Exercices sur le problème d'acquisition comprimée

Guillaume Lecué

8 mars 2016

Table des matières

1 Complexité

****

Exercice 1.1 (Argument volumique)
Soit $\|\cdot\|$ une norme sur $\mathbb{R}^n$. On note par B sa boule unité. Soit $0 < \varepsilon \leq 1$ et $\Lambda \subset B$ tels que pour tout $x, y \in \Lambda$

$$\|x - y\| \geq \varepsilon. \tag{1}$$

Alors nécessairement, le cardinal de Λ est tel que

$$|\Lambda| \leq \Big(1 + \frac{2}{\varepsilon}\Big)^n$$

****

Correction de l'exercice 1.1 On a

$$\bigcup_{x\in\Lambda} \big(x + (\varepsilon/2)\mathring{B}\big) \subset B + (\varepsilon/2)B = (1 + \varepsilon/2)B. \tag{2}$$

Par ailleurs, les ensembles $(x + (\varepsilon/2)\mathring{B})_{x\in\Lambda}$ sont disjoints. Alors, en prenant le volume dans (2), on obtient

$$\sum_{x\in\Lambda} \mathrm{Vol}\big(x + (\varepsilon/2)\mathring{B}\big) \leq \mathrm{Vol}\big((1 + \varepsilon/2)B\big).$$

Le volume étant invariant par translation, on a pour tout $x \in \Lambda$, $\mathrm{Vol}\big(x + (\varepsilon/2)\mathring{B}\big) = \mathrm{Vol}((\varepsilon/2)\mathring{B})$. De plus, par homothétie on a $\mathrm{Vol}((\varepsilon/2)\mathring{B}) = (\varepsilon/2)^n \mathrm{Vol}(\mathring{B})$ et $\mathrm{Vol}\big((1 + \varepsilon/2)B\big) = (1 + \varepsilon/2)^n \mathrm{Vol}(B)$. Par ailleurs, le volume du bord étant de mesure nulle, on a $\mathrm{Vol}(B) = \mathrm{Vol}(\mathring{B})$. On obtient alors

$$|\Lambda|(\varepsilon/2)^n \mathrm{Vol}(B) \leq (1 + \varepsilon/2)^n \mathrm{Vol}(B).$$

Donc $|\Lambda| \leq (1+2/\varepsilon)^n$.

****

Exercice 1.2 (Varshamov-Gilbert)

Soit $s \leq N/2$. Il existe une famille $\mathcal{S}$ d'ensembles de $\{1,\dots,N\}$ telle que

1. $\forall S \in \mathcal{S}$, $|S| = s$,
2. $\forall S_1, S_2 \in \mathcal{S}$, $S_1 \neq S_2 \Rightarrow |S_1 \cap S_2| \leq \lfloor s/2 \rfloor$,
3. $\log|\mathcal{S}| \geq \lfloor \frac{s}{2} \rfloor \log\left(\lfloor \frac{N}{8es} \rfloor\right)$.

****

Correction de l'exercice 1.2 On va faire une énumération successive des sous-ensembles de cardinal s de $\{1,\dots,N\}$ et retirer les éléments qui ne nous intéresse pas. Sans perte de généralité, on suppose que $s/2$ est un entier.

Soit S_1 un sous-ensemble de $\{1,...,N\}$ de cardinal s. On va "jeter" tous les sous-ensembles J de $\{1,...,N\}$ de cardinal s tels que $|S_1 \cap J| \leq s/2$. Il y en a

$$\sum_{k=s/2}^{s} \binom{s}{k}\binom{N-s}{s-k}$$

et puisque $s \leq N/2$, on a

$$\sum_{k=s/2}^{s} \binom{s}{k}\binom{N-s}{s-k} \leq 2^s \max_{s/2\leq k\leq s} \binom{N-s}{s-k} \leq 2^s \binom{N}{s/2}.$$

On prend S_2 de cardinal s parmi les ensembles restants et comme précédemment on "jete" tous les sous-ensembles qui ont plus de $s/2$ éléments en commun avec S_2. On réitère l'argument pour obtenir une famille $\mathcal{S} = \{S_1, S_2, \dots, S_p\}$ de sous-ensembles de cardinal s qui sont mutuellement s-séparés pour la distance de Hamming et tels que

$$|\mathcal{S}| \geq \left\lfloor \binom{N}{s} \Big/ 2^s \binom{N}{s/2} \right\rfloor.$$

Comme $s \leq N/2$, on a $\left(\frac{N}{s}\right)^s \leq \binom{N}{s} \leq \left(\frac{eN}{s}\right)^s$ et on obtient

$$|\mathcal{S}| \geq \left\lfloor \frac{(N/s)^s}{2^s (Ne/(s/2))^{(s/2)}} \right\rfloor \geq \left\lfloor \left(\frac{N}{8es}\right)^{s/2} \right\rfloor \geq \left\lfloor \frac{N}{8es} \right\rfloor^{\lfloor s/2 \rfloor}$$

ce qui conclut la preuve.

****

Exercice 1.3 (Calcul de l'intégrale de Dudley de B_1^N par rapport à ℓ_2^N)

On considère les boules unités $B_1^N = \{x \in \mathbb{R}^N : \sum_{i=1}^N |x_i| \leq 1\}$ et $B_2^N = \{x \in \mathbb{R}^N : \sum_{i=1}^N x_i^2 \leq 1\}$. Pour tout $\varepsilon > 0$, le nombre minimal de translatés de la boule εB_2^N nécessaires pour

couvrir B_1^N est noté $N(B_1^N, \varepsilon B_2^N)$. On rappelle qu'une mesure de la complexité de B_1^N pour la métrique ℓ_2^N est l'intégrale de Dudley définie par :

$$I(B_1^N, \ell_2^N) := \int_0^\infty \sqrt{\log N(B_1^N, \varepsilon B_2^N)} d\varepsilon.$$

On va montrer qu'il existe une constante absolue $c_0 > 0$ telle que $I(B_1^N, \ell_2^N) \leq c_0 (\log N)^{3/2}$. Pour cela on utilise un argument probabiliste.

On calcul d'abord le nombre d'entropie de B_1^N par rapport à ℓ_2^N. Soit $\varepsilon > 0$. On cherche à construire un ε-réseau de B_1^N pour ℓ_2^N. C'est-à-dire, on cherche un nombre minimal de points $x_1, \dots, x_p \in B_1^N$ tels que pour tout $x \in B_1^N$ il existe $i_0 \in \{1, \dots, p\}$ tel que $\|x - x_{i_0}\|_2 \leq \varepsilon$. Soit $x \in B_1^N$. On écrit $x = \sum_{i=1}^N \lambda_i e_i$ où $(e_1, \dots, e_N)$ est la base canonique de $\mathbb{R}^N$ et où $\sum_{i=1}^N |\lambda_i| \leq 1$. On considère la variable aléatoire X à valeurs dans $\{\pm e_1, \dots, \pm e_N, 0\}$ définie par

$$\mathbb{P}[X = \mathrm{Sign}(\lambda_i) e_i] = |\lambda_i|, \, \forall i = 1, \dots, N \text{ et } \mathbb{P}[X = 0] = 1 - \|x\|_1 .$$

1. **Déterminer la moyenne de X.**
2. **Soit n un entier à déterminer plus tard et $X_1, \dots, X_n$ des variables aléatoires indépendantes et de même loi que X. Montrer que $\mathbb{E} \left\| n^{-1} \sum_{i=1}^n X_i - \mathbb{E}X \right\|_2^2 \leq n^{-1}$**
3. **En déduire que si $n = \lfloor \varepsilon^{-2} \rfloor$ alors**

$$\Lambda_\varepsilon := \left\{ \frac{1}{n} \sum_{i=1}^n z_i : z_1, \dots, z_n \in \{\pm e_1, \dots, \pm e_N, 0\} \right\}$$

est un ε-réseau de B_1^N par rapport à ℓ_2^N. Déterminer alors une borne pour $\log N(B_1^N, \varepsilon B_2^N)$.

4. **Montrer que pour tout $\eta \geq \varepsilon$, on a**

$$\log N(B_1^N, \varepsilon B_2^N) \leq \log N(B_1^N, \eta B_2^N) + N \log \left(\frac{3\eta}{\varepsilon} \right).$$

5. **En déduire que pour tout $0 < \varepsilon \leq N^{-1/2}$,**

$$\log N(B_1^N, \varepsilon B_2^N) \leq c_0 N \log \left(c_1 / N\varepsilon^2 \right)$$

6. **Finalement, prouver que**

$$\log N(B_1^N, \varepsilon B_2^N) \leq c_0 \begin{cases} 0 & \text{quand } \varepsilon \geq 1, \\ \frac{\log(eN)}{\varepsilon^2} & \text{quand } N^{-1/2} \leq \varepsilon \leq 1, \\ N \log \left(\frac{e}{N\varepsilon^2} \right) & \text{quand } 0 < \varepsilon \leq N^{-1/2}. \end{cases}$$

Conclure qu'il existe bien $c_0 > 0$ tel que $I(B_1^N, \ell_2^N) \leq c_0 (\log N)^{3/2}$.

