MP 2 **Samedi 23 Mars 2024**

Devoir surveillé 6 (4h)

Exercice : caractérisation des fonctions homogènes de degré k

Soient une fonction $f : \mathbb{R}^n \setminus \{0\} \to \mathbb{R}$, différentiable en tout point, et k une constante réelle.

a) On suppose que f est homogène de degré k, ce qui signifie par définition

$$\forall\, t > 0,\ \forall\, x \in \mathbb{R}^n \setminus \{0\}, \quad f(tx) = t^k f(x) \tag{E1}$$

En dérivant la fonction $t \mapsto t^{-k} f(tx)$ montrer que f vérifie la relation suivante due à Euler :

$$\forall\, x \in \mathbb{R}^n \setminus \{0\}, \quad \sum_{i=1}^{n} x_i \frac{\partial f}{\partial x_i}(x) = k f(x) \tag{E2}$$

b) Montrer que la réciproque du a) est vraie, en remarquant qu'on peut appliquer (E2) à des variables qu'on note tx à la place de x.

c) Bonus, un peu de chimie. L'enthalpie libre G est une fonction homogène de degré 1 des différents n_i (nombre de moles des différents constituants). Quelle relation donne alors l'identité d'Euler (E2) ?

Problème : ε-isométries et probabilités

Notations :

— Dans tout le problème N, k et d désignent des entiers supérieurs ou égaux à deux.

— Pour tous entiers naturels non nuls p et q, on note $\mathcal{M}_{p,q}(\mathbb{R})$ l'ensemble des matrices à p lignes et q colonnes à coefficients réels.

— On note $A^\top$ la transposée d'une matrice A.

— Pour tous entiers naturels p et q, avec $p \leqslant q$, la notation $[\![p, q]\!]$ désigne l'ensemble $\{i \in \mathbb{N} \mid p \leqslant i \leqslant q\}$.

— Dans tout le problème on note $(\Omega, \mathcal{A}, \mathbb{P})$ un espace probabilisé **fini**. Toutes les variables aléatoires considérées sont définies sur Ω.

— Pour tout événement A de probabilité non nulle, et pour tout événement B, on note $\mathbb{P}_A(B)$ ou $\mathbb{P}(B \mid A)$ la probabilité conditionnelle de B sachant A.

— Étant donnée une variable aléatoire Z à valeurs réelles, on note $\mathbb{E}(Z)$ son espérance.

— On dit qu'une variable aléatoire Z est une variable de Rademacher lorsque $Z(\Omega) = \{-1, 1\}$ et

$$\mathbb{P}(Z = -1) = \mathbb{P}(Z = 1) = \frac{1}{2}$$

— De façon générale, si E est un espace euclidien, son produit scalaire et sa norme seront respectivement notés $\langle \cdot \mid \cdot \rangle$ et $\|\cdot\|$. Ces notations seront utilisées notamment pour $\mathbb{R}^d$ et $\mathbb{R}^k$, munis de leurs structures euclidiennes canoniques.

Problématique :

On s'intéresse à la question suivante : étant donnés N points dans un espace euclidien de grande dimension, est-il possible de les envoyer linéairement dans un espace de petite dimension sans trop modifier les distances entre ces points ? Pour préciser cette question, considérons N vecteurs distincts $v_1, \dots, v_N$ dans $\mathbb{R}^d$. Pour tout réel ε tel que $0 < \varepsilon < 1$, on dit qu'une application linéaire $f : \mathbb{R}^d \to \mathbb{R}^k$ est une ε-isométrie pour $v_1, \dots, v_N$ lorsque :

$$\forall (i,j) \in [\![1, N]\!]^2, \quad (1-\varepsilon)\, \|v_i - v_j\| \leqslant \|f(v_i) - f(v_j)\| \leqslant (1+\varepsilon)\, \|v_i - v_j\|$$

La question peut se reformuler ainsi : Pour quelles valeurs de k existe-t-il $f : \mathbb{R}^d \to \mathbb{R}^k$ qui soit une ε-isométrie pour $v_1, \ldots, v_N$?

Ce problème, adapté d'un sujet de Centrale conseillé par Ylan à cause de l'actualité, se propose d'établir le théorème suivant, démontré par William B. Johnson et Joram Lindenstrauss en 1984 :

Il existe une constante absolue c strictement positive telle que : quels que soient N et d, entier naturels supérieurs ou égaux à 2 et quels que soient $v_1, \ldots, v_N$ distincts dans $\mathbb{R}^d$, il suffit que

$$k \geqslant c\frac{\ln(N)}{\varepsilon^2}$$

pour qu'il existe une ε-isométrie $f : \mathbb{R}^d \to \mathbb{R}^k$ pour $v_1, \ldots, v_N$.

