CLUSTERING PAR QUANTIFICATION*

Documents de référence :

- sur la quantification : livre de Graff et Lushgy, "Foundations of Quantization for Probability Distributions", 2000.
- sur la méthode des k-means : document de Linder, "Learning-Theoretic Methods in Vector Quantization", 2004.

1 La problématique du clustering

Le problème

- Observations : $x_1, \cdots, x_n \in \mathscr{H}$, avec $(\mathscr{H}, \|.\|)$ Hilbert séparable
- Objectif :

  ▷ former une typologie dans la population $\{x_1, \cdots, x_n\}$, i.e. une partition de la population en k groupes (k fixé), ou types

  ▷ Exemples d'utilisation :

  - Médecine : former des groupes de patients au comportement homogène. Cas $\mathscr{H} = \mathbb{R}^d$ ou $\mathscr{H} = L_2([0,1])$;
  - Internet : prétraitement des moteurs de recherche.

Terminologie

- Typologie = *Clustering*, Type = *Cluster*

Principe de la méthode des k-means

- Erreur commise en résumant les observations à k points $(c_1, \cdots, c_k) \in \mathscr{H}^k$, appelé *k-centre* :

$$E(c_1, \cdots, c_k) := \sum_{i=1}^{n} \min_{1 \leq j \leq k} \|x_i - c_j\|^2$$

*Benoît Cadre, ENS Cachan Bretagne

- Si il existe un k-centre qui minimise cette erreur :

$$(c_1^\star, \cdots, c_k^\star) \in \arg \min_{c_1,\cdots,c_k \in \mathscr{H}} E(c_1, \cdots, c_k),$$

on note $\mathscr{P}^\star = \{A_1^\star, \cdots, A_k^\star\}$ la partition de Voronoi associée :

$$\begin{aligned} A_1^\star &= \left\{x \in \mathscr{H} : \|x - c_1^\star\| \leq \|x - c_j^\star\|, \forall j = 1, \cdots, k\right\}, \text{ et} \\ A_\ell^\star &= \left\{x \in \mathscr{H} : \|x - c_\ell^\star\| \leq \|x - c_j^\star\|, \forall j = 1, \cdots, k\right\} \setminus \bigcup_{k=1}^{\ell-1} A_k^\star, \end{aligned}$$

pour $\ell = 2, \cdots, k$.

- Dans la méthode des k-means, le ℓ-ème cluster est $A_\ell^\star \cap \{x_1, \cdots, x_n\}$.

Dans le cadre d'une modélisation statistique, il faudra se doter d'outils permettant d'évaluer les performances statistiques de la méthode. Notamment, des outils permettant d'évaluer sa stabilité face à la loi des observations ainsi que sa vitesse de convergence. Auparavant, il est nécessaire de faire un détour par l'origine théorique de la méthode, i.e. le principe (probabiliste) de la *quantification*.

2 Principe de la quantification

Le principe de *quantification* est un principe probabiliste dont l'objectif est de compresser l'information contenue dans une probabilité. On fixe dorénavant P une probabilité sur $\mathscr{H}$ d'ordre 2, i.e.

$$\int_{\mathscr{H}} \|x\|^2 P(\mathrm{d}x) < \infty.$$

Définition. *Un quantifieur q d'ordre k est une fonction mesurable $q : \mathscr{H} \to \mathscr{C} \subset \mathscr{H}$ avec $|\mathscr{C}| = k$.*

Un quantifieur q d'ordre k est donc caractérisé par :

▷ un alphabet $\mathscr{C} = \{c_1, \cdots, c_k\}$

▷ une partition $\mathscr{P} = \{A_1, \cdots, A_k\}$, avec la numérotation imposée par

$$q(x) = c_\ell \Leftrightarrow x \in A_\ell$$

On écrira donc dans la suite $q = (\mathscr{C}, \mathscr{P})$.

Un quantifieur apparaît donc comme un outil de compression de l'information. De ce fait, il faut se doter d'un outil qui va mesurer la pertinence de q en tant qu'outil de compression de l'information :

Définition. *La distorsion d'un quantifieur $q = (\mathscr{C}, \mathscr{P})$ d'ordre k est définie par :*

$$D(P,q) = \int_{\mathscr{H}} \|x - q(x)\|^2 P(\mathrm{d}x).$$

La distorsion minimale de P à l'ordre k est

$$D^\star(P) = \inf_q D(P,q),$$

l'inf étant pris sur tous les quantifieurs d'ordre k.

