CONCOURS CENTRALE•SUPÉLEC

Mathématiques 1

MP

4 heures Calculatrices autorisées

2018

Aplatissement aléatoire d'un ensemble de points en grande dimension

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 $v_1, ..., v_N$ 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 $v_1, ..., v_N$ 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 :

Objectif

Pour quelles valeurs de k existe-t-il $f : \mathbb{R}^d \to \mathbb{R}^k$ qui soit une ε-isométrie pour $v_1, ..., v_N$?

On 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 $v_1, \dots, v_N$ 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 $v_1, \dots, v_N$.

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 d'une inégalité de concentration, qui est utilisée dans la partie III où le théorème de Johnson-Lindenstrauss est démontré.

I Préliminaires

I.A – Projection sur un convexe fermé

Soit E un espace euclidien.

Q 1. Soient a et b dans E. Montrer la relation suivante et en donner une interprétation géométrique :

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

Q 2. En déduire que si u, v et v' dans E vérifient $v \neq v'$ et $\|u-v\| = \|u-v'\|$ alors $\|u - \dfrac{v+v'}{2}\| < \|u-v\|$.

Q 3. Soient F un fermé non vide de E et u dans E. Montrer qu'il existe v dans F tel que

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

Q 4. 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

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

On dira que v est le projeté de u sur C et on notera $d(u,C) = \|u-v\|$.

I.B – Inégalité de Hölder pour l'espérance

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

Q 5. Montrer que, pour tous réels positifs a et b,

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

On pourra utiliser la concavité du logarithme.

Q 6. En déduire que si X et Y sont deux variables aléatoires réelles sur l'espace probabilisé fini $(\Omega, \mathcal{A}, \mathbb{P})$ alors

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

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

I.C – Espérance conditionnelle

Soit $X : \Omega \to \mathbb{R}$ une variable aléatoire à valeurs réelles.

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.

Q 7. Soit $(A_1, ..., A_m)$ un système complet d'événements de probabilités non nulles. Montrer que

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

I.D – Variables aléatoires à queue sous-gaussienne

Soit $X : \Omega \to \mathbb{R}$ une variable aléatoire réelle.

On suppose qu'il existe deux réels strictement positifs a et b tels que, pour tout réel positif t,

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

Q 8. Montrer que

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

On pourra noter $X^2(\Omega) = \{y_1, ..., y_n\}$ avec $0 \leqslant y_1 < y_2 < \cdots < y_n$.

Q 9. Montrer que le moment d'ordre deux de X est inférieur ou égal à $\dfrac{a}{b}$.

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

Q 10. Justifier que, pour tout réel t,

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

Q 11. Montrer que, pour tout réel t,

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

Q 12. En déduire que pour tout réel t tel que $t \geqslant |\delta|$ on a

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

Q 13. Justifier que l'inégalité précédente reste valable si $0 \leqslant t < |\delta|$.

II L'inégalité de concentration de Talagrand

Soit E un espace euclidien de dimension $n \geqslant 1$ muni d'une base orthonormée $(e_1, ..., e_n)$.

Soient $\varepsilon_1, \ldots, \varepsilon_n : \Omega \to \{-1, 1\}$ 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

Q 14. Traiter le cas 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.

Q 15. Montrer que $\dfrac{1}{4}d(X,u)^2$ suit une loi binomiale de paramètres n et $1/2$.

Q 16. En déduire l'espérance de $\exp\left(\dfrac{1}{8}d(X,u)^2\right)$ et montrer qu'elle est inférieure ou égale à 2^n.

Q 17. Justifier que $d(X,C) \leqslant d(X,u)$ 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 se propose de démontrer l'inégalité (II.1) par récurrence sur la dimension n de E.

Q 18. Traiter le cas $n = 1$.

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 $E' = \mathrm{Vect}(e_1, ..., e_{n-1})$ et π la projection orthogonale sur E'

$$\pi : \left| \begin{array}{c} E \to E' \\ \displaystyle\sum_{i=1}^{n} x_i e_i \mapsto \sum_{i=1}^{n-1} x_i e_i \end{array} \right.$$

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 $E' + te_n$;

— $C_t = \pi(C \cap H_t)$.

Q 19. Montrer, pour $x' \in E'$ et $t \in \{-1, 1\}$, que $x' \in C_t \iff x' + te_n \in C$.

Q 20. Montrer que C_{+1} et C_{-1} sont des convexes fermés non vides de E'.

Pour t dans $\{-1,1\}$, on note Y_t le projeté de X' sur le convexe fermé non vide C_t. C'est une variable aléatoire à valeurs dans E'.

Q 21. Montrer que

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

II.D – Une inégalité cruciale

Soit λ un réel tel que $0 \leqslant \lambda \leqslant 1$.

Q 22. Montrer que

$$d(X,C) \leqslant \|(1-\lambda)(Y_{\varepsilon_n} + \varepsilon_n e_n) + \lambda(Y_{-\varepsilon_n} - \varepsilon_n e_n) - X\|$$

Q 23. En déduire que

$$d(X,C)^2 \leqslant 4\lambda^2 + \|(1-\lambda)(Y_{\varepsilon_n} - X') + \lambda(Y_{-\varepsilon_n} - X')\|^2$$

puis que

$$d(X,C)^2 \leqslant 4\lambda^2 + (1-\lambda)\|Y_{\varepsilon_n} - X'\|^2 + \lambda\|Y_{-\varepsilon_n} - X'\|^2$$

Ainsi, on a montré l'inégalité

$$d(X,C)^2 \leqslant 4\lambda^2 + (1-\lambda)d(X', C_{\varepsilon_n})^2 + \lambda d(X', C_{-\varepsilon_n})^2$$

II.E – Espérances conditionnelles

On note

$$p_+ = \mathbb{P}(X' \in C_{+1}) \qquad \text{et} \qquad p_- = \mathbb{P}(X' \in C_{-1})$$

On va supposer, sans perte de généralité, que $p_+ \geqslant p_-$.

