{c}}_{n,\lambda}, .) + \lambda I_w(\hat{\mathbf{c}}_{n,\lambda}) + \lambda\varepsilon \\ &\leq \inf_{m=1,\dots,K(\varepsilon)} \inf_{I_w(\mathbf{c}) \leq m\varepsilon} \left(P\bar{\gamma}(\mathbf{c}, .) + \lambda(2m+1)\varepsilon\right). \end{aligned}$$

Adding $-P\bar{\gamma}(\mathbf{c}^*, .)$ on both sides leads to

$$\begin{aligned} \ell(\hat{\mathbf{c}}_{n,\lambda}, \mathbf{c}^*) &\leq \inf_{m=1,\dots,K(\varepsilon)} \inf_{I_w(\mathbf{c}) \leq m\varepsilon} \left(\ell(\mathbf{c}, \mathbf{c}^*) + \lambda(2m+1)\varepsilon\right) \\ &\leq \inf_{r>0} \inf_{I_w(\mathbf{c}) \leq r} \left(\ell(\mathbf{c}, \mathbf{c}^*) + \lambda(2m+3)\varepsilon\right). \end{aligned}$$

Choosing $w_p = 1$ concludes the proof for the Lasso k-means procedure.

4.5.6 Proof of Theorem 4.3

The proof of Theorem 4.3 is almost the same as the proof of Theorem 4.1, with weights $w_p = \sigma_p$, leading to $T(w) = T$. To avoid confusion, $I_{WL}(\mathbf{c})$ will denote $I_w(\mathbf{c})$ with weights $w_p = \sigma_p$, and $\hat{I}_{WL}(\mathbf{c})$ will denote $I_w(\mathbf{c})$ with weights $w_p = \hat{\sigma}_p$. Let λ be larger than $\frac{1}{\sqrt{1-\alpha(y)}} \frac{6kM_\infty\sqrt{2\log(d)}}{\sqrt{n}} \left(1 + \frac{1}{2k}\sqrt{\frac{x}{\log(d)}}\right)$, then, with probability larger than $1 - e^{-y} - \left(\frac{\sqrt{k}dT}{\varepsilon} + 1\right)e^{-x}$ we have, for every $\mathbf{c}$ in C^k,

$$P_n\bar{\gamma}(\hat{\mathbf{c}}_{n,\lambda},.) + \lambda\hat{I}_{WL} \leq P_n\bar{\gamma}(\mathbf{c},.) + \sqrt{1+\alpha(y)}\lambda I_{WL}(\mathbf{c}).$$

It follows that

$$\begin{aligned}
&P_n\bar{\gamma}(\hat{\mathbf{c}}_{n,\lambda},.) + \lambda\hat{I}_{WL}(\hat{\mathbf{c}}_{n,\lambda}) \\
&\leq \inf_{m=1,\ldots,K(\varepsilon)} \inf_{I_{WL}(\mathbf{c})\leq m\varepsilon} P_n\bar{\gamma}(\mathbf{c},.) + \sqrt{1+\alpha(y)}m\varepsilon \\
&\leq \inf_{m=1,\ldots,K(\varepsilon)} \inf_{I_{WL}(\mathbf{c})\leq m\varepsilon} P\bar{\gamma}(\mathbf{c},.) + (\sqrt{1+\alpha(y)}m\varepsilon + \sqrt{1-\alpha(y)}m\varepsilon \\
&\leq \inf_{r>0} \inf_{I_{WL}(\mathbf{c})\leq r} P\bar{\gamma}(\mathbf{c},.) + \sqrt{1+\alpha(y)}(r+\varepsilon) + \sqrt{1-\alpha(y)}(r+\varepsilon),
\end{aligned}$$

where the middle inequality follows from Proposition 4.10. On the other hand, it may be written that

$$\begin{aligned}
P\bar{\gamma}(\hat{\mathbf{c}}_{n,\lambda},.) &\leq P_n\bar{\gamma}(\hat{\mathbf{c}}_{n,\lambda},.) + \sqrt{1-\alpha(y)}\lambda\hat{m}\varepsilon \\
&\leq P_n\bar{\gamma}(\hat{\mathbf{c}}_{n,\lambda},.) + \sqrt{1-\alpha(y)}\lambda\varepsilon + \lambda\hat{I}_{WL}(\hat{\mathbf{c}}_{n,\lambda}),
\end{aligned}$$

according to Proposition 4.4. Combining these two inequalities and taking into account that $\sqrt{1-\alpha(y)} + \sqrt{1+\alpha(y)} \leq 2$ leads to the result.

4.5.7 Proof of Theorem 4.2

As done in the previous Subsection, the results are derived for a generic penalty function

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

The main argument of this proof relies on a comparison between $(P - P_n)(\bar{\gamma}(\mathbf{c},.) - \bar{\gamma}(\mathbf{c}',.))$ and $I_w(\mathbf{c} - \mathbf{c}')$, stated in the following Proposition.

Proposition 4.11. *Let w be a deterministic d-dimensional vector, and let u denote the quantity $\log\left(\frac{\|w\|^2\sqrt{n}}{\sqrt{\log(kd)}}\right)$. If we denote by*

$$\lambda_0 = 16\sqrt{2\pi}\sqrt{\frac{k\log(kd)}{n}}T(w),$$

then, for every $x > 0$, denoting by

$$\lambda_1 = e\lambda_0\left(1 + \sqrt{\frac{u+x}{k\log kd}}\right),$$

we have, for any fixed $\mathbf{c}'$ *in* C^k*, with probability larger than* $1-e^{-x}$*,*

$$\sup_{I_w(\mathbf{c}-\mathbf{c}')\le 2\bar{M}(w)} \frac{\left|(P-P_n)(\gamma(\mathbf{c},.)-\gamma(\mathbf{c}',.))\right|}{I_w(\mathbf{c}-\mathbf{c}')\vee\lambda_0} \le \lambda_1, \tag{4.11}$$

where we recall that $\bar{M}(w)=\sqrt{k}\|w\|^2 T(w)$*.*

The proof of Proposition 4.11 relies on Section 3.4 in [vdG13], and is postponed to the next Section. The consistency result also relies on the following Lemma, which connects the L_1 penalty to the size of the support. For any subset $S\subset\{1,\dots,d\}$ and vector x in $\mathbb{R}^d$, the truncated vector x_S is defined by

$$x_S^{(p)} = x^{(p)}\mathbb{1}_{p\in S}.$$

Moreover, let $S(\mathbf{c})$ denote the support of $\mathbf{c}$, that is the set of coordinates such that $(c_1^{(p)},\dots,c_k^{(p)})\neq(0,\dots,0)$. At last, for a fixed $\mathbf{c}'$ in C^k, following the notation introduced in [vdG13], with a slight abuse of notation, we denote by $I_{w,1}(\mathbf{c}-\mathbf{c}')$ and $I_{w,2}(\mathbf{c}-\mathbf{c}')$ the quantities

$$\begin{cases} I_{w,1}(\mathbf{c}-\mathbf{c}') &= I_w((\mathbf{c}-\mathbf{c}')_{S(\mathbf{c}')}), \\ I_{w,2}(\mathbf{c}-\mathbf{c}') &= I_w((\mathbf{c}-\mathbf{c}')_{S^c(\mathbf{c}')}). \end{cases}$$

The following result is derived from Lemma A.4 in [vdG08].