****

Correction de l'exercice 1.3

1.

$$\mathbb{E}X = \sum_{y \in \{\pm e_1, \dots, \pm e_N, 0\}} y \mathbb{P}[X = y] = \sum_{i=1}^N \mathrm{Sign}(\lambda_i) e_i |\lambda_i| = x.$$

2. On a :

$$\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^{n}X_i-\mathbb{E}X\right\|_2^2=\frac{\mathbb{E}\,\|X-\mathbb{E}X\|_2^2}{n}=\frac{\mathbb{E}\,\|X\|_2^2-\|\mathbb{E}X\|_2^2}{n}=\frac{\|x\|_1^2-\|x\|_2^2}{n}\leq\frac{1}{n}.$$

3. On a vu que $\mathbb{E}X=x$ et $\mathbb{E}\left\|n^{-1}\sum_{i=1}^{n}X_i-\mathbb{E}X\right\|_2^2\leq\varepsilon^2$ donc il existe $\omega\in\Omega$ tel que

$$\left\|n^{-1}\sum_{i=1}^{n}X_i(\omega)-x\right\|_2\leq\varepsilon.$$

Or $n^{-1}\sum_{i=1}^{n}X_i(\omega)\in\Lambda_\varepsilon$. Donc il existe bien un $y\in\Lambda_\varepsilon$ tel que $\|y-x\|_2\leq\varepsilon$. Ceci étant vrai pour tout $x\in B_1^N$, Λ_ε est bien un ε-réseau de B_1^N pour ℓ_2^N.

On en déduit que $N(B_1^N,\varepsilon B_2^N)$ est plus petit que le cardinal de Λ_ε. Comme $|\Lambda_\varepsilon|\leq(2N+1)^n$, on a $\log N(B_1^N,\varepsilon B_2^N)\leq n\log(2N+1)\leq 2\varepsilon^{-2}\log(2N+1)$.

Note : En fait, on peut montrer que $|\Lambda_\varepsilon|\leq(c_0N/n+1)^n$ et donc $\log N(B_1^N,\varepsilon B_2^N)\lesssim\varepsilon^{-2}\log(c_0N\varepsilon^2)$ quand $N^{-1/2}\leq\varepsilon\leq 1$.

4. Pour recouvrir B_1^N par des boules εB_2^N, on peut d'abord recouvrir B_1^N par des boules ηB_2^N puis recouvrir chacune des boules ηB_2^N par des boules εB_2^N. On a alors

$$N(B_1^N,\varepsilon B_2^N)\leq N(B_1^N,\eta B_2^N)N(\eta B_2^N,\varepsilon B_2^N).$$

On sait aussi qu'un argument volumique donne

$$N(\eta B_2^N,\varepsilon B_2^N)=N\Big(B_2^N,\frac{\varepsilon}{\eta}B_2^N\Big)\leq\Big(1+\frac{2\eta}{\varepsilon}\Big)^N.$$

5. On a d'aprés 3) et 4), pour tout $\eta\geq\varepsilon$,

$$\log N(B_1^N,\varepsilon B_2^N)\leq 2\eta^{-2}\log(2N+1)+N\log\Big(\frac{3\eta}{\varepsilon}\Big).$$

Alors quand $\varepsilon\leq N^{-1/2}$ et $\eta=\sqrt{(\log N)/N}$, on obtient

$$\log N(B_1^N,\varepsilon B_2^N)\leq c_0N\log\Big(\frac{c_1}{N\varepsilon^2}\Big).$$

6. On a

$$\begin{aligned}I(B_1^N,\ell_2^N)&\leq\int_0^{N^{-1/2}}\sqrt{N\log\Big(\frac{e}{N\varepsilon^2}\Big)}d\varepsilon+\int_{N^{-1/2}}^1\sqrt{\frac{\log(eN)}{\varepsilon^2}}d\varepsilon\\&\leq\int_0^1\sqrt{\log\Big(\frac{e}{u}\Big)}du+\sqrt{\log(eN)}\int_{N^{-1/2}}^1\frac{d\varepsilon}{\varepsilon}\leq c_0(\log(eN))^{3/2}.\end{aligned}$$

****

Exercice 1.4 (Borne inférieure sur l'intégrale de Dudley de B_1^N par rapport à ℓ_2^N)
En utilisant la borne de Varshamov-Gilbert, montrer que

$$I(B_1^N,\ell_2^N)=\int_0^\infty\sqrt{\log N(B_1^N,\varepsilon B_2^N)}d\varepsilon\geq c_1(\log N)^{3/2}$$

pour une certaine constante absolue $c_1 > 0$.

On pourra d'abord montrer que si $N(B_1^N, \varepsilon B_2^N)$ est le nombre minimal de translatées de εB_2^N nécessaires pour couvrir B_1^N et $M(B_1^N, \varepsilon B_2^N)$ est le nombre maximal de points ε-écartés dans B_1^N par rapport à ℓ_2^N alors

$$N(B_1^N, \varepsilon B_2^N) \geq M(B_1^N, \varepsilon B_2^N).$$

****

Correction de l'exercice 1.4 Soit Λ_ε un ensemble ε-écarté de B_1^N par rapport à ℓ_2^N de cardinale maximal. On a par définition $|\Lambda_\varepsilon| = M(B_1^N, \varepsilon B_2^N)$ et si on ajoute un point $x \in B_1^N$ à Λ_ε alors soit $x \in \Lambda_\varepsilon$ soit $\Lambda_\varepsilon \cup \{x\}$ n'est plus ε-écarté et donc dans les deux cas il va exister $y \in \Lambda_\varepsilon$ tel que $\|x - y\|_2 \leq \varepsilon$. Donc Λ_ε est aussi un ε-réseau de B_1^N par rapport à ℓ_2^N. On en déduit que $N(B_1^N, \varepsilon B_2^N) \geq |\Lambda_\varepsilon|$ et donc $N(B_1^N, \varepsilon B_2^N) \geq M(B_1^N, \varepsilon B_2^N)$.

Soit $1 \leq s \leq N/2$. D'après la borne de Varshamov-Gilbert, il existe une collection $\mathcal{S}$ de sous-ensembles de $\{1, \ldots, N\}$ telle que :

1. pour tout $S \in \mathcal{S}$, $|S| = s$
2. pour tout $S_1 \neq S_2 \in \mathcal{S}$, $|S_1 \cap S_2| \leq s/2$
3. $\log |\mathcal{S}| \geq \lfloor \frac{s}{2} \rfloor \log\left(\lfloor \frac{N}{8es} \rfloor\right)$.

On considère une telle collection et on pose

$$\Lambda_s = \left\{x(S) : S \in \mathcal{S}\right\} \text{ où } x(S) = \frac{1}{s}\sum_{j \in S} e_j$$

où $(e_1, \ldots, e_N)$ est la base canonique de $\mathbb{R}^N$.

Par construction pour tout $S_1 \neq S_2 \in \mathcal{S}$, on a

$$\|x(S_1) - x(S_2)\|_2 = \sqrt{\sum_{j \in S_1 \Delta S_2} (1/s)^2} \geq \frac{1}{\sqrt{s}}.$$

De plus, pour tout $S \in \mathcal{S}, \|x(S)\|_1 = 1$ donc Λ_s est un ensemble $(1/\sqrt{s})$-écarté de B_1^N pour ℓ_2^N. Pour $\varepsilon = 1/\sqrt{s}$, on en déduit que

$$M(B_1^N, \varepsilon B_2^N) \geq |\Lambda_{1/\varepsilon^2}| \geq \frac{c_0 \log(c_1 N \varepsilon^2)}{\varepsilon^2}.$$

On en déduit que l'intégrale de Dudley vérifie

$$I(B_1^N, \ell_2^N) \geq \int_{1/\sqrt{N}}^1 \sqrt{\log N(B_1^N, \varepsilon B_2^N)} d\varepsilon \gtrsim \int_{1/\sqrt{N}}^1 \frac{\sqrt{\log(c_1 N \varepsilon^2)}}{\varepsilon} d\varepsilon \sim [\log(c_1 N)]^{3/2}.$$

****

Exercice 1.5 (Sur la complexité du cube combinatoire par rapport à la distance de Hamming)

On considère un entier N et le cube combinatoire $\mathcal{C}_N = \{0,1\}^N$. On munit $\mathcal{C}_N$ de la distance de Hamming : pour tout $x, y \in \mathcal{C}_N$,

$$\rho(x, y) = \sum_{i=1}^N I(x_i \neq y_i)$$

où $I(x_i \neq y_i) = 1$ quand $x_i \neq y_i$ et $I(x_i \neq y_i) = 0$ quand $x_i = y_i$. On fixe un entier $1 \leq k < N/2$. On s'intéresse aux sous-ensembles de points de $\mathcal{C}_N$ qui sont k-écartés par rapport à la distance de Hamming.