Les seules méthodes connues à ce jour pour démontrer ce théorème sont de nature probabiliste. Dans la partie I, on établit des résultats préliminaires portant sur la convexité et les probabilités. La partie II est consacrée à la démonstration (ici tronquée pour garder une taille de problème moins déraisonnable) d'une inégalité de concentration, due à Michel Talagrand, prix Abel 2024, qui est utilisée dans la partie III où le théorème de Johnson-Lindenstrauss est démontré.

Partie I

Projection sur un convexe fermé d'un espace euclidien :

Soit E un espace euclidien.

Q1) Soient a et b dans E. Montrer la relation suivante et en donner une interprétation géométrique dans un parallélogramme :

$$\|a+b\|^2 + \|a-b\|^2 = 2\left(\|a\|^2 + \|b\|^2\right)$$

Q2) En déduire que si u, v et v' dans E vérifient $v \neq v'$ et $\|u-v\| = \|u-v'\|$ alors $\left\|u - \frac{v+v'}{2}\right\| < \|u-v\|$.

Q3) Soient F un fermé non vide de E et u dans E. Montrer qu'il existe v dans F tel que

$$\forall w \in F, \quad \|u-v\| \leqslant \|u-w\|$$

On pourra noter $d = \inf\{||u-w||,\ w \in F\}$ et considérer une suite $(w_n) \in F^{\mathbb{N}}$ telle que $||u - w_n|| \underset{n\to+\infty}{\longrightarrow} d$.

Q4) En déduire que si C est un convexe fermé non vide de E et u est un vecteur de E alors il existe un unique v dans C tel que

$$\forall w \in C, \quad \|u-v\| \leqslant \|u-w\|$$

Définition : On dira que v est le projeté de u sur C et on notera $d(u, C) = \|u-v\|$. Autrement dit, on a défini une application $p : E \to C$ appelée *projection* sur le convexe fermé C. Bien sûr dans le cas où C est un sous-espace vectoriel de E, p est la projection orthogonale sur ce sous-espace.

Inégalité de Hölder pour l'espérance

Soient p et q deux réels strictement positifs tels que $\frac{1}{p} + \frac{1}{q} = 1$.

Q5) Montrer que, pour tous réels positifs a et b,

$$ab \leqslant \frac{a^p}{p} + \frac{b^q}{q}$$

Q6) En déduire que si X et Y sont deux variables aléatoires réelles sur un espace probabilisé **fini** $(\Omega, \mathcal{A}, \mathbb{P})$ alors

$$\mathbb{E}(|XY|) \leqslant \mathbb{E}\left(|X|^p\right)^{1/p} \mathbb{E}\left(|Y|^q\right)^{1/q}$$

On pourra d'abord montrer ce résultat lorsque $\mathbb{E}\left(|X|^p\right) = \mathbb{E}\left(|Y|^q\right) = 1$.

Espérance conditionnelle

Soit $X : \Omega \to \mathbb{R}$ une variable aléatoire à valeurs réelles défini sur un univers **fini** Ω. Pour tout événement $A \subset \Omega$ de probabilité non nulle, l'espérance conditionnelle de X sachant A, notée $\mathbb{E}(X \mid A)$, est par définition le réel

$$\mathbb{E}(X \mid A) = \sum_{x \in X(\Omega)} \mathbb{P}_A(X = x) \cdot x$$

En d'autres termes, $\mathbb{E}(X \mid A)$ est l'espérance de X dans l'espace $(\Omega, \mathcal{A}, \mathbb{P}_A)$. Les propriétés usuelles de linéarité et de positivité de l'espérance, qu'on ne demande pas de redémontrer, sont ainsi valables pour l'espérance conditionnelle sachant A.