L'objectif est alors d'atteindre la distorsion minimale. Bien sûr, la qualité d'une quantification est d'autant meilleure que k est grand. Ce phénomène est précisé ci-dessous. On rappelle que $\mathscr{H}$ est séparable.

Proposition. *Supposons que $\mathscr{H}$ est complet, et notons $D_k^\star(P)$ la distorsion minimale à l'ordre k. Alors, $D_k^\star(P) \searrow 0$ si $k \nearrow \infty$.*

Preuve. Tout d'abord, il est clair que la distorsion minimale décroit à mesure que son ordre augmente. Puis, comme $\mathscr{H}$ est un espace Polonais, la mesure bornée μ définie pour tout borélien A de $\mathscr{H}$ par

$$\mu(A) = \int_A \|x\|^2 P(\mathrm{d}x)$$

est tendue, i.e. pour tout $\varepsilon > 0$, il existe un compact K tel que $\mu(K) \geq 1 - \varepsilon$. On note $\{c_1, c_2, \cdots\}$ un sous-ensemble dénombrable dense. Comme K est compact, il existe $k \in \mathbb{N}$ tel que

$$K \subset B := \bigcup_{i=1}^{k} B(c_i, \sqrt{\varepsilon}).$$

On a donc $\mu(B) \geq 1 - \varepsilon$. Notons maintenant q_{k+1} le quantifieur d'ordre $k+1$ d'alphabet $\{c_1, \cdots, c_k, 0\}$ et de partition $\{A_1, \cdots, A_k, B^c\}$ avec $A_1 = B(c_1, \sqrt{\varepsilon})$ et

pour $i = 2, \cdots, k : A_i = B(c_i, \sqrt{\varepsilon}) \setminus A_{i-1}$. Comme $\|x - c_i\| \leq \sqrt{\varepsilon}$ si $x \in A_i$, on a :

$$\begin{aligned} D_{k+1}^{\star}(P) \leq D_{k+1}(P, q_{k+1}) &= \int_{\mathscr{H}} \|x - q_{k+1}(x)\|^2 P(\mathrm{d}x) \\ &= \sum_{i=1}^{k} \int_{A_i} \|x - c_i\|^2 P(\mathrm{d}x) + \int_{B^c} \|x\|^2 P(\mathrm{d}x) \\ &\leq \varepsilon P\left(\bigcup_{i=1}^{k} A_i\right) + \mu(B^c) \leq 2\varepsilon, \end{aligned}$$

ce qui achève la preuve. □

La classe de quantifieurs les plus intéressants est la suivante. Dans la suite, on suppose que les quantifieurs sont d'ordre k et on note, pour un alphabet $\mathscr{C} \subset \mathscr{H}$ de taille k, $\mathscr{P}_{\mathrm{V}}(\mathscr{C})$ la partition de Voronoi associée à $\mathscr{C}$.

Définition. *Un quantifieur d'ordre k est un quantifieur de type plus proches voisins (PPV) si sa partition est une partition de Voronoi associée à son alphabet. En d'autres termes, un quantifieur PPV s'écrit $q = (\mathscr{C}, \mathscr{P}_{\mathrm{V}}(\mathscr{C}))$, avec $\mathscr{C} \subset \mathscr{H}$ de taille finie.*

Ainsi, un quantifieur PPV est caractérisé par son alphabet. On notera les propriétés élémentaires suivantes :

Proposition. *Soit q_{ppv} un quantifieur PPV d'alphabet $\mathscr{C} = \{c_1, \cdots, c_k\}$. Alors,*

$$D(P, q_{\mathrm{ppv}}) = \int_{\mathscr{H}} \min_{1 \leq \ell \leq k} \|x - c_\ell\|^2 P(\mathrm{d}x),$$

et de plus, pour tout quantifieur $q = (\mathscr{C}, \mathscr{P})$, on a $D(P, q_{\mathrm{ppv}}) \leq D(P, q)$.