1. **On considère X la variable aléatoire uniformément distribuée sur $\mathcal{C}_N$. Soit $x \in \mathcal{C}_N$. Montrer que**
$$2^N \mathbb{P}\big[\rho(X,x) \leq k\big] = |\{y \in \mathcal{C}_N : \rho(y,x) \leq k\}|$$
où pour tout ensemble E, le cardinal de E est noté $|E|$.
2. **Étant donné $x \in \mathcal{C}_N$, déterminer la loi de $\rho(X,x)$.**
3. **Étant donné un point $x \in \mathcal{C}_N$, utiliser 1), 2) et la borne de Chernoff pour majorer le nombre de points qui sont à distance au plus k de x pour la distance de Hamming.**
4. **Construire un ensemble de points k écartés de cardinal au moins**
$$\left\lceil \frac{2^N}{\big(eN/k\big)^k} \right\rceil.$$

****

Correction de l'exercice 1.5

1. Comme X est uniformément distribué sur $\mathcal{C}_N$, pour tout $y \in \mathcal{C}_N$, on a $\mathbb{P}[X = y] = 2^{-N}$ et donc
$$\mathbb{P}\big[\rho(X,x) \leq k\big] = \sum_{y \in \mathcal{C}_N} I(\rho(y,x) \leq k)\mathbb{P}[X = y] = \frac{1}{2^N}|\{y \in \mathcal{C}_N : \rho(y,x) \leq k\}|.$$
2. Soit $x \in \mathcal{C}_N$. On a $\rho(X,x) = \sum_{i=1}^N I(X_i \neq x_i)$. Comme X est uniformément distribuée sur $\mathcal{C}_N$, pour tout $i \in \{1,\dots,N\}$, X_i est uniformément distribuée sur $\{0,1\}$, c'est donc une loi de Bernoulli de paramètre $1/2$. De même quelque soit $x_i \in \{0,1\}$, $I(X_i \neq x_i)$ est aussi une Bernoulli de paramètre $1/2$. On voit aussi que les X_i sont indépendantes et donc les $\delta_i = I(X_i \neq x_i)$ sont aussi indépendantes. Donc $\rho(X,x)$ est une somme de N variables de Bernoulli indépendantes de paramètre $1/2$, c'est donc un loi multinomiale de paramètre $1/2$.
3. Soit $\delta_1,\dots,\delta_N$ des variables de Bernoulli indépendantes de paramétre $1/2$. D'après 1), 2) et la borne de Chernoff, on a
$$|\{y \in \mathcal{C}_N : \rho(y,x) \leq k\}| = 2^N \mathbb{P}\big[\rho(X,x) \leq k\big] = 2^N \mathbb{P}\Big[\sum_{i=1}^N \delta_i \leq k\Big]$$
$$= 2^N \mathbb{P}\Big[\frac{1}{N}\sum_{i=1}^N \frac{1}{2} - \delta_i \geq \frac{1}{2} - \frac{k}{N}\Big] \leq 2^N \exp\big(-N h_{1/2}(1/2 - k/N)\big)$$
où
$$h_{1/2}(1/2 - k/N) = \log 2 - \frac{k}{N}\log\Big(\frac{N}{k}\Big) - \Big(1 - \frac{k}{N}\Big)\log\Big(\frac{N}{N-k}\Big).$$
On a donc
$$|\{y \in \mathcal{C}_N : \rho(y,x) \leq k\}| \leq \Big(\frac{N-k}{k}\Big)^k \Big(\frac{N}{N-k}\Big)^N \leq \Big(\frac{eN}{k}\Big)^k.$$

4. On considère $x_1 = (0, \dots, 0) \in \mathcal{C}_N$. On note

$$\Omega_1 = \{y \in \mathcal{C}_N : \rho(y, x_1) \leq k\}.$$

On a $|\Omega_1| \leq \big(eN/k\big)^k$. On choisit x_2 dans $\mathcal{C}_N - \Omega_1$. Il y a au moins $2^N - \big(eN/k\big)^k \geq 1$ choix possibles pour x_2. On note

$$\Omega_2 = \{y \in \mathcal{C}_N : \rho(y, x_2) \leq k\}.$$

On choisit un point x_3 dans $\mathcal{C}_N - \Omega_1 \cup \Omega_2$. Il y a au moins $2^N - 2\big(eN/k\big)^k \geq 1$ choix possibles pour x_3 de plus x_3 est k-écarté de x_1 et x_2. On réitère le procédé jusqu'à l'étape p telle que

$$2^N - (p+1)\big(eN/k\big)^k < 1 \leq 2^N - p\big(eN/k\big)^k.$$

On en déduit le résultat.

2 Concentration

****

Exercice 2.1 (Estimé de Bernoulli)

Soit $(A_i)_{i\geq 1}$ une suite d'événements indépendants tel que $a := \sum_{i\geq 1} \mathbb{P}(A_i) < \infty$. Pour tout n, l'*estimé de Bernoulli* dit que

$$\mathbb{P}\Big[\sum_{i\geq 1} I_{A_i} \geq n\Big] \leq \frac{a^n}{n!} \leq \Big(\frac{ea}{n}\Big)^n.$$

****

Correction de l'exercice 2.1 Par la borne de l'union et l'indépendance

$$\mathbb{P}\big[\sum_{i\geq 1} I_{A_i} \geq n\big] = \mathbb{P}\Big[\cup_{i_1<\dots<i_n} A_{i_1} \cap \dots \cap A_{i_n}\Big] \leq \sum_{i_1<\dots<i_n} \prod_{j=1}^n \mathbb{P}(A_j)$$
$$= \frac{1}{n!} \sum_{i_1\neq\dots\neq i_n} \prod_{j=1}^n \mathbb{P}(A_{i_j}) \leq \frac{1}{n!} \sum_{(i_1,\dots,i_n)} \prod_{j=1}^n \mathbb{P}(A_{i_j}) = \frac{1}{n!}\Big(\sum_i \mathbb{P}(A_i)\Big)^n = \frac{a^n}{n!}.$$

On conclut vu que $n! \geq (n/e)^n$.

****

Exercice 2.2 (Inégalité de concentration de Chernoff)

Soit $\delta_1, \dots, \delta_n$ des variables i.i.d. de Bernoulli de moyenne δ (aussi appelé des *sélecteurs*) définies par $\mathbb{P}[\delta_1 = 1] = 1 - \mathbb{P}[\delta_1 = 0] = \delta$. L'inégalité de concentration de Chernoff dit que pour tout $0 \leq t < 1 - \delta$,

$$\mathbb{P}\Big[\frac{1}{n}\sum_{i=1}^n \delta_i - \delta \geq t\Big], \mathbb{P}\Big[\frac{1}{n}\sum_{i=1}^n \delta_i - \delta \leq -t\Big] \leq \exp(-n h_\delta(t)),$$

où $h_\delta(t) := (1-\delta-t)\log\Big(\frac{1-\delta-t}{1-\delta}\Big) + (\delta+t)\log\Big(\frac{\delta+t}{\delta}\Big)$.

****

Correction de l'exercice 2.2 On utilise la méthode de Cramer-Chernoff : pour tout $t > 0$,

$$\mathbb{P}\Big[\frac{1}{n}\sum_{i=1}^{n}\delta_i - \delta \geq t\Big] \leq \exp\big(-\psi^*_{\bar{\delta}_n}(t)\big),$$

où $\psi^*_{\bar{\delta}_n}$ est le conjugué convexe de la transformée de Laplace de $\bar{\delta}_n := n^{-1}\sum_{i=1}^n \delta_i - \delta$ définie par $\psi_{\bar{\delta}_n}(\lambda) := \log \mathbb{E}\exp(\lambda\bar{\delta}_n)$ pour tout $\lambda > 0$ et

$$\psi^*_{\bar{\delta}_n}(t) = \sup_{\lambda>0}\big[\lambda t - \psi_{\bar{\delta}_n}(\lambda)\big].$$

En utilisant l'indépendance, on a $\psi_{\bar{\delta}_n}(\lambda) = n\psi_{\delta_1-\delta}(\lambda/n)$ et donc $\psi^*_{\bar{\delta}_n}(t) = n\psi^*_{\delta_1-\delta}(t)$. On montre directement que pour tout $\lambda \geq 0$,

$$\psi_{\delta_1-\delta}(\lambda) = \log\big(\delta\exp(\lambda) + 1 - \delta\big) - \lambda\delta$$

et pour tout $0 \leq t < 1 - \delta$,

$$\psi^*_{\delta_1-\delta}(t) = (1-\delta-t)\log\Big(\frac{1-\delta-t}{1-\delta}\Big) + (\delta+t)\log\Big(\frac{\delta+t}{\delta}\Big).$$

Ce qui démontre la première inégalité. La deuxième inégalité s'obtient de la même manière en changeant δ_i par $-\delta_i$.

****

Exercice 2.3 (Les variables bornées sont sous-gaussiennes)
Soit X une variable aléatoire réelle centrée telle que $a \leq X \leq b$ p.s.. Alors pour tout $\lambda > 0$,

$$\mathbb{E}[e^{\lambda X}] \leq \exp\Big(\frac{\lambda^2(b-a)^2}{8}\Big).$$

****

Correction de l'exercice 2.3 Soit $\lambda > 0$. Par convexité de la fonction exponentielle, on a pour tout $x \in [a, b]$,

$$\exp(\lambda x) = \exp\Big[\Big(\frac{x-a}{b-a}\Big)\lambda b + \Big(\frac{b-x}{b-a}\Big)\lambda a\Big] \leq \Big(\frac{x-a}{b-a}\Big)\exp(\lambda b) + \Big(\frac{b-x}{b-a}\Big)\exp(\lambda a).$$

On a alors

$$\mathbb{E}\exp(\lambda X) \leq \Big(\frac{-a}{b-a}\Big)\exp(\lambda b) + \Big(\frac{b}{b-a}\Big)\exp(\lambda a) = \Big(1-p+p\exp(s(b-a))\Big)\exp(-ps(b-a)) = \exp(\phi(u))$$

pour $p = -a/(b-a)$, $u = s(b-a)$ et $\phi(u) = -pu + \log(1-p+pe^u)$.

Pour tout $\alpha, \beta > 0$, on a $\alpha\beta \leq (\alpha+\beta)^2/4$. Alors, pour tout $t > 0$

$$\phi'(t) = -p + \frac{p}{p+(1-p)e^{-t}} \text{ et } \phi''(t) = \frac{p(1-p)e^{-t}}{(p+(1-p)e^{-t})^2} \leq \frac{1}{4}.$$

Un développement de Taylor à l'ordre 2 donne : $\phi(u) \leq \phi(0) + u\phi'(0) + u^2/8 = s^2(b-a)^2/8$.