Lemma 4.1. *Let* $\mathbf{c}'$ *be a fixed codebook. Then, for every* $\mathbf{c}$ *in* C^k *and* $\delta>0$*,*

$$2\lambda I_{w,1}(\mathbf{c}-\mathbf{c}') \le \frac{1}{\delta}\ell(\mathbf{c},\mathbf{c}^*)+\frac{1}{\delta}\ell(\mathbf{c}',\mathbf{c}^*)+\frac{2\delta\lambda^2}{\kappa_0}\|w_{S(\mathbf{c}')}\|^2. \tag{4.12}$$

The proof of Lemma 4.1 can be found in [vdG08]. For the sake of completeness it is briefly recalled here.

Proof of Lemma 4.1. Using Cauchy-Schwarz inequality, it is easy to see that

$$\begin{aligned} 2\lambda I_{w,1}(\mathbf{c}-\mathbf{c}') &\le 2\lambda\sqrt{\sum_{p\in\S(\mathbf{c}')} w_p^2}\|\mathbf{c}-\mathbf{c}'\| \\ &\le 2\lambda\sqrt{\sum_{p\in\S(\mathbf{c}')} w_p^2}(\|\mathbf{c}-\mathbf{c}^*\|+\|\mathbf{c}'-\mathbf{c}^*\|). \end{aligned}$$

Using the inequality $2ab\le\frac{\kappa_0}{\delta}a^2+\frac{\delta}{\kappa_0}b^2$, and applying (4.2) leads to

$$2\lambda I_{w,1}(\mathbf{c}-\mathbf{c}') \le \frac{1}{\delta}(\ell(\mathbf{c},\mathbf{c}^*)+\ell(\mathbf{c},\mathbf{c}'))+\frac{2\delta\lambda^2}{\kappa_0}\|w_{S(\mathbf{c}')}\|^2.$$

□

Now turn to the case where $w=1$, so that $\|w_{S(\mathbf{c}')}\|^2=\|\mathbf{c}'\|_0$, and choose $\lambda\ge 2\lambda_1$. Let $\mathbf{c}'$ be a fixed codebook, to be chosen later. The fundamental Lasso inequality yields

$$P_n\gamma(\hat{\mathbf{c}}_{n,\lambda},.)+\lambda I_L(\hat{\mathbf{c}}_{n,\lambda})\le P_n\gamma(\mathbf{c}',.)+\lambda I_L(\mathbf{c}',.),$$

so that

$$P\gamma(\hat{\mathbf{c}}_{n,\lambda},.)+\lambda I_L(\hat{\mathbf{c}}_{n,\lambda})\le P\gamma(\mathbf{c}',.)+\lambda I_L(\mathbf{c}',.)+(P-P_n)(\gamma(\hat{\mathbf{c}}_{n,\lambda},.)-\gamma(\mathbf{c}',.)).$$

Splitting $I_L(\hat{\mathbf{c}}_{n,\lambda})$ in $I_{L,1}(\hat{\mathbf{c}}_{n,\lambda}) + I_{L,2}(\hat{\mathbf{c}}_{n,\lambda})$, it may be easily derived that $I_L(\mathbf{c}') - I_{L,1}(\hat{\mathbf{c}}_{n,\lambda}) \leq I_{L,1}(\hat{\mathbf{c}}_{n,\lambda} - \mathbf{c}')$ and $I_{L,2}(\hat{\mathbf{c}}_{n,\lambda}) = I_{L,2}(\hat{\mathbf{c}}_{n,\lambda} - \mathbf{c}')$. It follows that

$$P\gamma(\hat{\mathbf{c}}_{n,\lambda},.) + \lambda I_{L,2}(\hat{\mathbf{c}}_{n,\lambda} - \mathbf{c}') \leq P\gamma(\mathbf{c}',.) + \lambda I_{L,1}(\hat{\mathbf{c}}_{n,\lambda} - \mathbf{c}') + (P - P_n)(\gamma(\hat{\mathbf{c}}_{n,\lambda},.) - \gamma(\mathbf{c}',.)).$$

Consequently, Proposition 4.11 yields, with probability larger than $1 - e^{-x}$,

$$\begin{aligned} \ell(\mathbf{c},\mathbf{c}^*) + \lambda I_L(\hat{\mathbf{c}}_{n,\lambda} - \mathbf{c}') &\leq \ell(\mathbf{c}',\mathbf{c}^*) + 2\lambda I_{L,1}(\hat{\mathbf{c}}_{n,\lambda} - \mathbf{c}') + (P - P_n)(\gamma(\hat{\mathbf{c}}_{n,\lambda},.) - \gamma(\mathbf{c}',.)) \\ &\leq \ell(\mathbf{c}',\mathbf{c}^*) + \lambda_1 (I_L(\hat{\mathbf{c}}_{n,\lambda} - \mathbf{c}') \vee \lambda_0) + 2\lambda I_{L,1}(\hat{\mathbf{c}}_{n,\lambda} - \mathbf{c}'). \end{aligned} \tag{4.13}$$

Hence, applying Lemma 4.1 with $\delta = 2$ leads to

$$\ell(\hat{\mathbf{c}}_{n,\lambda},\mathbf{c}^*) + 2\lambda I_L(\hat{\mathbf{c}}_{n,\lambda} - \mathbf{c}') \leq 3\ell(\mathbf{c}',\mathbf{c}^*) + \frac{8\lambda^2}{\kappa_0}\|\mathbf{c}'\|_0 + 2\lambda_1 (I_L(\hat{\mathbf{c}}_{n,\lambda} - \mathbf{c}') \vee \lambda_0). \tag{4.14}$$

If $I(\hat{\mathbf{c}}_{n,\lambda} - \mathbf{c}') \leq \lambda_0$, then it is clear that $\lambda I(\hat{\mathbf{c}}_{n,\lambda} - \mathbf{c}') \leq \lambda^2$. Otherwise, we have

$$2\lambda I_L(\hat{\mathbf{c}}_{n,\lambda} - \mathbf{c}') \leq 3\ell(\mathbf{c}',\mathbf{c}^*)\frac{8\lambda^2}{\kappa_0}\|\mathbf{c}'\|_0 + 2\lambda_1 I(\hat{\mathbf{c}}_{n,\lambda} - \mathbf{c}').$$

Since $\lambda \geq 2\lambda_1$, the consistency result easily follows, taking $\mathbf{c}' = \mathbf{c}^*_\lambda$. Let us turn to the prediction result.