Preuve. Pour la première propriété, on a si $\mathscr{P}_{\mathrm{V}}(\mathscr{C}) = \{A_{\mathrm{V},1}, \cdots, A_{\mathrm{V},k}\}$:

$$\begin{aligned} D(P, q_{\mathrm{ppv}}) &= \int_{\mathscr{H}} \|x - q_{\mathrm{ppv}}(x)\|^2 P(\mathrm{d}x) = \sum_{j=1}^{k} \int_{A_{\mathrm{V},j}} \|x - c_j\|^2 P(\mathrm{d}x) \\ &= \int_{\mathscr{H}} \min_{1 \leq \ell \leq k} \|x - c_\ell\|^2 P(\mathrm{d}x) \end{aligned}$$

Puis, pour la 2nde propriété, si $\mathscr{P} = \{A_1, \cdots, A_k\}$:

$$\begin{aligned} D(P, q_{\text{ppv}}) &= \sum_{j=1}^{k} \int_{A_j} \min_{1 \le \ell \le k} \|x - c_\ell\|^2 P(\mathrm{d}x) \\ &\le \sum_{j=1}^{k} \int_{A_j} \|x - c_j\|^2 P(\mathrm{d}x) \\ &\le \int_{\mathscr{H}} \|x - q(x)\|^2 P(\mathrm{d}x) = D(P, q), \end{aligned}$$

par définition de la distorsion. □

La conséquence importante est que les quantifieurs de distorsion minimale, s'ils existent, sont à chercher parmi les quantifieurs du type $q_{\text{ppv}} = (\mathbf{c}, \mathscr{P}_{\text{V}}(\mathbf{c}))$ avec $\mathbf{c} = (c_1, \cdots, c_k) \in \mathscr{H}^k$ (noter l'abus de notation) de distorsion :

$$W(P, \mathbf{c}) := \int_{\mathscr{H}} \min_{1 \le j \le k} \|x - c_j\|^2 P(\mathrm{d}x) = D(P, q_{\text{ppv}})$$

Théorème. *Il existe un quantifieur de distorsion minimale.*

Preuve. On montre qu'il existe $\mathbf{c}^\star \in \mathscr{H}^k$ tel que

$$W(P, \mathbf{c}^\star) = \inf_{\mathbf{c} \in \mathscr{H}^k} W(P, \mathbf{c})$$

en 3 étapes :

1. $\mathscr{H}^k \ni \mathbf{c} \mapsto W(P, \mathbf{c})$ est faiblement s.c.i.

2. il existe $R > 0$ t.q.

$$\inf_{\mathbf{c} \in \mathscr{H}^k} W(P, \mathbf{c}) = \inf_{\|\mathbf{c}\|_{\mathscr{H}^k} \le R} W(P, \mathbf{c}).$$

Cette propriété est démontrée dans le livre de Graf et Luschgy (Foundation of Quantization for Probability Distributions, 2000). La preuve, donnée dans le cas d'un espace d'observation du type $\mathbb{R}^d$, s'adapte à notre cas.

3. Conclusion.

Preuve de 1. :

- $x \mapsto \|x - c_i\|$ convexe + continue $\Rightarrow$ faiblement s.c.i.

$$\Leftrightarrow \{x \in \mathscr{H} \,:\, \|x - c_i\| \leq t\} \text{ faiblement fermé } \forall t$$

- $x \mapsto \min_{1\leq i\leq k} \|x - c_i\|$ faiblement s.c.i

$$\Leftrightarrow \{x \in \mathscr{H} \,:\, \min_{1\leq i\leq k} \|x - c_i\| \leq t\} = \bigcup_{i=1}^{k} \{x \in \mathscr{H}^k \,:\, \|x - c_i\| \leq t\}$$

  est faiblement fermé

- pour $\mathbf{c}_0 = (c_{1,0}, \cdots, c_{k,0})$:

$$\begin{aligned} \liminf_{\mathbf{c}\to\mathbf{c}_0} W(P,\mathbf{c}) &\geq \int_{\mathscr{H}} \liminf_{\mathbf{c}\to\mathbf{c}_0} \min_{1\leq j\leq k} \|x - c_j\|^2 P(\mathrm{d}x) \text{ [Fatou]} \\ &\geq \int_{\mathscr{H}} \min_{1\leq j\leq k} \|x - c_{j,0}\|^2 P(\mathrm{d}x) \text{ [s.c.i.]} \\ &= W(P,\mathbf{c}_0) \end{aligned}$$

  donc $W(P,.)$ est faiblement s.c.i.