****

Exercice 2.4 (Inégalité de concentration de Hoeffding)

Soit $X_1, \dots, X_n$ des variables aléatoires réelles centrées indépendantes telles que pour tout $i = 1, \dots, n$, $a_i \leq X_i \leq b_i$. Pour tout $t > 0$,

$$\mathbb{P}\Big[\frac{1}{n}\sum_{i=1}^{n} X_i > t\Big] \leq \exp\Big(-\frac{2nt^2}{\frac{1}{n}\sum_{i=1}^{n}(b_i - a_i)^2}\Big).$$

****

Correction de l'exercice 2.4 On utilise la méthode de Cramer-Chernoff :

$$\mathbb{P}\Big[\frac{1}{n}\sum_{i=1}^{n} X_i > t\Big] \leq \inf_{\lambda>0} e^{-\lambda t}\mathbb{E}\exp\Big(\frac{\lambda}{n}\sum_{i=1}^{n} X_i\Big) = \inf_{\lambda>0} e^{-\lambda t}\prod_{i=1}^{n}\mathbb{E}\exp(\lambda X_i/n).$$

Comme les variables aléatoires X_i satisfont $a_i \leq X_i \leq b_i$, on a d'après l'exercice précédent :

$$\mathbb{E}\exp(\lambda X_i/n) \leq \exp\Big(\frac{\lambda^2(b_i - a_i)^2}{8n^2}\Big).$$

On obtient

$$\mathbb{P}\Big[\frac{1}{n}\sum_{i=1}^{n} X_i > t\Big] \leq \inf_{\lambda>0}\exp\Big(-\lambda t + \frac{\lambda^2}{8n^2}\sum_{i=1}^{n}(b_i - a_i)^2\Big) = \exp\Big(-\frac{2nt^2}{\frac{1}{n}\sum_{i=1}^{n}(b_i - a_i)^2}\Big).$$

****

Exercice 2.5 (Dénombrement et concentration)

Démontrer que pour tout N et tout $1 \leq D \leq N$, on a

$$\sum_{i=0}^{D}\binom{N}{i} \leq \Big(\frac{eN}{D}\Big)^D.$$

On pourra considérer N variables de Bernoulli $\delta_1, \dots, \delta_N$ de moyenne $1/2$ et utiliser une inégalité de concentration pour borner $\mathbb{P}[\sum_{i=1}^{N}\delta_i \leq D]$.

****

Correction de l'exercice 2.5 La fonction $D \to \big(eN/D\big)^D$ est croissante. Ainsi quand $D \geq N/2$, on a $\big(eN/D\big)^D \geq (\sqrt{2e})^N > 2^N$ donc l'inégalité est vérifiée dans ce cas. On peut maintenant supposer $D < N/2$.

Soit $\delta_1, \dots, \delta_N$ des variables de Bernoulli indépendantes de moyenne $1/2$. On note $S = \sum_{i=1}^{N}\delta_i$. On voit que

$$\frac{1}{2^N}\sum_{i=0}^{D}\binom{N}{i} = \mathbb{P}\big[S \leq D\big].$$

On a $\mathbb{E}S = N/2$ alors, comme $N/2 - D > 0$,

$$\mathbb{P}\big[S \leq D\big] = \mathbb{P}\Big[\sum_{i=1}^{N}\mathbb{E}\delta_i - \delta_i \geq \frac{N}{2} - D\Big] = \mathbb{P}\Big[\frac{1}{N}\sum_{i=1}^{N} X_i \geq t\Big]$$

où $X_i = \mathbb{E}\delta_i - \delta_i$ et $t = 1/2 - D/N > 0$. On s'est donc ramené à un problème de concentration de la moyenne empirique. On peut appliquer l'inégalité de Hoeffding : $|X_i| \leq 1/2$ p.s. donc

$$\mathbb{P}\Big[\frac{1}{N}\sum_{i=1}^{N} X_i \geq t\Big] \leq \exp\Big(\frac{-2Nt^2}{(1/2)^2}\Big) = \exp\big(-8Nt^2\big).$$

L'inégalité de concentration de Bernstein donne : $|X_i| \leq 1/2$ p.s. et $\mathrm{var}(X_i) = 1/4$ donc

$$\mathbb{P}\Big[\frac{1}{N}\sum_{i=1}^{N} X_i \geq t\Big] \leq \exp\big(-c_0 N \min(t^2, t)\big).$$

On voit notamment que dans ce cas l'inégalité de Hoeffding donne un meilleure résultat. En effet, le contrôle sur la variance des X_i n'apporte pas d'information suplémentaire au fait que les X_i sont bornées.

On peut aussi appliquer l'inégalité de concentration de Chernoff qui donne pour $\delta_1, \ldots, \delta_N$ des Bernoulli indépendantes de paramètre δ,

$$\mathbb{P}\Big[\frac{1}{N}\sum_{i=1}^{N} \delta - \delta_i \geq t\Big] \leq \exp\Big(-N h_\delta(t)\Big)$$

où

$$h_\delta(t) = (1-\delta-t)\log\Big(\frac{1-\delta-t}{1-\delta}\Big) + (\delta+t)\log\Big(\frac{\delta+t}{\delta}\Big).$$

On a donc grâce à l'estimé de Bernoulli,

$$\begin{aligned}\sum_{i=0}^{D}\binom{N}{i} &= 2^N \mathbb{P}\big[S \leq D\big] \leq 2^N \exp\big(-N h_{1/2}(1/2 - D/N)\big) \\ &= \exp\Big[D\log\Big(\frac{N}{D}\Big) + (N-D)\log\Big(\frac{N}{N-D}\Big)\Big] \leq \Big(\frac{eN}{D}\Big)^D.\end{aligned}$$

****

Exercice 2.6 (Lemme de Johnson-Lindenstrauss)

Ce lemme est un fameux résultat de géometrie en grandes dimensions. Etant donné p points $x_1, \ldots, x_p$ dans l'espace $\mathbb{R}^N$ (qu'on imagine de grande dimension), il existe une matrice $A : \mathbb{R}^N \to \mathbb{R}^k$ telle que $k \sim \log p$ et pour tout $i, j = 1, \ldots, p$, $\|A(x_i - x_j)\|_2 \sim \|x_i - x_j\|_2$.

En d'autres termes, on peut projeter p points d'un espace de grande dimension dans un espace de dimension seulement $\log p$ tout en conservant les distances euclidiennes d'origines dans $\mathbb{R}^N$. C'est donc un résultat de réduction de dimension.

Plus précisément le résultat s'énonce de la manière suivante : *Il existe une constante absolue $c_0 > 0$ telle que pour tout $0 < \varepsilon < 1$, pour tout ensemble de p points $x_1, \ldots, x_p$ de $\mathbb{R}^N$, et pour tout $k \geq c_0(\log p)/\varepsilon^2$, il existe un opérateur $A : \mathbb{R}^N \to \mathbb{R}^k$ tel que pour tout $i, j = 1, \ldots, p$,*

$$\sqrt{1-\varepsilon}\,\|x_i - x_j\|_2 \leq \|A(x_i - x_j)\|_2 \leq \sqrt{1+\varepsilon}\,\|x_i - x_j\|_2 .$$

****

Correction de l'exercice 2.6 L'argument est probabiliste car il est basé sur les propriétés de concentration des variables de Rademacher (on peut aussi utiliser les variables gaussiennes ; la preuve en est d'ailleurs plus simple mais on souhaite mettre ici en avant l'inégalité de Hoeffding). On considère en effet comme opérateur A la matrice aléatoire dont les entrées sont des variables de Rademacher indépendantes $(\varepsilon_{ij} : 1 \leq i \leq k, 1 \leq j \leq N)$ telles que $\mathbb{P}[\varepsilon_{ij} = 1] = \mathbb{P}[\varepsilon_{ij} = -1] = 1/2$ normalisée par $k^{-1/2}$. On note

$$A = \frac{1}{\sqrt{k}}\big(\varepsilon_{ij}\big)_{\substack{1\leq i\leq k\\ 1\leq j\leq N}} = \frac{1}{\sqrt{k}}\begin{pmatrix} \boldsymbol{\varepsilon}_1^\top \\ \vdots \\ \boldsymbol{\varepsilon}_k^\top \end{pmatrix} : \mathbb{R}^N \to \mathbb{R}^k$$

où $\boldsymbol{\varepsilon}_1, \dots, \boldsymbol{\varepsilon}_k$ sont k vecteurs aléatoires indépendants de Rademacher dans $\mathbb{R}^N$ (càd, dont les coordonnées sont N variables de Rademacher i.i.d.).

Dans un premier temps, on voit que pour tout $x \in \mathbb{R}^N$ on a $\mathbb{E}\langle \boldsymbol{\varepsilon}, x\rangle^2 = \|x\|_2^2$. En effet,

$$\mathbb{E}\langle \boldsymbol{\varepsilon}, x\rangle^2 = \mathbb{E}\Big(\sum_{j=1}^N \varepsilon_j x_j\Big)^2 = \sum_{j_1,j_2=1}^N \mathbb{E}\varepsilon_{j_1}\varepsilon_{j_2}x_{i_1}x_{i_2} = \sum_{j=1}^N x_j^2 = \|x\|_2^2 .$$

On dit que le vecteur aléatoire $\boldsymbol{\varepsilon}$ est isotrope.