If $I_L(\hat{\mathbf{c}}_{n,\lambda} - \mathbf{c}^*) \leq \lambda_0$, then $I_{L,1}(\hat{\mathbf{c}}_{n,\lambda} - \mathbf{c}^*) \leq \lambda_0$. Consequently, taking $\mathbf{c}' = \mathbf{c}^*$ in (4.13) yields

$$\ell(\hat{\mathbf{c}}_{n,\lambda},\mathbf{c}^*) \leq \lambda_0\lambda_1 + 2\lambda\lambda_0.$$

Since $\lambda_0 \leq \lambda_1 \leq \lambda/2$, it may be easily derived that $\ell(\hat{\mathbf{c}}_{n,\lambda},\mathbf{c}^*) \leq 2\lambda^2$. If $I_L(\hat{\mathbf{c}}_{n,\lambda} - \mathbf{c}^*) > \lambda_0$, then taking $\mathbf{c}' = \mathbf{c}^*$ in (4.14) ensures that

$$\ell(\hat{\mathbf{c}}_{n,\lambda},\mathbf{c}^*) + 2(\lambda - \lambda_1) I_L(\hat{\mathbf{c}}_{n,\lambda} - \mathbf{c}^*) \leq \frac{8\lambda^2}{\kappa_0}\|\mathbf{c}^*\|_0.$$

4.5.8 Proof of Theorem 4.4

Throughout this Subsection, the sequence w will be chosen as $w_p = \sigma_p$, so that $T(w) = T$ and $\bar{M}(w) = \sqrt{k}\sigma^2 T$. Choose $\lambda \geq \frac{2}{\sqrt{1-\alpha(y)}}\lambda_1$, where λ_1 is defined in Proposition 4.11. By definition of the Weighted Lasso k-means procedure, we have

$$P_n\gamma(\hat{\mathbf{c}}_{n,\lambda},.) + \lambda\hat{I}_{WL}(\hat{\mathbf{c}}_{n,\lambda}) \leq P_n\gamma(\mathbf{c}',.) + \lambda\hat{I}_{WL}(\mathbf{c}').$$

As in the previous Subsection, this leads to

$$\ell(\hat{\mathbf{c}}_{n,\lambda},\mathbf{c}^*) + \lambda\hat{I}_{WL}(\hat{\mathbf{c}}_{n,\lambda} - \mathbf{c}') \leq \ell(\mathbf{c}',\mathbf{c}^*) + 2\lambda\hat{I}_{WL,1}(\hat{\mathbf{c}}_{n,\lambda} - \mathbf{c}') + (P - P_n)(\gamma(\hat{\mathbf{c}}_{n,\lambda},.) - \gamma(\mathbf{c}',.)).$$

Using Proposition 4.4 and Proposition 4.11, it easily follows that, with probability larger than $1 - e^{-x} - e^{-y}$,

$$\ell(\hat{\mathbf{c}}_{n,\lambda},\mathbf{c}^*) + \lambda\sqrt{1-\alpha(y)}I_{WL}(\hat{\mathbf{c}}_{n,\lambda} - \mathbf{c}') \leq \ell(\mathbf{c}',\mathbf{c}^*) + \lambda_1(I_{WL}(\hat{\mathbf{c}}_{n,\lambda} - \mathbf{c}') \vee \lambda_0) + 2\sqrt{1+\alpha(y)}I_{WL,1}(\hat{\mathbf{c}}_{n,\lambda} - \mathbf{c}'). \tag{4.15}$$

Now, applying Lemma 4.1 with $\delta = \frac{1}{2\sqrt{1+\alpha(y)}}$ and choosing $\mathbf{c}' = \mathbf{c}^*_\lambda$ leads to

$$\frac{1}{2\ell(\hat{\mathbf{c}}_{n,\lambda}, \mathbf{c}^*)} + \lambda\sqrt{1-\alpha(y)} I_{WL}(\hat{\mathbf{c}}_{n,\lambda} - \mathbf{c}^*_\lambda) \leq \frac{3}{2}\ell(\mathbf{c}^*_\lambda, \mathbf{c}^*) + \frac{4(1+\alpha(y))\lambda^2 \sigma^2_{S(\mathbf{c}^*_\lambda)}}{\kappa_0} + \lambda_1 (I_{WL}(\hat{\mathbf{c}}_{n,\lambda} - \mathbf{c}^*_\lambda) \vee \lambda_0).$$

Recalling that $\lambda \geq \frac{2}{\sqrt{1-\alpha(y)}}\lambda_1$ and $\lambda_1 \geq \lambda_0$, if $I(\hat{\mathbf{c}}_{n,\lambda} - \mathbf{c}^*_\lambda) \leq \lambda_0$, then

$$\lambda I_{WL}(\hat{\mathbf{c}}_{n,\lambda} - \mathbf{c}^*_\lambda) \leq \sqrt{1-\alpha(y)}\lambda^2.$$

Otherwise, we have

$$\lambda I_{WL}(\hat{\mathbf{c}}_{n,\lambda} - \mathbf{c}^*_\lambda) \leq \frac{1}{\sqrt{1-\alpha(y)}} \left[3\ell(\mathbf{c}^*_\lambda, \mathbf{c}^*) + \frac{8(1+\alpha(y))\lambda^2 \sigma^2_{S(\mathbf{c}^*_\lambda)}}{\kappa_0} \right].$$

Let us turn now to the prediction result. Suppose that $I_{WL}(\hat{\mathbf{c}}_{n,\lambda} - \mathbf{c}^*) \leq \lambda_0$. Then, if $\mathbf{c}' = \mathbf{c}^*$, the Lasso inequality combined with Proposition 4.11 ensures that

$$\ell(\hat{\mathbf{c}}_{n,\lambda}, \mathbf{c}^*) + \lambda \hat{I}_{WL,2}(\hat{\mathbf{c}}_{n,\lambda} - \mathbf{c}^*) \leq \lambda_1\lambda_0 + \lambda \hat{I}_{WL,1}(\hat{\mathbf{c}}_{n,\lambda} - \mathbf{c}^*),$$

which leads to, applying Proposition 4.4,

$$\ell(\hat{\mathbf{c}}_{n,\lambda}, \mathbf{c}^*) \leq \lambda_1\lambda_0 + \sqrt{1+\alpha(y)}\lambda\lambda_0 \leq \frac{\lambda^2}{2}\left(\sqrt{(1-\alpha(y))(1+\alpha(y))} + \frac{1}{2}(1-\alpha(y)) \right).$$

Since $\sqrt{1-\alpha(y)} + \sqrt{1+\alpha(y)} \leq 2$, it is easy to see that

$$\ell(\hat{\mathbf{c}}_{n,\lambda}, \mathbf{c}^*) \leq \sqrt{1-\alpha(y)}\lambda^2.$$