Q7) Soit $(A_1, \ldots, A_m)$ un système complet d'événements de probabilités non nulles. Montrer que

$$\mathbb{E}(X) = \sum_{i=1}^{m} \mathbb{P}(A_i) \cdot \mathbb{E}(X \mid A_i)$$

Variables aléatoires à queue sous-gaussienne

Soit $X : \Omega \to \mathbb{R}$ une variable aléatoire réelle définie sur un univers fini Ω. On suppose qu'il existe deux réels strictement positifs a et b tels que, pour tout réel positif t,

$$\mathbb{P}(|X| \geqslant t) \leqslant a \exp\left(-bt^2\right)$$

Q8) Le but de cette question est de montrer que

$$\mathbb{E}\left(X^2\right) = 2 \int_0^{+\infty} t\, \mathbb{P}(|X| \geqslant t) \mathrm{d}t$$

Pour cela, on note : $X^2(\Omega) = \{y_1, \ldots, y_n\}$ avec $0 \leqslant y_1 < y_2 < \cdots < y_n$.

a) Tracer le graphe de la fonction $t \mapsto \mathbb{P}(X^2 \geq t)$ pour $t \in \mathbb{R}^+$ en plaçant les points $y_1, \ldots, y_n$ en abscisse, et justifier qu'il s'agit d'une fonction continue par morceaux.

b) En déduire que :

$$\int_0^{+\infty} P\left(X^2 \geqslant t\right) \mathrm{d}t = y_1 + \sum_{i=1}^{n-1} P\left(X^2 \geqslant y_{i+1}\right)(y_{i+1} - y_i)$$

c) En déduire que :

$$\int_0^{+\infty} P\left(X^2 \geqslant t\right) \mathrm{d}t = \mathbb{E}(X^2)$$

d) En déduire la formule annoncée en début de question.

Q9) Montrer que le moment d'ordre deux de X est inférieur ou égal à $\frac{a}{b}$.

Soit δ un réel tel que $0 \leqslant |\delta| \leqslant \sqrt{\frac{a}{b}}$.

Q10) Justifier que, pour tout réel t,

$$\mathbb{P}(|X + \delta| \geqslant t) \leqslant \mathbb{P}(|X| \geqslant t - |\delta|)$$

Q11) Montrer que, pour tout réel t,

$$-b(t - |\delta|)^2 \leqslant a - \frac{1}{2} bt^2$$

Q12) En déduire que pour tout réel t tel que $t \geqslant |\delta|$ on a

$$\mathbb{P}(|X + \delta| \geqslant t) \leqslant a \exp(a) \exp\left(-\frac{1}{2} bt^2\right)$$

N.B on admettra que l'inégalité précédente reste valable si $0 \leqslant t < |\delta|$.

Partie II L'inégalité de concentration de Talagrand

Soit E un espace euclidien de dimension $n \geqslant 1$ muni d'une base orthonormée $(e_1, \dots, e_n)$. Soient $\varepsilon_1, \dots, \varepsilon_n : \Omega \to \{-1, 1\}$ des variables aléatoires de Rademacher indépendantes dans leur ensemble. On pose

$$X = \sum_{i=1}^{n} \varepsilon_i e_i.$$

L'objectif de cette partie est de montrer, pour tout convexe fermé non vide C de E,

$$\mathbb{P}(X \in C) \cdot \mathbb{E}\left(\exp\left(\frac{1}{8} d(X, C)^2\right)\right) \leqslant 1 \tag{II.1}$$

II A Étude de deux cas particuliers

Q13) Traiter le cas facile où C est un convexe fermé de E ne rencontrant pas $X(\Omega)$.

On suppose, dans la suite de cette sous-partie II.A uniquement, que C est un convexe fermé de E qui rencontre $X(\Omega)$ en un seul vecteur u.

Q14) Montrer que la variable aléatoire $\frac{1}{4}\|X - u\|^2$ suit une loi binomiale $\mathcal{B}(n, 1/2)$.

Q15) En déduire l'espérance de $\exp\left(\frac{1}{8}\|X - u\|^2\right)$ et montrer qu'elle est inférieure ou égale à 2^n.

Q16) Justifier que $d(X, C) \leqslant \|X - u\|$ et en déduire l'inégalité (II.1) dans ce cas.

II.B - Initialisation

On suppose désormais que C est un convexe fermé de E tel que $C \cap X(\Omega)$ contient au moins deux éléments. Quitte à permuter les vecteurs de la base, on peut supposer que ces deux vecteurs diffèrent par leur dernière coordonnée. On va montrer l'inégalité (II.1) par récurrence sur $n = \dim(E)$.

Q17) Montrer que dans le cas $n = 1$, l'inégalité (II.1) est une égalité.