On fixe $r, s \in \{1, \dots, p\}$. On s'intéresse au comportement de la norme euclidienne de $A(x_r - x_s)$. On a

$$\mathbb{E}\,\|A(x_r - x_s)\|_2^2 = \mathbb{E}\frac{1}{k}\sum_{i=1}^k \langle \boldsymbol{\varepsilon}_i, x_r - x_s\rangle^2 = \|x_r - x_s\|_2^2 .$$

On a donc pour tout $t > 0$,

$$\mathbb{P}\Big[\big|\,\|A(x_r - x_s)\|_2^2 - \|x_r - x_s\|_2^2\,\big| \geq t\Big] = \mathbb{P}\Big[\Big|\frac{1}{k}\sum_{i=1}^k X_i - \mathbb{E}X_i\Big| \geq t\Big]$$

où $X_i = \langle \boldsymbol{\varepsilon}_i, x_r - x_s\rangle^2$. On a donc (pour ceux qui connaissent les normes d'Orlicz) $\|X_i\|_{\psi_1} = \big\|\langle \boldsymbol{\varepsilon}_i, x_r - x_s\rangle^2\big\|_{\psi_1} = \big\|\langle \boldsymbol{\varepsilon}_i, x_r - x_s\rangle\big\|_{\psi_2}^2$. Par ailleurs, l'inégalité de Hoeffding donne :

$$\big\|\langle \boldsymbol{\varepsilon}_i, x_r - x_s\rangle\big\|_{\psi_2} \leq c_0 \|x_r - x_s\|_2 .$$

On en déduit alors que pour tout $0 < \varepsilon < 1$,

$$\mathbb{P}\Big[\big|\,\|A(x_r - x_s)\|_2^2 - \|x_r - x_s\|_2^2\,\big| \geq \varepsilon\,\|x_r - x_s\|_2^2\Big] \leq 2\exp\big(-ck\varepsilon^2\big).$$

Le résultat précédent ne porte que sur un seul couple (x_r, x_s). Il y a au plus p^2 tels couples. En appliquant la borne de l'union, on a

$$\mathbb{P}\Big[\forall 1 \leq s, r \leq p : \big|\,\|A(x_r - x_s)\|_2^2 - \|x_r - x_s\|_2^2\,\big| \geq \varepsilon\,\|x_r - x_s\|_2^2\Big] \leq 2p^2\exp\big(-ck\varepsilon^2\big).$$

En particulier, quand $k \geq (2/c)(\log p)/\varepsilon^2$, on a avec probabilité au moins $1/2$, pour tout $1 \leq s, r \leq p$,

$$(1-\varepsilon)\,\|x_s - x_r\|_2^2 \leq \|A(x_r - x_s)\|_2^2 \leq (1+\varepsilon)\,\|x_r - x_s\|_2^2$$

****

Exercice 2.7 (Le problème du collectionneur de coupons)

On dispose d'une urne contenant n coupons différents. A chaque tirage, on choisit au hasard un coupon qui est ensuite replacé dans l'urne. On se demande combien de tirages doivent être effectués pour être à peu prêt sûr (càd avec probabilité au moins $1/2$) d'avoir tiré chacun des coupons au moins une fois. La réponse est un $\mathcal{O}(n \log n)$.

On rappelle que la loi géométrique de paramètre $0 < \delta < 1$ est la loi du nombre N du premier succès dans une suite de Bernoulli indépendantes de paramètre δ. On peut la définir par

$$N = \min(n : \delta_n = 1)$$

où $(\delta_n)_n$ est une suite de Bernoulli indépendantes de paramètre δ.

1 **Déterminer la loi, l'espérance, la variance et les déviation de N : $\mathbb{P}[N > k]$ pour tout $k \in \mathbb{N}^*$.**

Soit $i \in \{1, \dots, n\}$. On suppose qu'on vient juste d'observer $i-1$ coupons différents. On se demande combien de tirages vont être nécessaires pour observer un nouveau coupon. On note par T_i ce nombre de tirages ; c'est-à-dire le premier instant où on observe un i-éme coupon différent des $i-1$ précédemment observés.

2 **Quelle est la loi de T_i ? Quelle est la moyenne de T_1 et T_n ? (Interpréter).**

3 **On note par T l'instant de la première fois où on a observé tous les coupons de l'urne. Déterminer T en fonction des T_i. Déterminer sa moyenne.**

4 **Calculer la variance de T et, grâce à l'inégalité de Chebishev, majorer la probabilité que T dévie de sa moyenne.**

****

Correction de l'exercice 2.7

1. On a que pour tout $k \in \mathbb{N}^*$

$$\mathbb{P}\big[N = k\big] = (1-\delta)^{k-1}\delta.$$

Notamment, $\mathbb{E}N = \delta^{-1}$, $\mathrm{var}(N) = (1-\delta)\delta^{-2}$et pour tout $k \in \mathbb{N}^*$

$$\mathbb{P}\big[N > k\big] = 1 - \sum_{i=1}^{k}(1-\delta)^{k-1}\delta = (1-\delta)^k = \exp\Big(-k\log\Big(\frac{1}{1-\delta}\Big)\Big),$$

c'est-à-dire, N est une variable sous-exponentielle.

2. La probabilité d'observer un nouveau coupon est à ce stade $(n-i+1)/n$. Donc T_i est une loi géométrique de paramètre

$$\delta_i = 1 - \frac{i-1}{n}.$$

On a logiquement $\delta_1 = 1$: on est sûr d'observer un nouveau coupon lors du premier tirage et on voit que $\delta_n = 1/n$ donc, une fois observer $n-1$ coupons différents, il faudra en moyenne n nouveaux tirages juste pour observer le dernier coupon (car $\mathbb{E}T_n = \delta_n^{-1} = n$).

3. Le premier instant où on aura observé au moins une fois tous les n coupons de l'urne est donné par $T = T_1 + T_2 + \ldots + T_n$: c'est une somme de lois géométriques indépendantes de paramètres $\delta_1 = 1, \delta_2 = 1 - 1/n, \cdots, \delta_n = 1/n$.

   On a donc un premier résultat concernant le nombre moyen de tirages à effectuer pour observer au moins une fois tous les coupons de l'urne :

$$\mathbb{E}T = \sum_{i=1}^{n} \mathbb{E}T_i = \sum_{i=1}^{n} \delta_i^{-1} = \frac{n}{n} + \frac{n}{n-1} + \cdots + \frac{n}{1} = n \sum_{i=1}^{n} \frac{1}{i} = nH_n$$

   où $H_n = \sum_{i=1}^{n} i^{-1}$ est le nombre harmonique. On a quand n tend vers l'infini

$$H_n = \log n + \gamma + \frac{1}{2n} + o(1/n)$$

   où $\gamma \sim 0.577$ est la constante d'Euler. On a donc bien

$$\mathbb{E}T = n \log n + \gamma n + \frac{1}{2} + o(1).$$

   C'est-à-dire, en moyenne l'instant du premier tirage où on aura observer tous les coupons de l'urne est en $n \log n$.

4. La variance de T est donnée par

$$\mathrm{var}(T) = \sum_{i=1}^{n} \mathrm{var}(T_i) = \sum_{i=1}^{n} \frac{n(i-1)}{n-i+1} \leq n^2 \sum_{i=1}^{n} \frac{1}{i^2} \leq 2n^2.$$

   L'inégalité de Chebitshev donne, pour tout $t > 0$,

$$\mathbb{P}[|T - \mathbb{E}T| \geq tn] \leq \frac{\mathrm{var}(T)}{(tn)^2} \leq \frac{2}{t^2}$$

   donc avec probabilité plus grande que $1 - 2t^{-2}$, $T \leq n \log n + (\gamma + t)n + 1/2 + o(1)$.

****

Exercice 2.8 (Inégalité de Bernstein avec variance empirique)

Première partie

Soient V, $V_1, \ldots, V_n$ des v.a. i.i.d. à valeurs négatives ou nulles. Soit $\bar{V} = \frac{1}{n} \sum_{i=1}^{n} V_i$.

1. **Montrer que $\exp(u) \leq 1 + u + \frac{u^2}{2}$ pour $u \leq 0$.**
2. **Montrer que $\log \mathbb{E} \exp(sV) \leq s\mathbb{E}V + \frac{s^2}{2}\mathbb{E}(V^2)$ pour tout $s \geq 0$.**
3. **En déduire par l'argument de Chernoff que pour tout $\lambda > 0$, on a**

$$\mathbb{P}\big(\bar{V} - \mathbb{E}V > t\big) \leq \exp\bigg(- \lambda t + \frac{\lambda^2 \mathbb{E}(V^2)}{2n} \bigg).$$

4. **En déduire qu'avec probabilité au moins $1 - \varepsilon$, on a**

$$\bar{V} - \mathbb{E}V \leq \sqrt{\frac{2\mathbb{E}(V^2) \log(1/\varepsilon)}{n}}.$$

Deuxième partie

Soient X, $X_1, \dots, X_n$ des v.a. i.i.d. à valeurs dans $[0,1]$. Soient $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$, $\sigma = \sqrt{\mathbb{E}[(X-\mathbb{E}X)^2]}$ l'écart type de X, $\hat{\sigma} = \sqrt{\frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2}$ l'écart type empirique, $\varepsilon > 0$ et $L = \frac{\log(1/\varepsilon)}{n}$.

5. **En utilisant la question précédente, montrer qu'avec probabilité au moins $1-\varepsilon$,**

$$\sigma^2 \leq \frac{1}{n}\sum_{i=1}^n (X_i - \mathbb{E}X)^2 + \sigma\sqrt{2L}.$$

6. **Montrer qu'avec probabilité au moins $1-2\varepsilon$, on a**

$$\left|\bar{X} - \mathbb{E}X\right| \leq \sigma\sqrt{2L} + \frac{L}{3} \tag{3}$$

et qu'avec probabilité au moins $1-3\varepsilon$, on a simultanément (3) et

$$\sigma^2 \leq \hat{\sigma}^2 + \left|\bar{X} - \mathbb{E}X\right|^2 + \sigma\sqrt{2L} \tag{4}$$

7. **L'objectif de cette question est de montrer que lorsque (3) et (4) sont vraies, on a**

$$\sigma \leq \hat{\sigma} + 1,8\sqrt{L}. \tag{5}$$

7.a Montrer que (5) est vraie pour $1,8\sqrt{L} \geq \frac{1}{2}$.