Now if $I_{WL}(\hat{\mathbf{c}}_{n,\lambda} - \mathbf{c}^*) > \lambda_0$, choosing $\mathbf{c}' = \mathbf{c}^*$ in (4.15) and applying Lemma 4.1, with $\delta = \frac{1}{2\sqrt{1+\alpha(y)}}$, leads to

$$\frac{1}{2}\ell(\hat{\mathbf{c}}_{n,\lambda}, \mathbf{c}^*) + \frac{\sqrt{1-\alpha(y)}}{2}\lambda I_{WL}(\hat{\mathbf{c}}_{n,\lambda} - \mathbf{c}^*) \leq \frac{4(1+\alpha(y))\sigma^2_{S(\mathbf{c}^*)}\lambda^2}{\kappa_0}.$$

4.5.9 Proofs of Proposition 4.6, Proposition 4.7, Proposition 4.8 and Proposition 4.9

According to Proposition 4.3 and Proposition 4.5, it is easy to see that, for every variable p in $\{1, \ldots, d\}$, a sufficient condition for the p-th coordinate of $\mathbf{c}^*_\lambda$ to be driven to 0 is

$$u\left(\sigma_p^2 - R_p^*\right) < \frac{8(1+\alpha)}{3\kappa_0}\lambda^2, \tag{4.16}$$

where for the Lasso k-means strategy $u = 1$ and $\alpha = 0$, whereas for the Weighted Lasso k-means strategy $u = 1/\sigma_p^2$ and α is defined as in Theorem 4.3. Therefore, to derive a sufficient condition for a variable to be driven to 0, it suffices to give an upper bound on σ_p^2 and a lower bound on R_p^*. The following Lemma provides theses bounds.

Lemma 4.2. *Let P denote the distribution of an unidimensional Gaussian mixture, with weights $\theta_1 = 0.3$, $\theta_2 = 0.2$, $\theta_3 = 0.2$ and $\theta_4 = 0.3$, means m_1, m_2, m_3, m_4 such that $m_2 = 0$ and $|m_1| = |m_3| = |m_4| := m$ (m can be equal to 0), and common variance s^2. Then*

$$\begin{cases} \sigma^2 & = & s^2 + (1-\theta_2)m^2, \\ R^*(P) & \geq & s^2/k^2, \end{cases}$$

where $\sigma^2 = Px^2$ and $R^*(P)$ denotes the optimal k quantization error for P.

The first bound of Lemma 4.2 derives from straightforward computation, the second bound follows from a lower bound for Gaussian process exposed in [LP02]. A short proof is given in Section 4.6.4.

Applying Lemma 4.2 for every variable, in Model 1 and Model 2, and plugging the obtained quantities in (4.16) for the Lasso k-means and Weighted Lasso k-means strategies leads to the result.

4.6 Technical results

4.6.1 Proof of Proposition 4.10

This proof is a slight modification of a result in [BDL08], namely Lemma 4.3. Introducing some independent Rademacher variables ε_i, $i = 1, \dots, n$, such that $\varepsilon_i = \pm 1$ with probability 1/2, and applying the symmetrization principle (see, e.g., Section 2.2 in [Kol06]) leads to

$$\mathbb{E} \sup_{I_w(\mathbf{c}) \leq r} (P - P_n)\bar{\gamma}(\mathbf{c}, .) \leq 2\mathbb{E}_X \mathbb{E}_\varepsilon \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \min_{j=1,\dots,k} -2\langle X_i, c_j \rangle + \|c_j\|^2,$$

where $\mathbb{E}_Z$ means expectation with respect to the law of Z, for some random variable Z. Let us denote by $I_w(c)$ the norm $I_w(c) = \sum_{p=1}^{d} w_p |c_p|$, for a code point c in C. Proceeding by induction on k as done in Lemma 4.3 $ii)$ in [BDL08], we may write

$$\mathbb{E}_\varepsilon \sup_{I_w(c) \leq r} \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \min_{j=1,\dots,k} -2\langle X_i, c_j \rangle + \|c_j\|^2 \leq 2k \left[\mathbb{E}_\varepsilon \sup_{I_w(c) \leq r} \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \langle X_i, c \rangle + \frac{rT(w)}{2\sqrt{n}} \right].$$

At last, it is immediate that

$$\mathbb{E}_{X,\varepsilon} \sup_{I(c) \leq r} \left\langle c, \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i X_i \right\rangle \leq r \mathbb{E}_{X,\varepsilon} \sup_{p=1,\dots,d} \left| \sum_{i=1}^{n} \varepsilon_i \frac{X_i^{(p)}}{w_p} \right|.$$

When $X_1, \dots, X_n$ is fixed, Hoeffding's inequality ensures that, for every $p = 1, \dots, d$, $\sum_{i=1}^{n} \varepsilon_i \frac{X_i^{(p)}}{w_p}$ is subgaussian with variance $\frac{T(w)^2}{n}$. For a comprehensive introduction to subgaussian variables and its application to empirical processes theory the interested reader is referred to [Mas07]. Applying Theorem 3.12 of [Mas07] ensures that

$$\mathbb{E}_\varepsilon \sup_{p=1,\dots,d} \left| \sum_{i=1}^{n} \varepsilon_i \frac{X_i^{(p)}}{w_p} \right| \leq \sqrt{2\log(d)} \frac{T(w)}{\sqrt{n}},$$

which leads to

$$\mathbb{E}\sup_{I_w(\mathbf{c})\le r}(P-P_n)\bar{\gamma}(\mathbf{c},.)\le\frac{2kT(w)}{\sqrt{n}}r+\frac{4k\sqrt{2\log(d)}T(w)}{\sqrt{n}}r.$$

Applying a bounded difference concentration inequality such as Theorem 5.1 in [Mas07] leads to the desired result.

4.6.2 Proof of Proposition 4.11

For a fixed $\mathbf{c}'$ in $\mathbf{c}^k$, denote by $Z_r(\mathbf{c}')$ the following random variable

$$Z_r(\mathbf{c}')=\sup_{I_w(\mathbf{c}-\mathbf{c}')\le r}\left|(P-P_n)(\gamma(\mathbf{c},.)-\gamma(\mathbf{c}',.)\right|$$

The following Proposition offers a bound on $Z_r(\mathbf{c}')$.