Conclusion :

- $B_{\mathscr{H}^k}(0,R)$ faiblement compacte + $W(P,.)$ faiblement s.c.i. $\Rightarrow$ il existe $\mathbf{c}^\star \in \mathscr{H}^k$ minimum de $W(P,.)$
- $q^\star = (\mathbf{c}^\star, \mathscr{P}_{\text{ppv}}(\mathbf{c}^\star))$ est un quantifieur de distorsion minimale car

$$W(P,\mathbf{c}^\star) = \inf_{\mathscr{H}^k} W(P,.) = \inf_{q} D(P,q) = D^\star(P).$$

□

3 Clustering par quantification

3.1 Principe général

Le contexte et les outils utilisés dorénavant sont précisés ci-dessous :

- $x_1, \cdots, x_n \in \mathscr{H}$ réalisations de v.a. $X_1, \cdots, X_n$ i.i.d. de loi P (d'ordre 2)
- P_n désigne la mesure empirique des observations i.e.

$$P_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i}.$$

- Distorsion empirique du quantifieur q (d'ordre k) :

$$D(P_n, q) = \int_{\mathscr{H}} \|x - q(x)\|^2 P_n(\mathrm{d}x) = \frac{1}{n} \sum_{i=1}^{n} \|X_i - q(X_i)\|^2$$

- Distorsion empirique de $q_{\mathrm{ppv}} = (\mathbf{c}, \mathscr{P}_{\mathrm{V}}(\mathbf{c}))$, avec $\mathbf{c} = (c_1, \cdots, c_k) \in \mathscr{H}^k$:

$$D(P_n, q_{\mathrm{ppv}}) = W(P_n, \mathbf{c}) = \frac{1}{n} \sum_{i=1}^{n} \min_{1 \leq j \leq k} \|X_i - c_j\|^2$$

Principe général d'une méthode de clustering par quantification

▷ Trouver un quantifieur empirique $\hat{q} = (\hat{\mathscr{C}}, \hat{\mathscr{P}})$

▷ Les clusters sont $\hat{A} \cap \{X_1, \cdots, X_n\}$, $\hat{A} \in \hat{\mathscr{P}}$

Pour se doter d'outils qui assurent que la méthode de quantification est performante, on introduit la définition qui suit :

Définitions. *Soit $\hat{q}$ un quantifieur empirique. On dit qu'il est*

▷ *consistant, si $\mathbb{E}D(P, \hat{q}) \to D^\star(P)$*

▷ *de vitesse $(v_n)_n$ si $\mathbb{E}D(P, \hat{q}) - D^\star(P) = O(1/v_n)$, avec $v_n \to \infty$*

On aura noté au passage que, puisque $D(P, \hat{q}) \geq D^\star(P)$, la propriété $\mathbb{E}D(P, \hat{q}) \to D^\star(P)$ est équivalente à $D(P, \hat{q}) \to D^\star(P)$ dans L^1.

3.2 La méthode des k-means

Principe de la méthode

▷ calcul des centres optimaux $\hat{\mathbf{c}} = (\hat{c}_1, \cdots, \hat{c}_k)$ tels que

$$W(P_n, \hat{\mathbf{c}}) = \min_{\mathbf{c} \in \mathscr{H}^k} W(P_n, \mathbf{c}) \qquad (\star)$$

▷ si $\hat{A}_\ell$ est le ℓ-ème élément de $\mathscr{P}_{\mathrm{V}}(\hat{\mathbf{c}})$, le ℓ-ème cluster est constitué des $\{X_1, \cdots, X_n\} \cap \hat{A}_\ell$ i.e. des observations X_i telles que

$$\|X_i - \hat{c}_\ell\| \leq \|X_i - \hat{c}_j\|, \quad \forall j = 1, \cdots, k$$