7.b Pour $1,8\sqrt{L} < \frac{1}{2}$, montrer

$$\sigma^2 \leq \hat{\sigma}^2 + \frac{\sqrt{L}\sigma}{3,6} + \frac{2}{3\times(3,6)^2}\sigma\sqrt{2L} + \frac{L}{9\times(3,6)^2} + \sigma\sqrt{2L},$$

et en déduire $\sigma \leq \frac{1,77}{2}\sqrt{L} + \sqrt{\hat{\sigma}^2 + 0,8L}$.

7.c Conclure.

8 **Déduire des questions précédentes qu'avec probabilité au moins $1-\varepsilon$, on a**

$$\left|\bar{X} - \mathbb{E}X\right| \leq \hat{\sigma}\sqrt{\frac{2\log(3\varepsilon^{-1})}{n}} + \frac{3\log(3\varepsilon^{-1})}{n},$$

et comparer avec l'inégalité (3).

****

Correction de l'exercice 2.8

1. étudier la fonction $u \mapsto \exp(u) - 1 - u - \frac{u^2}{2}$.
2. appliquer la question 1) en utilisant également $\log(1+u) \leq 1+u$.
3. $\mathbb{P}(\bar{V} - \mathbb{E}V > t) = \mathbb{P}\big(e^{\lambda(\bar{V}-\mathbb{E}V)} > e^{\lambda t}\big) \leq e^{-\lambda(t+\mathbb{E}V)}\mathbb{E}[e^{\lambda\bar{V}}] \leq e^{-\lambda t}e^{\frac{\lambda^2\mathbb{E}(V^2)}{2n}}$ en appliquant le résultat de la question 2) et le caractère i.i.d des v.a. $V_1, \dots, V_n$.
4. Pour $\mathbb{E}(V^2) = 0$, le résultat est trivial, sinon optimiser en λ (donc prendre $\lambda = nt/\mathbb{E}V^2$), et prendre t tel que $e^{-nt^2/(2\mathbb{E}V^2)} = \varepsilon$.

5. On applique 4) pour $V = -(X - \mathbb{E}X)^2$ en utilisant que $\mathbb{E}V^2 \leq \sigma^2$.
6. On écrit l'inégalité de Bernstein classique pour X et $-X$, et on applique la borne de l'union associée aux deux inégalités ainsi obtenues.

   La deuxième inégalité est juste l'inégalité de la question 5) après développement et simplification de la somme. Par la borne de l'union, puisque l'inégalité de la question 5) et l'inégalité (1) sont respectivement valables avec probabilité au moins $1-\varepsilon$ et $1-2\varepsilon$, on a bien (1) et (2) simultanément vraies sur un événement de probabilité supérieure ou égale à $1 - 3\varepsilon$.

7.a La variance d'une v.a. à valeurs dans $[0, 1]$ est inférieure ou égale à $1/2$.

7.b Utiliser (1) à l'intérieur de (2) et utiliser (lorsque nécessaire) $\sqrt{L} < \frac{1}{3,6}$ et $\sigma \leq 1/2$. Pour la deuxième partie, on se ramène à l'inégalité d'ordre 2 : $\sigma^2 - 1,77\sqrt{L}\sigma - (\hat{\sigma}^2 + \frac{L}{100}) \leq 0$.

7.c En utilisant $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ pour a et b positifs, on montre l'inégalité (3) en constatant que $1,77/2 + \sqrt{0,8} < 1,8$.

7. On combine les inégalités (1) et (3), et conclut en notant que $1,8\sqrt{2} + 1/3 < 3$. Au prix d'une constante plus grande (3 au lieu de 1/3) et d'un niveau de confiance légèrement différent, on a donc réussi à montrer que (1) reste vraie lorsque la variance est remplacée par son équivalent empirique, c'est-à-dire une inégalité "à la Bernstein" mais avec la variance empirique. C'est utile pour construire des intervalles de confiance (observables) non asymptotiques plus précis que ceux issus de la borne de Hoeffding.

3 Notions en Compressed Sensing

****

Exercice 3.1 (Notion de mesures incohérentes via le problème des douze pièces)

Cet exercice a pour but de se familiariser avec les notions de mesures, parcimonie, dictionnaire et incohérence au travers du problème des douze pièces. Pour plus de détails, on renvoie le lecteur intéressé à

site web sur le problème des douze pièces.

Le problème des douze pièces s'énonce de la manière suivante : on se donne douze pièces identiques en taille, forme, etc. mais dont une seule parmi les douze a une masse différente des onze autres. On ne sait pas si cette pièce est plus lourde ou plus légère que les autres ; on sait seulement qu'il y en a une qui est de masse différente des autres. On dispose d'un balance à deux plateaux. Comment retrouver cette pièce parmi les douze en ne faisant que 3 mesures au plus.

1. **Modéliser ce problèmes dans le cadre du CS et indiqué les différences avec le cadre classique du CS ;**
2. **Construire les mesures "triviales" permettant de résoudre le problème en 7 pesées ;**
3. **Définir la notion de parcimonie et le dictionnaire associé pour ce problème ;**
4. **proposer 3 vecteurs de mesure (non-adaptatifs) qui permettent de résoudre ce problème.**

****

Correction de l'exercice 3.1

1. On peut modéliser ce problème dans le cadre du CS de la manière suivante. On définit $x \in \mathbb{R}^{12}$ comme étant le vecteur des poids des 12 pièces. Le "signal" x est de la forme $x_i = p_0$ pour tout $i \in \{1, \dots, 12\} - \{i_1\}$ sauf pour un indice i_1 pour lequel $x_{i_1} = p_1$, où $p_0 \neq p_1$. La i-ième pesée est donc associée à un vecteur de mesure $X_i \in \mathbb{R}^{12}$ où

$$(X_i)_j = \begin{cases} 1 & \text{si la pièce } j \text{ est placée dans le plateau de gauche,} \\ -1 & \text{si la pièce } j \text{ est placée dans le plateau de droite,} \\ 0 & \text{si la pièce } j \text{ n'est pas utilisée au cours de la pesée } i. \end{cases}$$

   On observe ici seulement le signe des mesures $\langle x, X_i \rangle$ càd

$$\text{signe}(\langle x, X_i \rangle) = \begin{cases} 1 & \text{quand la balance penche à gauche,} \\ -1 & \text{quand la balance penche à droite,} \\ 0 & \text{quand la balance reste à l'équilibre.} \end{cases}$$

   Contrairement au cadre du CS, on n'observe que le signe des mesures $\langle x, X_i \rangle$. De plus, on ne souhaite pas reconstruire x mais seulement déterminer l'endroit où x a une coordonnée différente, càd, on cherche à reconstruire son support.

   Ces différences avec l'énoncé classique du CS, comme introduit dans les chapitres précédents, ont été étudiées ces dernières années : l'observation du signe des mesures est une forme de quantification des mesures qui s'appelle le "*one bit compressed sensing*" ; le problème qui consiste à ne retrouver que le support de x est appelé le problème de *reconstruction de support*. On ne présente pas ces deux variantes du CS dans le cours. Cependant on retrouve dans tous ces problèmes le rôle centrale de l'incohérence des mesures avec la base (resp., le dictionnaire) dans laquelle (respectivement, dans lequel) est exprimé la parcimonie du signal d'intérêt.

   Par ailleurs, l'énoncé classique du problème des douze pièces autorise des mesures **adaptatives**, càd pour lesquelles on est autorisé à adapter notre plan d'expérience (càd à choisir les vecteurs mesures) en fonction des résultats précédents. Ce type de plan d'expérience n'est pas le cadre classique du CS qui ne considère que des mesures "non-adaptatives" (càd dont le choix des vecteurs mesures ne dépend pas des résultats précédents). On verra en fait qu'il existe une solution non-adaptative au problème des douze pièces, càd les 3 mêmes mesures sont utilisables dans tous les cas.

2. L'approche triviale en 7 pesées consiste à prendre les pièces deux à deux et à les mettre chacune sur un plateau. Dans le pire cas, il faut 7 mesures ; en effet, quand la pièce de masse différente se trouve dans le dernier lot de deux pièces, il faut effectuer 6 pesées pour identifier le couple contenant la pièce de masse différente puis il faut une dernière pesée pour déterminer laquelle de ces deux pièces a une masse différente de celle des autres. Dans cette stratégie, les vecteurs de mesure sont, pour les 6 premiers,

$$(1, -1, 0, \dots, 0)^\top, (0, 0, 1, -1, 0, \dots, 0)^\top, \cdots, (0, \dots, 0, 1, -1)^\top$$

et le septième est construit suivant la mesure qui a permit d'identifier quel couple contient la pièce de masse différente.

3. La stratégie précédente nécessite beaucoup trop de mesures car les vecteurs de mesure ne sont pas incohérents avec la "base" dans laquelle le signal x est sparse. En effet, il nous faut d'abord identifier quelle est la structure de petite dimension associée à ce problème. Ici le vecteur sparse est le gradient de x

$$\nabla x = (x_2 - x_1, \dots, x_{12} - x_{11})^\top \in \mathbb{R}^{11}.$$

En effet, on voit que $\|\nabla x\|_0 \leq 2$.