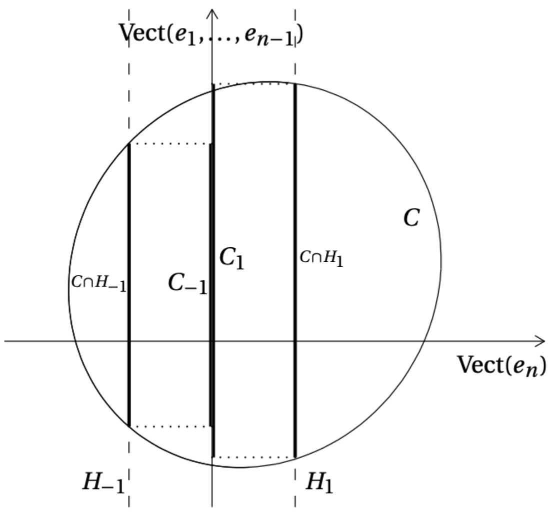
II.C - Propriétés de C_{+1} et C_{-1}

Soit n un entier tel que $n \geqslant 2$. On suppose à présent que (II.1) est vérifiée au rang $n - 1$. On note $E' = \operatorname{Vect}(e_1, \dots, e_{n-1})$ et π la projection orthogonale sur E', i.e. $\pi : \left| \begin{array}{ccc} & E \to E' & \\ \displaystyle\sum_{i=1}^{n} x_i e_i & \mapsto & \displaystyle\sum_{i=1}^{n-1} x_i e_i \end{array} \right.$

On pose $X' = \pi \circ X = \sum_{i=1}^{n-1} \varepsilon_i e_i$. C'est une variable aléatoire à valeurs dans E'. Pour t dans $\{-1, 1\}$ on note :

— H_t l'hyperplan affine $E' + te_n$

— $C_t = \pi(C \cap H_t)$.



Q18) Montrer, pour $x' \in E'$ et $t \in \{-1, 1\}$, l'équivalence : $x' \in C_t \Longleftrightarrow x' + te_n \in C$.

Q19) Montrer que C_{+1} et C_{-1} sont des convexes fermés non vides de E'.

Q20) Montrer que

$$\mathbb{P}(X \in C) = \frac{1}{2}\mathbb{P}\left(X' \in C_{+1}\right) + \frac{1}{2}\mathbb{P}\left(X' \in C_{-1}\right)$$

II.D - Suite de la démonstration de l'hérédité : admise ici, et application.

On admet qu'on a démontré la propriété donnée par (II.1), en fait il y a encore pas mal de travail ! !

Q21) A l'aide de (II.1) montrer l'inégalité de Talagrand suivante :

$$\boxed{\text{Pour tout } C \text{ convexe fermé non vide de } E \text{ et pour tout réel } t \text{ strictement positif} \quad \mathbb{P}(X \in C) \cdot \mathbb{P}(d(X, C) \geqslant t) \leqslant \exp\left(-\frac{t^2}{8}\right)}$$

Partie III : démonstration du théorème de Johnson-Lindenstrauss

Dans cette partie on considère l'espace $E = \mathcal{M}_{k,d}(\mathbb{R})$ muni du produit scalaire défini par

$$\forall (A, B) \in E^2, \quad \langle A \mid B \rangle = \operatorname{tr}\left(A^\top \cdot B\right)$$

On notera $\|\cdot\|_F$ la norme euclidienne associée. On rappelle que $\mathbb{R}^d$ et $\mathbb{R}^k$ sont munis de leurs normes euclidiennes canoniques, notées indistinctement $\|\cdot\|$. On identifie $\mathbb{R}^d$ à $\mathcal{M}_{d,1}(\mathbb{R})$, de sorte qu'un vecteur quelconque $x = (x_1, \dots, x_d)$ de $\mathbb{R}^d$ peut être identifié à la matrice colonne $(x_1 \dots x_d)^\top$. On fixe un vecteur $(u_1, \dots, u_d)$ dans $\mathbb{R}^d$, identifié comme ci-dessus à la matrice colonne $(u_1 \dots u_d)^\top$ de $\mathcal{M}_{d,1}(\mathbb{R})$), et tel que $\|u\| = 1$. On définit l'application

$$g : \left| \begin{array}{l} \mathcal{M}_{k,d}(\mathbb{R}) \to \mathbb{R} \\ M \mapsto \|M \cdot u\| \end{array} \right.$$

Soit $X = (\varepsilon_{ij})_{1 \leqslant i \leqslant k, 1 \leqslant j \leqslant d}$ une variable aléatoire à valeurs dans $\mathcal{M}_{k,d}(\mathbb{R})$, dont les coefficients ε_{ij} sont des variables aléatoires de Rademacher indépendantes dans leur ensemble.