Un des avantages de cette méthode est qu'elle ne s'appuie pas sur un calcul de la partition de Voronoi (ce qui est numériquement infaisable, mème pour des dimensions relativement petites). En revanche, son écueil principal est que l'étape de minimisation est difficile à mettre en oeuvre numériquement, surtout en grande dimension. Pour cette étape de minimisation, la méthode standard, appelée "itération de Llyod", est basée sur la remarque suivante :

Condition des centres. Pour une partition $\mathscr{P} = \{A_1, \cdots, A_k\}$ et $\mathbf{c} = (c_1, \cdots, c_k) \in \mathscr{H}^k$, on note $q = (\mathbf{c}, \mathscr{P})$ et $\hat{q}' = (\hat{\mathbf{c}}', \mathscr{P})$ avec $E(\mathbf{c}) := \mathbf{c}' = (c'_1, \cdots, c'_k)$ tel que

$$c'_j = \arg\min_{y \in \mathscr{H}} \sum_{i=1}^{n} \|X_i - y\|^2 \mathbf{1}\{X_i \in A_j\}$$

Noter que c'_j est, à un facteur près, une espérance conditionnelle pour la mesure empirique. On a alors :

$$\begin{aligned} D(P_n, q) &= \frac{1}{n} \sum_{j=1}^{k} \sum_{i=1}^{n} \|X_i - c_j\|^2 \mathbf{1}\{X_i \in A_j\} \\ &\geq \frac{1}{n} \sum_{j=1}^{k} \sum_{i=1}^{n} \|X_i - c'_j\|^2 \mathbf{1}\{X_i \in A_j\} \\ &= D(P_n, q'). \end{aligned}$$

En d'autres termes, cette opération permet de faire décroitre la distorsion.

La remarque précédente nous donne une technique de minimisation dans la méthode des k-means :

Calcul numérique de $\hat{\mathbf{c}}$ défini par $(\star)$. Dans la méthode des k-means, le calcul numérique de $\hat{\mathbf{c}}$ est effectué par décroissance de la distorsion utilisant la remarque qui précède. C'est l'" Itération de Llyod" : de l'itération ℓ à l'itération $\ell+1$, on passe d'un k-centre à un autre comme suit

$$k-\text{centre } \mathbf{c}_\ell \to \text{partition Voronoi associée } \to k-\text{centre } \mathbf{c}_{\ell+1} = E(\mathbf{c}_\ell).$$

Cependant, même s'il est assuré que la distorsion décroît entre 2 itérations, rien ne nous dit que l'algorithme est convergent ...

4 Consistance de la méthode des k-means

L'outil indispensable pour établir la consistance de la méthode des k-means est la distance de Wasserstein :

Définition. *La distance de Wasserstein ρ_W est définie pour ν_1, ν_2 proba sur $\mathscr{H}$ d'ordre 2 par :*

$$\rho_W(\nu_1,\nu_2) = \inf_{X\sim\nu_1, Y\sim\nu_2} \sqrt{\mathbb{E}\|X-Y\|^2}.$$

Il s'agit d'une distance usuelle en probabilité. Mentionnons 2 de ses propriétés fondamentales (on renvoie au livre de Dudley, "Real Analysis and Probability", pour les preuves) :

Propriétés.

1. *Soient ν_n, ν des probabilités d'ordre 2 sur $\mathscr{H}$. On a $\rho_W(\nu_n,\nu) \to 0$ si*

$$\nu_n \Rightarrow \nu \quad \text{et} \quad \int_{\mathscr{H}} \|x\|^2 \nu_n(\mathrm{d}x) \to \int_{\mathscr{H}} \|x\|^2 \nu(\mathrm{d}x).$$

2. *Pour ν_1, ν_2 des probabilités d'ordre 2 sur $\mathscr{H}$, il existe (X_0, Y_0) tel que $X_0 \sim \nu_1$ et $Y_0 \sim \nu_2$ tel que*

$$\rho_W(\nu_1,\nu_2) = \sqrt{\mathbb{E}\|X_0 - Y_0\|^2}.$$

Le lien entre l'étude qui nous intéresse et la distance de Wasserstein est établi ci-dessous :