Proposition 4.12. *Suppose that w is deterministic. Let $x>0$, and $\mathbf{c}'$ be a fixed codebook. Then, with probability larger than $1-e^{-x}$,*

$$Z_r(\mathbf{c}')\le 16\sqrt{2\pi}\sqrt{\frac{k\log(kd)}{n}}rT(w)\left(1+\frac{1}{4\sqrt{\pi}}\sqrt{\frac{x}{k\log(kd)}}\right).$$

The proof of Proposition 4.12 can be found in the next Subsection. Proposition 4.11 derives from a peeling argument, as in Section 3.4 of [vdG13], combined with Proposition 4.12. Let a be such that $e^{-(a-1)}2\bar{M}\le\lambda_0$, and take $u_0=\log(a)$. Then it is easy to see that $u_0\le u$, where u is defined in Proposition 4.11. We may write

$$\begin{aligned}
&\mathbb{P}\left(\sup_{I_w(\mathbf{c}-\mathbf{c}')\le 2\bar{M}(w)}\frac{\left|(P-P_n)(\gamma(\mathbf{c},.)-\gamma(\mathbf{c}',.))\right|}{I_w(\mathbf{c}-\mathbf{c}')\vee\lambda_0}\ge\lambda_1\right)\\
&\le\sum_{j=2}^{a}\mathbb{P}\left(\sup_{\substack{I_w(\mathbf{c}-\mathbf{c}')\le 2e^{-(j-1)}\bar{M}(w)\\ I_w(\mathbf{c}-\mathbf{c}')\ge 2e^{-j}\bar{M}(w)}}\frac{\left|(P-P_n)(\gamma(\mathbf{c},.)-\gamma(\mathbf{c}',.))\right|}{2e^{-j}\bar{M}(w)}\ge\lambda_1\right)\\
&\qquad+\mathbb{P}\left(\sup_{I_w(\mathbf{c}-\mathbf{c}')\le\lambda_0}\frac{\left|(P-P_n)(\gamma(\mathbf{c},.)-\gamma(\mathbf{c}',.))\right|}{2e^{-(a-1)}\bar{M}(w)}\ge\lambda_1\right)\\
&\le\sum_{j=1}^{a}\mathbb{P}\left(Z_{2e^{-(j-1)}\bar{M}(w)}\ge 2e^{-(j-1)}\bar{M}(w)\lambda_0\left(1+\sqrt{\frac{u+x}{k\log kd}}\right)\right)\\
&\le ae^{-u}e^{-x},
\end{aligned}$$

where the last inequality follows from Proposition 4.12. Noticing that $ae^{-u}\le 1$ proves the result.

4.6.3 Proof of Proposition 4.12

This proof is a slight modification of the proof of Theorem 3.1 in Chapter 3, and mainly relies on the use of Gaussian complexities combined with Slepian's Lemma (see, e.g., Theorem 3.14 in [Mas07]). First, it may be easily noticed that, for every $j=1,\dots,k$, if $I_w(\mathbf{c}-\mathbf{c}')\le r$, then, for all x in $\mathbb{R}^d$,

$$\left|-2\langle x,c_j\rangle+\|c_j\|^2+2\left\langle x,c_j'\right\rangle-\|c_j'\|^2\right|\le 4rT(w),$$

which leads to

$$\|\gamma(\mathbf{c},.)-\gamma(\mathbf{c}',.)\|_\infty \le 4rT(w).$$

As a consequence, a bounded difference concentration inequality (see, e.g., Theorem 5.1 in [Mas07]) yields, with probability larger than $1-e^{-x}$,

$$Z_r \le \mathbb{E}Z_r + 4rT(w)\sqrt{\frac{2x}{n}}.$$

It remains to bound from above $\mathbb{E}Z_r$. According to the symmetrization principle (see, e.g., Section 2.2 of [Kol06]), combined with Lemma 4.5 of [LT91], we may write

$$\mathbb{E}Z_r \le 2\sqrt{\frac{\pi}{2}}\mathbb{E}_X\mathbb{E}_g \sup_{I_w(\mathbf{c}-\mathbf{c}')\le r} \frac{1}{n}\sum_{i=1}^n g_i(\gamma(\mathbf{c},X_i)-\gamma(\mathbf{c}',X_i)),$$

where the g_i's are independent standard Gaussian variables. Let $\mathbf{c}'$ and $X_1,\dots,X_n$ be fixed, and define, for $\mathbf{c}$ such that $I_w(\mathbf{c}-\mathbf{c}')\le r$ the Gaussian process

$$Y_{\mathbf{c}} = \sum_{i=1}^n g_i(\gamma(\mathbf{c},X_i)-\gamma(\mathbf{c}',X_i)).$$

Since, for every codebooks $\mathbf{c}_1$ and $\mathbf{c}_2$,

$$(\gamma(\mathbf{c}_1,X_i)-\gamma(\mathbf{c}_2,X_i))^2 \le \max_{j=1,\dots,k} 8\langle c_{1,j}-c_{2,j},X_i\rangle^2 + 2(\|c_{1,j}\|^2-\|c_{2,j}|^2)^2,$$

it is easy to see that

$$\mathrm{Var}(Y_{\mathbf{c}_1}-Y_{\mathbf{c}_2}) \le \sum_{i=1}^n\sum_{j=1}^k 8\langle c_{1,j}-c_{2,j},X_i\rangle^2 + 2n\sum_{j=1}^k(\|c_{1,j}\|^2-\|c_{2,j}\|^2)^2.$$

To derive bounds on the Gaussian complexity defined above, the following comparison result between Gaussian processes is needed.

Theorem 4.5 (Slepian's Lemma)**.** *Let Y_t and N_t, t in $\mathcal{V}$, be some centered real Gaussian processes. Assume that*

$$\forall t_1,t_2\in\mathcal{V} \quad \mathrm{Var}(Y_{t_1}-Y_{t_2}) \le \mathrm{Var}(N_{t_1}-N_{t_2}),$$

then

$$\mathbb{E}\sup_{t\in\mathcal{V}} Y_t \le 2\mathbb{E}\sup_{t\in\mathcal{V}} N_t.$$

A proof of Theorem 4.5 can be found in Theorem 3.14 of [Mas07]. Denote by $\mathcal{V}$ the set of codebooks $\mathbf{c}$ in C^k such that $I_w(\mathbf{c}-\mathbf{c}')\le r$. Now introduce, for $\mathbf{c}$ such that $I_w(\mathbf{c}-\mathbf{c}')\le r$, the following Gaussian process

$$N_{\mathbf{c}} = 2\sqrt{2}\sum_{i=1}^n\sum_{j=1}^k \left\langle c_j-c'_j,X_i\right\rangle \xi_{i,j} + \sqrt{2n}\sum_{j=1}^k(\|c_j\|^2-\|c'_j\|^2)\xi'_j,$$

where the ξ's and ξ''s are independent standard Gaussian random variables. It is worth noticing that, for all $\mathbf{c}_1$ and $\mathbf{c}_2$ in $\mathcal{V}$, $\mathrm{Var}(Y_{\mathbf{c}_1}-Y_{\mathbf{c}_2}) \le \mathrm{Var}(N_{\mathbf{c}_1}-N_{\mathbf{c}_2})$. Consequently, applying Theorem 4.5 yields

$$\mathbb{E}_g \sup_{I_w(\mathbf{c}-\mathbf{c}')\le r} Y_{\mathbf{c}} \le 2\mathbb{E}_{\xi,\xi'} \sup_{I_w(\mathbf{c}-\mathbf{c}')\le r} N_{\mathbf{c}}.$$