Essayons maintenant de traduire cette notion de parcimonie de ∇x en une notion de parcimonie sur x pour un bon choix de dictionnaire (l'objectif, ensuite, est de construire des vecteurs incohèrents avec les éléments de ce dictionnaire). On peut écrire l'opérateur de différentiation discret ∇ sous forme matricielle :

$$\nabla = \begin{pmatrix} -1 & 1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & -1 & 1 & 0 & \cdots & 0 & 0 \\ \vdots & & & & \cdots & \vdots & \\ 0 & 0 & 0 & 0 & \cdots & -1 & 1 \end{pmatrix} : \mathbb{R}^{12} \mapsto \mathbb{R}^{11} \text{ et pour } f_0 = \begin{pmatrix} 1 \\ 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} \in \mathbb{R}^{12}$$

on obtient la formule d'intégration discrète :

$$x = \nabla^{-1} \nabla x + x_1 f_0 \tag{6}$$

où $\nabla^{-1} : \mathbb{R}^{11} \mapsto \mathbb{R}^{12}$ est l'opérateur d'intégration donnée sous forme matricielle par

$$\nabla^{-1} = \begin{pmatrix} 0 & 0 & 0 & 0 & \cdots & 0 & 0 \\ 1 & 0 & 0 & 0 & \cdots & 0 & 0 \\ 1 & 1 & 0 & 0 & \cdots & 0 & 0 \\ 1 & 1 & 1 & 0 & \cdots & 0 & 0 \\ \vdots & & & & \cdots & \vdots & \\ 1 & 1 & 1 & 1 & \cdots & 1 & 1 \end{pmatrix} \in \mathbb{R}^{12 \times 11}$$

Au passage, on peut faire le parallèle entre les objets précédemment introduits et la différentiation et l'intégration d'une fonction à variable continue : Δ^{-1} est l'opérateur d'intégration discrète, ∇ est l'opérateur de différentiation discrète et f_{12} joue le rôle de la "fonction" constante égale à 1. Ainsi (6) est la formulation discrète de la formule d'intégration

$$f(v) = \int_0^v f'(t)dt + f(0), \text{ pour tout } v > 0.$$

Le dictionnaire dans lequel x est s-sparse est donné par les colonnes de ∇^{-1} (la composante sur f_0 étant toujours non nulle). On définit alors le dictionnaire $\{f_1, \dots, f_{11}\}$ formé des colonnes de

∇^{-1} :

$$f_1 = \begin{pmatrix} 0 \\ 1 \\ 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}, f_2 = \begin{pmatrix} 0 \\ 0 \\ 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}, \cdots, f_{11} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 0 \\ 1 \end{pmatrix}.$$

En effet, on a

$$x = (\nabla x)_1 f_1 + \cdots (\nabla x)_{11} f_{11} + x_1 f_0$$

où les coefficients $(\nabla x)_i, i = 1, \ldots, 11$ sont les coordonnées du gradient ∇x et sont donc s-sparse avec $s \in \{1, 2\}$.

4. Une solution (non-adaptative) est donnée par :

$$X_1 = \begin{pmatrix} 1 \\ 0 \\ -1 \\ 1 \\ 0 \\ -1 \\ 1 \\ 0 \\ -1 \\ 1 \\ 0 \\ -1 \end{pmatrix}, \quad X_2 = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ -1 \\ -1 \\ -1 \\ -1 \\ 1 \\ 0 \end{pmatrix} \text{ et } X_3 = \begin{pmatrix} -1 \\ -1 \\ -1 \\ 1 \\ 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ -1 \\ 1 \\ 0 \end{pmatrix}$$

Pour chaque pesée, on choisit quatre pièces pour le plateau de gauche et quatre pour le plateau de droite. Quatre pièces sont laissées de côté à chaque mesure.

On peut vérifier que ces mesures sont bien effectives sur quelques exemples.

Quand la pièce 1 est plus lourde : le résultat des pesées est donné par

$$\text{pesée } 1 : 1 \quad \text{pesée } 2 : 1 \quad \text{pesée } 2 : -1$$

De ces pesées, on en déduit que 1 est bien plus lourde que les autres.

Quand la pièce 11 est plus légère : le résultat des pesées est donné par

$$\text{pesée } 1 : 0 \quad \text{pesée } 2 : -1 \quad \text{pesée } 2 : -1$$

De ces pesées, on en déduit que 11 est bien plus légère.

****

Exercice 3.2 (Principe d'incertitude discret et Dirac Comb)

La matrice de Fourier (transformée de Fourier discrète) est donnée par :

$$\Gamma : \left\{ \begin{array}{ccc} \mathbb{C}^N & \to & \mathbb{C}^N \\ x & \to & \Gamma x = \hat{x} \end{array} \right. \quad \textbf{où } \Gamma = \Big(\frac{w^{(p-1)(q-1)}}{\sqrt{N}} \Big)_{1 \leq p,q \leq N} \textbf{ et } w = \exp(-2\mathbf{i}\pi/N).$$

On note par $\bar{\Gamma}_1, \ldots, \bar{\Gamma}_N$ les vecteurs lignes de Γ. On a donc $\hat{x}_i = \langle \Gamma_i, x \rangle$ pour tout $i = 1, \ldots, N$.

Le principe d'incertitude discret dit que si $x \in \mathbb{C}^N$ est non nul alors

$$|\mathrm{supp}(x)| \times |\mathrm{supp}(\hat{x})| \geq N.$$

En particulier, si x est un vecteur à petit support alors nécessairement sa transformée de Fourier $\hat{x}$ a un grand support. C'est la dualité classique entre vecteurs parcimonieux et transformée de Fourier "bien étalée".

L'objectif de l'exercice est de démontrer ce principe d'incertitude et de montrer qu'il est optimal pour les signaux *Dirac comb*.

1. **Écrire l'inégalité de Plancherel discrète.**
2. **Montrer que $\|x\|_\infty \leq \frac{1}{\sqrt{N}} \|\hat{x}\|_1$**
3. **Déduire le résultat grâce aux inégalités**

$$\|x\|_2^2 \leq |\mathrm{supp}(x)| \, \|x\|_\infty^2 \textbf{ et } \|\hat{x}\|_1 \leq \sqrt{|\mathrm{supp}(\hat{x})|} \, \|\hat{x}\|_2 .$$

4. **On suppose que $N = k^2$ pour un certain entier k. On définit le signal x appelé *Dirac comb* comme étant l'indicateur de l'ensemble $\{0, k, 2k, \cdots, k^2 - k\}$ (ici les coordonnées sont indexées par $\{0, 1, \ldots, N-1\}$).**

   4.a Montrer que pour le Dirac comb, on a $x = \hat{x}$

   4.b Montrer que pour le Dirac comb, on a $|\mathrm{supp}(x)| \times |\mathrm{supp}(\hat{x})| = N$

****

Correction de l'exercice 3.2

1. L'inégalité de Plancherel s'écrit ici $\|x\|_2^2 = \|\hat{x}\|_2^2$. Ce qui traduit que la transformée de Fourier Γ est une isométrie.
2. $\|x\|_\infty = \max_{1 \leq j \leq N} |\langle e_j, x \rangle|$ où $(e_1, \ldots, e_N)$ est la base canonique de $\mathbb{C}^N$. Comme Γ est une isométrie, on a

$$\|x\|_\infty = \max_{1 \leq j \leq N} |\langle e_j, x \rangle| = \max_{1 \leq j \leq N} |\langle \Gamma e_j, \Gamma x \rangle| = \max_{1 \leq j \leq N} \|\Gamma e_j\|_\infty \, \|\Gamma x\|_1 = \frac{1}{\sqrt{N}} \|\hat{x}\|_1 .$$

3. On note I_x (resp. $I_{\hat{x}}$) le support de x (resp. $\hat{x}$). On a

$$\|x\|_2^2 \leq |I_x| \, \|x\|_\infty^2 \leq \frac{|I_x| \, \|\hat{x}\|_1^2}{N} \leq \frac{|I_x||I_{\hat{x}}| \, \|\hat{x}\|_2^2}{N} = \frac{|I_x||I_{\hat{x}}| \, \|x\|_2^2}{N}.$$

   Le résultat s'en déduit quand $x \neq 0$.

4.a Soit $j \in \{0, 1, \ldots, N-1\}$, on a

$$\hat{x}_j = \frac{1}{\sqrt{N}} \sum_{r=0}^{N-1} w^{jr} x_r = \frac{1}{\sqrt{N}} \sum_{r=0}^{k-1} \exp(-2\mathbf{i}\pi jrk/k^2) = \frac{1}{k} \sum_{r=0}^{k-1} \exp(-2\mathbf{i}\pi jr/k)$$
$$= \begin{cases} 1 & \text{si } j \in \{0, k, 2k, \cdots, k^2-k\} \\ 0 & \text{sinon} \end{cases}$$

où $w = \exp(-2\mathbf{i}\pi/N)$. On a donc bien $x = \hat{x}$.

4.b On a $|\text{supp}(x)| = k$ et donc $|\text{supp}(x)| \times |\text{supp}(\hat{x})| = k^2 = N$.

****

Exercice 3.3 (Parcimonie du Basis Pursuit)
Soit $A \in \mathbb{R}^{m \times N}$ et $y \in \mathbb{R}^m$. On suppose qu'il existe une seule et unique solution au problème

$$\mathbf{argmin}_{At=y} \|t\|_1 = \{\hat{x}\}. \tag{7}$$

On note par $a_1, \cdots, a_N$ les vecteurs colonnes de A dans $\mathbb{R}^m$.