Proposition. *Soient ν_1, ν_2 des probabilités d'ordre 2 sur $\mathscr{H}$. Si q est PPV, alors*

$$|D(\nu_1,q)^{1/2} - D(\nu_2,q)^{1/2}| \leq \rho_W(\nu_1,\nu_2).$$

Preuve. Soit (X_0, Y_0) tel que $X_0 \sim \nu_1$ et $Y_0 \sim \nu_2$ tel que

$$\rho_W(\nu_1,\nu_2) = \sqrt{\mathbb{E}\|X_0 - Y_0\|^2}.$$

Si $q = (\mathbf{c}, \mathscr{P}_{\mathrm{V}}(\mathbf{c}))$:

$$\begin{aligned} D(\nu_1,q)^{1/2} = W(\nu_1,\mathbf{c})^{1/2} &= \sqrt{\mathbb{E}\min_{1\leq j\leq k}\|X_0 - c_j\|^2} \\ &\leq \sqrt{\mathbb{E}\min_{1\leq j\leq k}(\|X_0 - Y_0\| + \|Y_0 - c_j\|)^2} \\ &\leq \sqrt{\mathbb{E}\|X_0 - Y_0\|^2} + \sqrt{\mathbb{E}\min_{1\leq j\leq k}\|Y_0 - c_j\|^2} \\ &= \rho_W(\nu_1,\nu_2) + D(\nu_2,q)^{1/2}, \end{aligned}$$

d'où la proposition. □

On fixe dorénavant les quantités issues de la méthode des k-means, i.e. $\hat{\mathbf{c}} = (\hat{c}_1, \cdots, \hat{c}_k)$ un minimum de $W(P_n, .)$ défini par $(\star)$:

$$\hat{q} = (\hat{\mathbf{c}}, \mathscr{P}_{\mathrm{V}}(\hat{\mathbf{c}}))$$

Théorème. *La méthodes des k-means est consistante, i.e. $\mathbb{E}D(P,\hat{q}) \to D^\star(P)$.*

Preuve. Si $q^\star$ est un quantifieur optimal PPV pour P, on a avec la proposition précédente :

$$\begin{aligned} &D(P,\hat{q})^{1/2} - D^\star(P)^{1/2} \\ &\quad = [D(P,\hat{q})^{1/2} - D(P_n,\hat{q})^{1/2}] + [D(P_n,\hat{q})^{1/2} - D(P,q^\star)^{1/2}] \\ &\quad \leq [D(P,\hat{q})^{1/2} - D(P_n,\hat{q})^{1/2}] + [D(P_n,q^\star)^{1/2} - D(P,q^\star)^{1/2}] \\ &\quad \leq 2\rho_W(P,P_n). \end{aligned}$$

Or, $\rho_W(P_n,P) \to 0$ p.s. car $\mathbb{P}(P_n \Rightarrow P) = 1$ (Th. Varadarajan) et p.s.

$$\int_{\mathscr{H}} \|x\|^2 P_n(\mathrm{d}x) \to \int_{\mathscr{H}} \|x\|^2 P(\mathrm{d}x).$$

On a donc $D(P,\hat{q}^\star) \to D^\star(P)$ p.s. i.e. $\hat{q}$ est consistant. $\square$

Comme d'habitude en statistique, cette propriété ne doit être vue que comme le minimum que toute méthode raisonnable doit vérifier.

5 Vitesse de convergence dans la méthode des k-means

L'hypothèse fondamentale du résultat de cette section est la *contrainte de pic*, qui exprime le fait que le support de la loi P est borné. Elle amène 2 commentaires :

- La contrainte de pic est classique en apprentissage, et plus généralement en statistique, car seules un nombre fini de données sont récoltées. Mais, il en résulte un manque d'unité entre la modélisation probabiliste et l'utilisation statistique du modèle :

  ▷ Modélisation probabiliste (signal, ...) mène souvent à des diffusions browniennes, pour lesquelles P n'est pas à support borné.

  ▷ Le traitement statistique se situe en aval de cette modélisation. Mais X est supposée bornée ...

- Quelle caractéristique de P remplace R dans l'inégalité du thèorème ci-dessous ?