III.A - Une inégalité de concentration

Q22) Montrer que $C = \{M \in \mathcal{M}_{k,d}(\mathbb{R}) \mid g(M) \leqslant r\}$ est une partie convexe et fermée de $\mathcal{M}_{k,d}(\mathbb{R})$.

Q23) Montrer que pour toute matrice M dans $\mathcal{M}_{k,d}(\mathbb{R})$

$$\|M \cdot u\| \leqslant \|M\|_F$$

Soient r et t deux réels, avec $t > 0$.

Q24) Montrer que pour toute matrice M dans $\mathcal{M}_{k,d}(\mathbb{R})$

$$d(M, C) < t \quad \Longrightarrow \quad g(M) < r + t$$

Q25) En déduire que

$$\mathbb{P}(g(X) \leqslant r) \cdot \mathbb{P}(g(X) \geqslant r + t) \leqslant \exp\left(-\frac{1}{8}t^2\right)$$

III.B - Médianes

On dit qu'un réel m est une médiane de la variable aléatoire $g(X)$ lorsque

$$\mathbb{P}(g(X) \geqslant m) \geqslant \frac{1}{2} \quad \text{et} \quad \mathbb{P}(g(X) \leqslant m) \geqslant \frac{1}{2}$$

Q26) Justifier que $g(X)$ admet au moins une médiane.
On pourra considérer la fonction G de $\mathbb{R}$ dans $\mathbb{R}$ telle que, pour tout réel $t, G(t) = \mathbb{P}(g(X) \leqslant t)$, et examiner l'ensemble $G^{-1}([1/2, 1])$. On se souviendra aussi que Ω étant fini, $g(X)$ prend un nombre fini *de valeurs.*

Q27) Déduire des deux questions précédentes que, pour tout réel strictement positif t :

$$\mathbb{P}(|g(X) - m| \geqslant t) \leqslant 4 \exp\left(-\frac{1}{8}t^2\right)$$

où m est une médiane de $g(X)$. Autrement dit que $|g(X) - m|$ est une variable aléatoire sous-gaussienne.

Q28) En déduire, à l'aide de la partie I, que $\mathbb{E}\left((g(X) - m)^2\right) \leqslant 32$.

Q29) En écrivant

$$g(X)^2 = \|Xu\|^2 = \sum_{i=1}^{k} \left(\sum_{\ell=1}^{d} \varepsilon_{i,\ell} u_\ell\right)^2$$

montrer que $\mathbb{E}\left(g(X)^2\right) = k$, et en déduire que $\mathbb{E}(g(X)) \leqslant \sqrt{k}$.

Q30) En déduire que $(\sqrt{k} - m)^2 \leqslant \mathbb{E}\left((g(X) - m)^2\right)$.

III. C - Un lemme-clé avec un petit calcul admis

Q31) A l'aide de ce qui précède et de la partie I, montrer que, pour tout réel $t > 0$:

$$\mathbb{P}(|g(X) - \sqrt{k}| \geqslant t) \leqslant 4 \exp(4) \exp\left(-\frac{1}{16}t^2\right)$$

On pose $A_k = \frac{X}{\sqrt{k}}$. Soient ε dans $]0, 1[$ et δ dans $]0, 1/2\,[$. On suppose que $k \geqslant 160 \frac{\ln(1/\delta)}{\varepsilon^2}$.

On admettra que, pour tout vecteur unitaire u dans $\mathbb{R}^d$, on peut en déduire (*par un calcul sans grâce*) :

$$\mathbb{P}\left(|\, \|A_k \cdot u\| - 1| > \varepsilon\right) < \delta$$

III.D - Conclusion

On conserve les notations et les hypothèses précédentes. Soient $v_1, \dots, v_N$ des vecteurs distincts dans $\mathbb{R}^d$. Pour tout $(i, j) \in [\![1, N]\!]^2$ tel que $i < j$ on note E_{ij} l'événement

$$(1 - \varepsilon) \|v_i - v_j\| \leqslant \|A_k \cdot v_i - A_k \cdot v_j\| \leqslant (1 + \varepsilon) \|v_i - v_j\|$$

Q32) Montrer que $\mathbb{P}\left(\overline{E_{ij}}\right) < \delta$, où $\overline{E_{ij}}$ désigne l'événement contraire de E_{ij}.

Q33) En déduire que $\mathbb{P}\left(\bigcap_{1 \leqslant i < j \leqslant N} E_{ij}\right) \geqslant 1 - \frac{N(N-1)}{2}\delta$.

Q34) En déduire le théorème de Johnson et Lindenstrauss.