It follows that

$$\mathbb{E}_{\xi,\xi'} \sup_{I_w(\mathbf{c}-\mathbf{c}')\le r} N_{\mathbf{c}} \le \mathbb{E}_{\xi} \sup_{I_w(\mathbf{c}-\mathbf{c}')\le r} 2\sqrt{2}\sum_{i=1}^{n}\sum_{j=1}^{k}\left\langle c_j - c'_j, X_i\right\rangle \xi_{i,j} + \mathbb{E}_{\xi'} \sup_{I_w(\mathbf{c}-\mathbf{c}')\le r} \sqrt{2n}\sum_{j=1}^{k}(\|c_j\|^2 - \|c'_j\|^2)\xi'_j.$$

The first term of the right side can be bounded as follows.

$$\begin{aligned}
&\mathbb{E}_{\xi} \sup_{I_w(\mathbf{c}-\mathbf{c}')\le r} 2\sqrt{2}\sum_{i=1}^{n}\sum_{j=1}^{k}\left\langle c_j - c'_j, X_i\right\rangle \xi_{i,j} \\
&\le 2\sqrt{2}\mathbb{E}_{\xi} \sup_{I_w(\mathbf{c}-\mathbf{c}')\le r} \sum_{j=1}^{k}\left\langle c_j - c'_j, \sum_{i=1}^{n}\xi_{i,j}X_i\right\rangle \\
&\le 2\sqrt{2}\mathbb{E}_{\xi} \sup_{I_w(\mathbf{c}-\mathbf{c}')\le r} \left(\sum_{j=1}^{k}\sum_{p=1}^{d} w_p |c_j^{(p)} - c_j'^{(p)}|\right) \max_{j,p}\left|\sum_{i=1}^{n}\frac{\xi_{i,j}X_i^{(p)}}{w_p}\right| \\
&\le 2\sqrt{2k}r\mathbb{E}_{\xi} \max_{j=1,\dots,k,p=1,\dots,d}\left|\sum_{i=1}^{n}\frac{\xi_{i,j}X_i^{(p)}}{w_p}\right|.
\end{aligned}$$

It is worth noticing that, for every (j,p), the random variable $\sum_{i=1}^{n}\frac{\xi_{i,j}X_i^{(p)}}{w_p}$ is Gaussian, with variance bounded by $nT^2(w)$. Consequently, applying Theorem 3.12 in [Mas07] gives

$$\mathbb{E}_{\xi} \max_{j=1,\dots,k,p=1,\dots,d}\left|\sum_{i=1}^{n}\frac{\xi_{i,j}X_i^{(p)}}{w_p}\right| \le T(w)\sqrt{2n\log(kd)}.$$

In turn, the second term of the right side may be bounded by

$$\begin{aligned}
&\mathbb{E}_{\xi'} \sup_{I_w(\mathbf{c}-\mathbf{c}')\le r} \sqrt{2n}\sum_{j=1}^{k}(\|c_j\|^2 - \|c'_j\|^2)\xi'_j \\
&\qquad \le \sqrt{2n}\mathbb{E}_{\xi'} \sup_{I_w(\mathbf{c}-\mathbf{c}')\le r} \sum_{j=1}^{k}\left(\sum_{p=1}^{d} w_p |c_j^{(p)} - c_j'^{(p)}|\frac{2M_p}{w_p}\right)\left|\xi'_j\right| \\
&\qquad \le 2\sqrt{2n}T(w)\mathbb{E}_{\xi'} \sup_{I_w(\mathbf{c}-\mathbf{c}')\le r} I(\mathbf{c}-\mathbf{c}')\sqrt{\sum_{j=1}^{k}\xi_j'^2} \\
&\qquad \le 2T(w)r\sqrt{2nk}.
\end{aligned}$$

Combining these two bounds leads to

$$\mathbb{E}Z_r(\mathbf{c}') \le 16\sqrt{2\pi}\sqrt{\frac{k\log(kd)}{n}}rT(w).$$

4.6.4 Proof of Lemma 4.2

The second moment bound follows from

$$\begin{aligned}
Px^2 &= \sum_{j=1}^{4}\theta_j(m_j^2 + s^2) \\
&= (1-\theta_2)m^2 + s^2.
\end{aligned}$$

The lower bound on the k quantization error is derived from a version of Proposition 4.9 in [LP02] adapted to the unidimensional normal distribution, namely

Proposition 4.13 (cf. Proposition 4.9 in [LP02])**.** *Suppose that P has a Gaussian distribution with variance s^2. Then*

$$R^*(P) \geq s^2/k^2.$$

It remains to relate the k quantization error of an unidimensional Gaussian mixture to the k quantization error of a Gaussian distribution. For any distribution Q denote by $R^*(Q)$ its optimal k quantization error. Also define the random variable Z by $\mathbb{P}(Z = i) = \theta_i$, for $i = 1, \ldots, 4$, and the Gaussian mixture P such that $P|\{Z = i\}$ is an unidimensional Gaussian distribution with mean m_i and variance s^2. Then it is easy to see that

$$\begin{aligned} R^*(P) &\geq \sum_{i=1}^{4} \theta_i R^*(P|Z = i) \\ &\geq R^*(\mathcal{N}(0, s^2)) \\ &\geq s^2/k^2, \end{aligned}$$

according to Proposition 4.13.

Bibliographie

[ABCP13] Anestis ANTONIADIS, Xavier BROSSAT, Jairo CUGLIARI et Jean-Michel POGGI : Clustering functional data using wavelets. *Int. J. Wavelets Multiresolut. Inf. Process.*, 11(1):1350003, 30, 2013.

[AF12] Benjamin AUDER et Aurélie FISCHER : Projection-based curve clustering. *J. Stat. Comput. Simul.*, 82(8):1145–1168, 2012.

[AGG05] András ANTOS, László GYÖRFI et András GYÖRGY : Individual convergence rates in empirical vector quantizer design. *IEEE Trans. Inform. Theory*, 51(11):4013–4022, 2005.

[Ant05] András ANTOS : Improved minimax bounds on the test and training distortion of empirically designed vector quantizers. *IEEE Trans. Inform. Theory*, 51(11):4022–4032, 2005.

[Bac08] Francis R. BACH : Consistency of the group lasso and multiple kernel learning. *J. Mach. Learn. Res.*, 9:1179–1225, 2008.

[Bad77] Adrian BADDELEY : Integrals on a moving manifold and geometrical probability. *Advances in Appl. Probability*, 9(3):588–603, 1977.

[BBM99] Andrew BARRON, Lucien BIRGÉ et Pascal MASSART : Risk bounds for model selection via penalization. *Probab. Theory Related Fields*, 113(3): 301–413, 1999.