Théorème. *Si* $\mathrm{supp}(P) \subset B(0,R)$,

$$\mathbb{E}D(P,\hat{q}) - D^\star(P) \leq 36k\frac{R^2}{\sqrt{n}}$$

On rappelle les notations suivantes :

- pour Q une mesure signée sur $\mathscr{H}$ et $\mathscr{F}$ un ensemble de fonctions réelles définies sur $\mathscr{H}$:
$$\|Q\|_{\mathscr{F}} = \sup_{f\in\mathscr{F}} |Q(f)|;$$

- si $\sigma_1,\cdots,\sigma_n$ désigne une suite de v.a. de Rademacher i.i.d. indépendantes des observations :
$$P_n^\sigma = \frac{1}{n}\sum_{i=1}^n \sigma_i\,\delta_{X_i}, \quad \text{la mesure empirique symétrisée.}$$

On mentionne tout d'abord un outil fondamental dans l'étude de la mesure empirique (Théorème 4.12 du livre de Ledoux et Talagrand, "Probability in Banach Spaces", 1991) :

Lemme. [Principe de contraction] *Soit $\mathscr{F}$ un ensemble de fonctions réelles définies sur $\mathscr{H}$. Si $|\mathscr{F}| = \{|f| : f \in \mathscr{F}\}$, on a*

$$\mathbb{E}\|P_n^\sigma\|_{|\mathscr{F}|} \leq 2\mathbb{E}\|P_n^\sigma\|_{\mathscr{F}}.$$

Remarques préliminaires :

- Si $\text{supp}(P) \subset B(0,R)$, alors les centres optimaux sont dans $B_R = B(0,R)$. En effet, si $\|c\| > R$ et p est la projection orthogonale sur B_R alors, par définition de la projection orthogonale, on a $\forall x \in B_R$:

$$\begin{aligned} \|x-c\|^2 &= \|x-p(c)\|^2 + \|p(c)-c\|^2 - 2\langle x-p(c), c-p(c)\rangle \\ &\geq \|x-p(c)\|^2. \end{aligned}$$

  On a donc une distorsion plus petite pour des centres dans B_R.

- Si $X \sim P$:

$$\begin{aligned} W(P,\mathbf{c}) &= \mathbb{E}\min_{1\leq j\leq k}\|X-c_j\|^2 \\ &= \mathbb{E}\|X\|^2 + \mathbb{E}\min_{1\leq j\leq k}\left[-2\langle X, c_j\rangle + \|c_j\|^2\right]. \end{aligned}$$

Ces 2 observations nous conduisent à la conclusion suivante : plutôt que de minimiser $W(P,.)$ sur $\mathscr{H}^k$, il suffit donc de minimiser, sur B_R^k :

$$\bar{W}(P,\mathbf{c}) = \mathbb{E}\min_{1\leq j\leq k} f_{c_j}(X), \text{ si } f_c(x) = -2\langle x, c\rangle + \|c\|^2.$$

La même observation est valable pour P_n au lieu de P.

Preuve. En utilisant la notation générique $\mathbf{c} = (c_1, \cdots, c_k) \in \mathscr{H}^k$:

$$\begin{aligned}
D(P,\hat{q}) - D^\star(P) &= W(P,\hat{\mathbf{c}}) - \inf_{\mathbf{c}\in B_R^k} W(P,\mathbf{c}) \\
&= \bar{W}(P,\hat{\mathbf{c}}) - \inf_{\mathbf{c}\in B_R^k} \bar{W}(P,\mathbf{c}) \\
&\leq [\bar{W}(P,\hat{\mathbf{c}}) - \bar{W}(P_n,\hat{\mathbf{c}})] + \big[\inf_{\mathbf{c}\in B_R^k} \bar{W}(P_n,\mathbf{c}) - \inf_{\mathbf{c}\in B_R^k} \bar{W}(P,\mathbf{c})\big] \\
&\leq 2 \sup_{\mathbf{c}\in B_R^k} \big|\bar{W}(P_n,\mathbf{c}) - \bar{W}(P,\mathbf{c})\big| \\
&= 2 \sup_{\mathbf{c}\in B_R^k} \frac{1}{n}\Big| \sum_{i=1}^n \big(\min_{1\leq j\leq k} f_{c_j}(X_i) - \mathbb{E} \min_{1\leq j\leq k} f_{c_j}(X)\big)\Big|.
\end{aligned}$$