Q 24. Montrer que $p_- > 0$.

Q 25. Montrer que pour tout λ dans $[0,1]$

$$\mathbb{E}\left(\exp\left(\frac{1}{8}d(X,C)^2\right) \mid \varepsilon_n = -1\right) \leqslant \exp\left(\frac{\lambda^2}{2}\right)\mathbb{E}\left(\left(\exp\left(\frac{1}{8}d(X', C_{-1})^2\right)\right)^{1-\lambda} \cdot \left(\exp\left(\frac{1}{8}d(X', C_{+1})^2\right)\right)^{\lambda}\right)$$

Q 26. En déduire que

$$\mathbb{E}\left(\exp\left(\frac{1}{8}d(X,C)^2\right) \mid \varepsilon_n = -1\right) \leqslant \exp\left(\frac{\lambda^2}{2}\right)\left(\mathbb{E}\left(\exp\left(\frac{1}{8}d(X', C_{-1})^2\right)\right)\right)^{1-\lambda} \cdot \left(\mathbb{E}\left(\exp\left(\frac{1}{8}d(X', C_{+1})^2\right)\right)\right)^{\lambda}$$

Q 27. À l'aide de l'hypothèse de récurrence, justifier que

$$\mathbb{E}\left(\exp\left(\frac{1}{8}d(X,C)^2\right) \mid \varepsilon_n = 1\right) \leqslant \frac{1}{p_+}$$

Q 28. Déduire de ce qui précède que pour tout λ dans $[0,1]$

$$\mathbb{E}\left(\exp\left(\frac{1}{8}d(X,C)^2\right)\right) \leqslant \frac{1}{2}\left(\frac{1}{p_+} + \exp\left(\frac{\lambda^2}{2}\right)\frac{1}{(p_-)^{1-\lambda}} \cdot \frac{1}{(p_+)^{\lambda}}\right)$$

II.F – Optimisation

Q 29. On pose $\lambda = 1 - \dfrac{p_-}{p_+}$. Montrer que

$$\mathbb{E}\left(\exp\left(\frac{1}{8}d(X,C)^2\right)\right) \leqslant \frac{1}{2p_+}\left(1 + \exp\left(\frac{\lambda^2}{2}\right)(1-\lambda)^{\lambda-1}\right)$$

Q 30. Montrer que pour tout $x \in [0,1[$

$$\frac{x^2}{2} + (x-1)\ln(1-x) \leqslant \ln(2+x) - \ln(2-x)$$

On pourra faire une étude de fonction.

Q 31. En déduire que pour tout $x \in [0,1[$

$$1+\exp\left(\frac{x^2}{2}\right)(1-x)^{x-1} \leqslant \frac{4}{2-x}$$

Q 32. Terminer la démonstration de l'inégalité (II.1).

II.G – Inégalité de Talagrand

Q 33. En déduire l'inégalité de Talagrand :

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

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

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

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 $x = (x_1,...,x_d)$ de $\mathbb{R}^d$ peut être identifié à la matrice colonne $(x_1 \ ... \ x_d)^\top$.

On fixe un vecteur $(u_1,...,u_d)$ dans $\mathbb{R}^d$, identifié comme ci-dessus à la matrice colonne $(u_1...u_d)^\top$ 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

Q 34. Montrer que $C = \{M \in \mathcal{M}_{k,d}(\mathbb{R}) \mid g(M) \leqslant r\}$ est une partie convexe et fermée de $\mathcal{M}_{k,d}(\mathbb{R})$.

Q 35. 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 $t>0$.

Q 36. Montrer que pour toute matrice M dans $\mathcal{M}_{k,d}(\mathbb{R})$

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

Q 37. En déduire que

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

III.B – Médianes

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

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

Q 38. 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 $G^{-1}([1/2,1])$.

Q 39. Déduire de ce qui précède que, pour tout réel strictement positif t

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

où m est une médiane de $g(X)$.

Q 40. En déduire que $\mathbb{E}\left(\left(g(X)-m\right)^2\right) \leqslant 32$.

Q 41. Montrer que $\mathbb{E}\left(g(X)^2\right) = k$, et en déduire que $\mathbb{E}(g(X)) \leqslant \sqrt{k}$.

Q 42. En déduire que $(\sqrt{k}-m)^2 \leqslant \mathbb{E}\left(\left(g(X)-m\right)^2\right)$.

III.C – Un lemme-clé

Q 43. Montrer que, pour tout réel strictement positif t

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

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

Q 44. Montrer que, pour tout vecteur unitaire u dans $\mathbb{R}^d$:

$$\mathbb{P}\Big(\big|\|A_k\cdot u\|-1\big| > \varepsilon\Big) < \delta$$

III.D – Conclusion

On conserve les notations et les hypothèses précédentes. Soient $v_1,...,v_N$ 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

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

Q 45. Montrer que $\mathbb{P}(\overline{E_{ij}}) < \delta$, où $\overline{E_{ij}}$ désigne l'événement contraire de E_{ij}.

Q 46. En déduire que $\mathbb{P}\left(\bigcap_{1\leqslant i<j\leqslant N} E_{ij}\right) \geqslant 1-\dfrac{N(N-1)}{2}\delta$.

Q 47. En déduire le théorème de Johnson et Lindenstrauss.

• • • FIN • • •