[BBM08] Gilles BLANCHARD, Olivier BOUSQUET et Pascal MASSART : Statistical performance of support vector machines. *Ann. Statist.*, 36(2):489–531, 2008.

[BDL08] Gérard BIAU, Luc DEVROYE et Gábor LUGOSI : On the performance of clustering in Hilbert spaces. *IEEE Trans. Inform. Theory*, 54(2):781–790, 2008.

[BG98] Toby BERGER et Jerry D. GIBSON : Lossy source coding. *IEEE Trans. Inform. Theory*, 44(6):2693–2723, 1998. Information theory : 1948–1998.

[BH89] Eric B. BAUM et David HAUSSLER : What size net gives valid generalization ? *Neural Comput.*, 1(1):151–160, mars 1989.

[BJM06] Peter L. BARTLETT, Michael I. JORDAN et Jon D. MCAULIFFE : Convexity, classification, and risk bounds. *J. Amer. Statist. Assoc.*, 101(473):138–156, 2006.

[BLL98] Peter L. BARTLETT, Tamás LINDER et Gábor LUGOSI : The minimax distortion redundancy in empirical quantizer design. *IEEE Trans. Inform. Theory*, 44(5):1802–1813, 1998.

[BM02] Peter L. BARTLETT et Shahar MENDELSON : Rademacher and Gaussian complexities : risk bounds and structural results. *J. Mach. Learn. Res.*, 3(Spec. Issue Comput. Learn. Theory):463–482, 2002.

[Bou02] Olivier BOUSQUET : A Bennett concentration inequality and its application to suprema of empirical processes. *C. R. Math. Acad. Sci. Paris*, 334(6):495–500, 2002.

[Bre11] Haim BREZIS : *Functional analysis, Sobolev spaces and partial differential equations*. Universitext. Springer, New York, 2011.

[BV04] Stephen BOYD et Lieven VANDENBERGHE : *Convex optimization*. Cambridge University Press, Cambridge, 2004.

[Cha12] Sohail CHAND : On tuning parameter selection of lasso-type methods - a monte carlo study. *In Proceedings of 9th IBCAST*, pages 120–129, Islamabad, Pakistan, 2012.

[Cho94] Philip A. CHOU : The distortion of vector quantizers trained on n vectors decreases to the optimum as $\mathcal{O}_p(1/n)$. *In Proc. IEEE Int. Symp. Inf. Theory*, page 457, Trondheim, Norway, 1994.

[CL06] Michaël CHICHIGNOUD et Sébastien LOUSTAU : Adaptive Noisy Clustering. 2013-06.

[CLG89] Philip A. CHOU, Tom LOOKABAUGH et Robert M. GRAY : Entropy-constrained vector quantization. *IEEE Trans. Acoust. Speech Signal Process.*, 37(1):31–42, 1989.

[CP12] Benoît CADRE et Quentin PARIS : On Hölder fields clustering. *TEST*, 21(2):301–316, 2012.

[CPR12] Guillermo D. CAÑAS, Tomaso POGGIO et Lorenzo ROSASCO : Learning manifolds with k-means and k-flats. *CoRR*, abs/1209.1121, 2012.

[CWLX14] Xiangyu CHANG, Yu WANG, Rongjian LI et Zongben XU : Sparse K-Means with ℓ_∞/ℓ_0 Penalty for High-Dimensional Data Clustering. *ArXiv e-prints*, 2014.

[DeS]

[DGL96] Luc DEVROYE, László GYÖRFI et Gábor LUGOSI : *A probabilistic theory of pattern recognition*, volume 31 de *Applications of Mathematics (New York)*. Springer-Verlag, New York, 1996.

[DGLP04] Sylvain DELATTRE, Siegfried GRAF, Harald LUSCHGY et Gilles PAGÈS : Quantization of probability distributions under norm-based distortion measures. *Statist. Decisions*, 22(4):261–282, 2004.

[Dud67] R. M. DUDLEY : The sizes of compact subsets of Hilbert space and continuity of Gaussian processes. *J. Functional Analysis*, 1:290–330, 1967.

[Dud02] R. M. DUDLEY : *Real analysis and probability*, volume 74 de *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge, 2002. Revised reprint of the 1989 original.

[Fis10] Aurélie FISCHER : Quantization and clustering with Bregman divergences. *J. Multivariate Anal.*, 101(9):2207–2221, 2010.

[FR02] Chris FRALEY et Adrian E. RAFTERY : Model-based clustering, discriminant analysis, and density estimation. *J. Amer. Statist. Assoc.*, 97(458):611–631, 2002.

[GG91] Allen GERSHO et Robert M. GRAY : *Vector quantization and signal compression*. Kluwer Academic Publishers, Norwell, MA, USA, 1991.

[GL00] Siegfried GRAF et Harald LUSCHGY : *Foundations of quantization for probability distributions*, volume 1730 de *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 2000.

[GLP03] Siegfried GRAF, Harald LUSCHGY et Gilles PAGÈS : Functional quantization and small ball probabilities for Gaussian processes. *J. Theoret. Probab.*, 16(4):1047–1062 (2004), 2003.

[GLP07] Siegfried GRAF, Harald LUSCHGY et Gilles PAGÈS : Optimal quantizers for Radon random vectors in a Banach space. *J. Approx. Theory*, 144(1): 27–53, 2007.

[GZ84] Evarist GINÉ et Joel ZINN : Some limit theorems for empirical processes. *Ann. Probab.*, 12(4):929–998, 1984. With discussion.

[HTF09] Trevor HASTIE, Robert TIBSHIRANI et Jerome FRIEDMAN : *The elements of statistical learning*. Springer Series in Statistics. Springer, New York, second édition, 2009. Data mining, inference, and prediction.

[Jun12] Stefan JUNGLEN : *Geometry of optimal codebooks and constructive quantization*. Thèse de doctorat, Universität Trier, Universitätsring 15, 54296 Trier, 2012.

[Kol06] Vladimir KOLTCHINSKII : Local Rademacher complexities and oracle inequalities in risk minimization. *Ann. Statist.*, 34(6):2593–2656, 2006.

[Lev13] Clément LEVRARD : Fast rates for empirical vector quantization. *Electron. J. Stat.*, 7:1716–1746, 2013.

[Lev14] Clément LEVRARD : Non Asymptotic Bounds for Vector Quantization. *ArXiv e-prints*, mai 2014.

[Lin02] Tamás LINDER : Learning-theoretic methods in vector quantization. *In Principles of nonparametric learning (Udine, 2001)*, volume 434 de *CISM Courses and Lectures*, pages 163–210. Springer, Vienna, 2002.

[Llo82] Stuart P. LLOYD : Least squares quantization in PCM. *IEEE Trans. Inform. Theory*, 28(2):129–137, 1982.

[LLZ94] Tamás LINDER, Gábor LUGOSI et Kenneth ZEGER : Rates of convergence in the source coding theorem, in empirical quantizer design, and in universal lossy source coding. *IEEE Trans. Inform. Theory*, 40(6):1728–1740, 1994.