D'après le théorème de symétrisation en moyenne (cf. Chapitre 2) :

$$\begin{aligned}
\mathbb{E}D(P,\hat{q}) - D^\star(P) &\leq 2\mathbb{E} \sup_{\mathbf{c}\in B_R^k} \frac{1}{n}\Big| \sum_{i=1}^n \big(\min_{1\leq j\leq k} f_{c_j}(X_i) - \mathbb{E} \min_{1\leq j\leq k} f_{c_j}(X)\big)\Big| \\
&\leq 4\mathbb{E} \sup_{\mathbf{c}\in B_R^k} \frac{1}{n}\Big| \sum_{i=1}^n \sigma_i \min_{1\leq j\leq k} f_{c_j}(X_i)\Big|.
\end{aligned}$$

Pour le traitement du dernier terme, nous allons procéder par itération sur k, en s'appuyant sur le principe de contraction. On note :

$$S_k := \mathbb{E} \sup_{(c_1,\cdots,c_k)\in B_R^k} \frac{1}{n}\Big| \sum_{i=1}^n \sigma_i \min_{1\leq j\leq k} f_{c_j}(X_i)\Big|.$$

Cas $k = 1$. Comme $\|X\| \leq R$:

$$\begin{aligned}
S_1 &= \mathbb{E} \sup_{c\in B_R} \frac{1}{n}\Big| \sum_{i=1}^n \sigma_i\big(-2\langle X_i, c\rangle + \|c\|^2\big)\Big| \\
&\leq 2\mathbb{E} \sup_{c\in B_R} \frac{1}{n}\Big| \sum_{i=1}^n \sigma_i\langle X_i, c\rangle\Big| + \mathbb{E} \sup_{c\in B_R} \frac{\|c\|^2}{n}\Big| \sum_{i=1}^n \sigma_i\Big| \\
&\leq 2\mathbb{E} \sup_{c\in B_R} \frac{1}{n}\Big| \sum_{i=1}^n \sigma_i\langle X_i, c\rangle\Big| + \frac{R^2}{n}\mathbb{E}\Big| \sum_{i=1}^n \sigma_i\Big| \\
&\leq 2\frac{R}{n}\mathbb{E}\| \sum_{i=1}^n \sigma_i X_i\| + \frac{R^2}{\sqrt{n}} \\
&\leq 2R\sqrt{\frac{\mathbb{E}\|X\|^2}{n}} + \frac{R^2}{\sqrt{n}} \leq \frac{3R^2}{\sqrt{n}}.
\end{aligned}$$

Cas $k=2$. Comme $\min(a,b) = (a+b)/2 - |a-b|/2$ pour $a,b \in \mathbb{R}$:

$$\begin{aligned} S_2 &= \mathbb{E} \sup_{(c_1,c_2)\in B_R^2} \frac{1}{2n} \Big| \sum_{i=1}^n \sigma_i (f_{c_1}(X_i) + f_{c_2}(X_i) - \big|f_{c_1}(X_i) - f_{c_2}(X_i)\big|) \Big| \\ &\leq S_1 + \mathbb{E} \sup_{(c_1,c_2)\in B_R^2} \frac{1}{2n} \Big| \sum_{i=1}^n \sigma_i \big| f_{c_1}(X_i) - f_{c_2}(X_i) \big| \Big|. \end{aligned}$$

En appliquant le principe de contraction, on obtient :

$$\begin{aligned} S_2 &\leq S_1 + \mathbb{E} \sup_{(c_1,c_2)\in B_R^2} \frac{1}{n} \Big| \sum_{i=1}^n \sigma_i \big(f_{c_1}(X_i) - f_{c_2}(X_i) \big) \Big| \\ &\leq 3S_1. \end{aligned}$$

Cas $k=3$. Comme $S_2 \leq 3S_1$,

$$\begin{aligned} S_3 &\leq \frac{S_1+S_2}{2} + S_1 + S_2 \\ &\leq 6S_1. \end{aligned}$$

En itérant le procédé, on trouve :

$$S_k \leq 3kS_1 \leq 9k\frac{R^2}{\sqrt{n}}.$$

Finalement :

$$\mathbb{E}D(P,\hat{q}^\star) - D^\star(P) \leq 4S_k \leq 36k\frac{R^2}{\sqrt{n}},$$

d'où le théorème. □