[LP02] Harald LUSCHGY et Gilles PAGÈS : Functional quantization of Gaussian processes. *J. Funct. Anal.*, 196(2):486–531, 2002.

[LT91] Michel LEDOUX et Michel TALAGRAND : *Probability in Banach spaces*, volume 23 de *Ergebnisse der Mathematik und ihrer Grenzgebiete (3) [Results in Mathematics and Related Areas (3)]*. Springer-Verlag, Berlin, 1991. Isoperimetry and processes.

[Lug02] Gabor LUGOSI : Pattern classification and learning theory. *In Principles of nonparametric learning (Udine, 2001)*, volume 434 de *CISM Courses and Lectures*, pages 1–56. Springer, Vienna, 2002.

[Mas07] Pascal MASSART : *Concentration inequalities and model selection*, volume 1896 de *Lecture Notes in Mathematics*. Springer, Berlin, 2007.

Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23, 2003, With a foreword by Jean Picard.

[Mey12] Pascal MEYNET, Caroline ; Massart : *Sélection de variables pour la classification non supervisée en grande dimension*. Thèse de doctorat, 2012. Thèse de doctorat Mathématiques Paris 11 2012.

[Mey13] Caroline MEYNET : An ℓ_1-oracle inequality for the Lasso in finite mixture Gaussian regression models. *ESAIM Probab. Stat.*, 17:650–671, 2013.

[MM11] Pascal MASSART et Caroline MEYNET : The Lasso as an ℓ_1-ball model selection procedure. *Electron. J. Stat.*, 5:669–687, 2011.

[MM12] Pascal MASSART et Caroline MEYNET : Some rates of convergence for the selected Lasso estimator. *In Algorithmic learning theory*, volume 7568 de *Lecture Notes in Comput. Sci.*, pages 17–33. Springer, Heidelberg, 2012.

[MM13] Cathy MAUGIS-RABUSSEAU et Bertrand MICHEL : Adaptive density estimation for clustering with Gaussian mixtures. *ESAIM Probab. Stat.*, 17:698–724, 2013.

[MN06] Pascal MASSART et Élodie NÉDÉLEC : Risk bounds for statistical learning. *Ann. Statist.*, 34(5):2326–2366, 2006.

[MP00] Geoffrey MCLACHLAN et David PEEL : *Finite mixture models*. Wiley Series in Probability and Statistics : Applied Probability and Statistics. Wiley-Interscience, New York, 2000.

[MT99] Enno MAMMEN et Alexandre B. TSYBAKOV : Smooth discrimination analysis. *Ann. Statist.*, 27(6):1808–1829, 1999.

[MV03] Shahar MENDELSON et Roman VERSHYNIN : Entropy and the combinatorial dimension. *Invent. Math.*, 152(1):37–55, 2003.

[MZ97] Neri MERHAV et Jacob ZIV : On the amount of statistical side information required for lossy data compression. *IEEE Trans. Inform. Theory*, 43(4):1112–1121, 1997.

[Pag98] Gilles PAGÈS : A space quantization method for numerical integration. *J. Comput. Appl. Math.*, 89(1):1–38, 1998.

[Pol81] David POLLARD : Strong consistency of k-means clustering. *Ann. Statist.*, 9(1):135–140, 1981.

[Pol82a] David POLLARD : A central limit theorem for empirical processes. *J. Austral. Math. Soc. Ser. A*, 33(2):235–248, 1982.

[Pol82b] David POLLARD : A central limit theorem for k-means clustering. *Ann. Probab.*, 10(4):919–926, 1982.

[Pol82c] David POLLARD : Quantization and the method of k -means. *IEEE Transactions on Information Theory*, 28(2):199–204, 1982.

[SB08] Douglas STEINLEY et Michael J. BRUSCO : Selection of variables in cluster analysis : an empirical comparison of eight procedures. *Psychometrika*, 73(1):125–144, 2008.

[Sim96] Hans Ulrich SIMON : General bounds on the number of examples needed for learning probabilistic concepts. *J. Comput. System Sci.*, 52(2):239–254, 1996. Sixth Annual Workshop on Computational Learning Theory (COLT) (Santa Cruz, CA, 1993).

[SWF12] Wei SUN, Junhui WANG et Yixin FANG : Regularized k-means clustering of high-dimensional data and its asymptotic consistency. *Electron. J. Stat.*, 6:148–167, 2012.

[Tal05] Michel TALAGRAND : *The generic chaining*. Springer Monographs in Mathematics. Springer-Verlag, Berlin, 2005. Upper and lower bounds of stochastic processes.

[Tar95] Thaddeus TARPEY : Principal points and self-consistent points of symmetric multivariate distributions. *J. Multivariate Anal.*, 53(1):39–51, 1995.

[TCKV10] Marieke E. TIMMERMAN, Eva CEULEMANS, Henk A. L. KIERS et Maurizio VICHI : Factorial and reduced K-means reconsidered. *Comput. Statist. Data Anal.*, 54(7):1858–1871, 2010.

[Ter12] Yoshikazu TERADA : Strong Consistency of Reduced K-means Clustering. *ArXiv e-prints*, 2012.

[Ter13] Yoshikazu TERADA : Strong Consistency of Factorial K-means Clustering. *ArXiv e-prints*, 2013.

[Tsy09] Alexandre B. TSYBAKOV : *Introduction to nonparametric estimation*. Springer Series in Statistics. Springer, New York, 2009. Revised and extended from the 2004 French original, Translated by Vladimir Zaiats.

[Vap82] Vladimir VAPNIK : *Estimation of dependences based on empirical data*. Springer Series in Statistics. Springer-Verlag, New York-Berlin, 1982. Translated from the Russian by Samuel Kotz.

[Vap00] Vladimir VAPNIK : *The nature of statistical learning theory*. Statistics for Engineering and Information Science. Springer-Verlag, New York, second édition, 2000.

[VC74] Vladimir VAPNIK et Alexey CHERVONENKIS : *Teoriya raspoznavaniya obrazov. Statisticheskie problemy obucheniya*. Izdat. “Nauka”, Moscow, 1974.

[vdG08] Sara A. van de GEER : High-dimensional generalized linear models and the lasso. *Ann. Statist.*, 36(2):614–645, 2008.

[vdG13] Sara A. van de GEER : Generic chaining and the ℓ_1-penalty. *J. Statist. Plann. Inference*, 143(6):1001–1012, 2013.

[VK01] Maurizio VICHI et Henk A. L. KIERS : Factorial k-means analysis for two-way data. *Comput. Statist. Data Anal.*, 37(1):49–64, 2001.

[WT10] Daniela M. WITTEN et Robert TIBSHIRANI : A framework for feature selection in clustering. *J. Amer. Statist. Assoc.*, 105(490):713–726, 2010.