1. **Montrer que le système $\{a_j : j \in \text{supp}(\hat{x})\}$ est linéairement indépendant.**
2. **En déduire que $\|\hat{x}\|_0 \leq m$.**

****

Correction de l'exercice 3.3

1. On raisonne par l'absurde. On note $J_{\hat{x}} \subset \{1, \ldots, N\}$ le support de $\hat{x}$. On suppose que la famille $\{a_j : j \in J_{\hat{x}}\}$ est liée. Soit $v \in \mathbb{R}^{J_{\hat{x}}}$ non nul tel que $\sum_{j \in J_{\hat{x}}} v_j a_j = 0$.
   Par unicité, on a pour tout $t \neq 0$, $\|\hat{x}\|_1 < \|\hat{x} + tv\|_1$ car $Av = 0$ et donc $A(\hat{x} + tv) = A\hat{x} = y$.
   Or, on a
   $$\|\hat{x} + tv\|_1 = \sum_{j \in J_{\hat{x}}} \text{sgn}(\hat{x}_j + tv_j)(\hat{x}_j + tv_j).$$
   Pour t tel que $|t| < \min_{j \in J_{\hat{x}}} |\hat{x}_j| / \|v\|_\infty$, on a $\text{sgn}(\hat{x}_j + tv_j) = \text{sgn}(\hat{x}_j)$ alors
   $$\|\hat{x} + tv\|_1 = \|\hat{x}\|_1 + t \sum_j v_j \text{sgn}(\hat{x}_j).$$
   Mais comme $\|\hat{x} + tv\|_1 > \|\hat{x}\|_1$, on a donc pour tout $|t| < \min_{j \in J_{\hat{x}}} |\hat{x}_j| / \|v\|_\infty$, $t \sum_j v_j \text{sgn}(\hat{x}_j) > 0$. Ce qui est une contradiction.
2. $\{a_j : j \in \text{supp}(\hat{x})\}$ est une système de vecteurs de $\mathbb{R}^m$ linéairement indépendant. Il ne peut pas y en avoir plus de m. Donc $\|\hat{x}\|_0 \leq m$.

****

Exercice 3.4 (RIP et nombre minimal de mesures)
Soit $A \in \mathbb{R}^{m \times N}$ telle que pour tout $x \in \Sigma_{2s}$, $(1/2)\|x\|_2 \leq \|Ax\|_2 \leq (3/2)\|x\|_2$. Alors il existe des constantes absolues $c_0, c_1 > 0$ telles que $m \geq c_0 s \log(c_1 N/s)$.

****

Correction de l'exercice 3.4 Pour tout $x \in \Sigma_{2s} \cap B_2^N$, on a $Ax \in (3/2)B_2^m$. De plus, pour tout $x, y \in \Sigma_s \cap B_2^N$, si $\|x-y\|_2 \geq \varepsilon$ alors $\|Ax - Ay\|_2 \geq (1/2)\|x-y\|_2 \geq \varepsilon/2$.

On en déduit que si Λ_ε est un ensemble ε-écarté de $\Sigma_{2s} \cap B_2^N$ pour ℓ_2^N alors $A\Lambda_\varepsilon$ est un ensemble $(\varepsilon/2)$-écarté de $(3/2)B_2^m$ pour ℓ_2^m.

Or, on sait qu'il est possible de construire une ensemble $(1/4)$-écarté de $\Sigma_{2s} \cap B_2^N$ pour ℓ_2^N de cardinal au moins $(c_0 N/s)^{s/c_1}$ où c_0, c_1 sont deux constantes absolues. Donc $(3/2)B_2^m$ contient un ensemble $1/8$-écarté pour ℓ_2^m de cardinal $(c_0 N/s)^{s/c_1}$. Or on sait qu'un tel ensemble doit forcément avoir un cardinal plus petit que 17^m.

Note : On sait que $RIP(65s)$ implique $RE(s)$ et que $RE(s)$ implique que $m \geq c_0 s \log(c_1 N/s)$.

****

Exercice 3.5 (Robustesse et borne inférieure de RIP)
Soit $A \in \mathbb{R}^{m\times N}$ une matrice de compression et $\Delta : \mathbb{R}^m \to \mathbb{R}^N$ un algorithme de décompression (non nécessairement linéaire). On dit que (A, Δ) est C-robuste d'ordre s quand pour tout $x \in \Sigma_s$ et tout $e \in \mathbb{R}^m$, on a

$$\|\Delta(Ax+e) - x\|_2 \leq C\|e\|_2 .$$

En particulier, si (A, Δ) est C-robuste d'ordre s alors pour tout $x \in \Sigma_s$, on a $\Delta(Ax) = x$. Donc Δ permet la reconstruction exacte de tout élément x de Σ_s à partir de Ax. La robustesse demande en plus que Δ s'adapte aux erreurs dans les observations. C'est-à-dire que le signal reconstruit $\Delta(Ax+e)$ à partir des observations bruitées $Ax+e$ soit proche de x à un terme de l'ordre de grandeur du bruit près.

Montrer que si (A, Δ) est C-robuste d'ordre s alors la borne inférieure de $RIP(2s)$ est nécessaire, càd, pour tout $x \in \Sigma_{2s}$

$$\|Ax\|_2 \geq \frac{1}{2}\|x\|_2 .$$

****

Correction de l'exercice 3.5 Soit $x, y \in \Sigma_s$. On note

$$e_y = \frac{A(x-y)}{2} \text{ et } e_x = \frac{A(y-x)}{2}.$$

On a $Ax + e_x = Ay + e_y$ donc $\hat{x} = \Delta(Ax+e_x) = \Delta(Ay+e_y) = \hat{y}$. On obtient :

$$\|x-y\|_2 = \|x - \hat{x} + \hat{y} - y\|_2 \leq \|x-\hat{x}\|_2 + \|y-\hat{y}\|_2 \leq C\|e_x\|_2 + C\|e_y\|_2 = C\|A(x-y)\|_2 .$$

****

Exercice 3.6 (ℓ_2-stabilité)
Soit $A \in \mathbb{R}^{m\times N}$ une matrice de mesures et $\Delta : \mathbb{R}^m \to \mathbb{R}^N$ une procédure de décompression. On dit que (A, Δ) est C-ℓ_2-stable d'ordre s quand pour tout $x \in \mathbb{R}^N$,

$$\|x - \Delta(Ax)\|_2 \leq C\sigma_{s,2}(x) \text{ où } \sigma_{s,2}(x) = \min_{z\in\Sigma_s} \|x-z\|_2 . \tag{8}$$

On va montrer que s'il existe (A, Δ) C-ℓ_2-stable d'ordre s (pour n'importe quel s, même $s = 1$) alors nécessairement $m \geq \big(1 - \sqrt{C^2 - 1}/C\big)N$.

1. **Montrer que pour tout $x \in \ker(A)$ et tout $j \in \{1, \dots, N\}$, $|x_i| \leq C' \|x\|_2$ pour $C' = \sqrt{C^2 - 1}/C$.**
2. **Soit L un sev de $\mathbb{R}^N$ tel que pour tout $x \in L$ et pour tout $i \in \{1, \dots, N\}$, $|x_i| \leq C' \|x\|_2$. Montre que $\dim(L) \leq C'N$.**
3. **En déduire que $m \geq (1 - \sqrt{C^2 - 1}/C)N$**

Note : La stabilité d'une procédure est importante quand les signaux à reconstruire ne sont pas exactement parcimonieux mais seulement proche de Σ_s – ces signaux sont dit *compressibles*. Le résultat de cet exercice dit qu'on ne peut pas avoir (8) sauf si on a un nombre de mesures de l'ordre de N – ce qui n'est pas envisageable en CS. C'est pour cette raison qu'on a introduit la stabilité par rapport à $\sigma_{s,1}(x)$.

****

Correction de l'exercice 3.6

1. On a $\Delta(0) = 0$ donc pour tout $x \in \ker(A)$, on a $\|x\|_2 \leq C\sigma_{s,2}(x)$. Pour $s = 1$, on a pour tout $x \in \ker(A)$ et tout $j = 1, \dots, N$
$$\|x\|_2^2 \leq C^2\big(\|x_2\| - |x_j|^2\big)$$
donc $|x_j|^2 \leq (C^2 - 1)/C^2 \|x\|_2^2$ càd $|x_j| \leq C' \|x\|_2$ pour $C' = \sqrt{C^2 - 1}/C$.
2. Soit P la projection sur L. On note par $(e_1, \dots, e_N)$ la base canonique de $\mathbb{R}^N$. Pour tout j, $Pe_j \in L$ donc, par hypothèse, $(Pe_j)_i \leq C' \|Pe_j\|$. On a
$$\dim(L) = \mathrm{Tr}(P) = \sum_{j=1}^{N} \langle Pe_j, e_j \rangle \leq \sum_{j=1}^{N} C' \|Pe_j\|_2 \leq C'N.$$
3. On a donc $N - m \geq \dim(\ker(A)) \leq C'N$ et donc $m \geq \big(1 - \sqrt{C^2 - 1}/C\big)N$.

4 Optimisation

****

Exercice 4.1 (Optimisation convexe sous contraintes linéaires)

Soit le problème d'optimisation suivant

$$\min\Big(x^2 + yz + y^2 + z^2 : x \leq -1, z \leq -1\Big).$$

1. **Réduire ce problème de minimisation à un problème de la forme**
$$\min_{t \in K} f(t) \qquad (P)$$
où f est une fonction à valeurs réelles et
$$K = \{t : g(t) \leq 0\}$$
avec $g : \mathbb{R}^2 \to \mathbb{R}$ convexe de classe $\mathcal{C}^1$.

2. **Montrer que f est convexe.**
3. **Montrer que la contrainte K est qualifiée.**
4. **Écrire le lagrangien de (P).**
5. **Écrire les conditions KKT de (P) et résoudre (P) à partir de ces équations.**
6. **Appliquer le théorème de dualité. Écrire le problème dual de (P), et retrouver ainsi la solution de (P) obtenue à la question 5).**

****

Correction de l'exercice 4.1

1. Dans un premier temps, on peut facilement minimiser la fonction en x, le minimum étant atteint en $x = -1$. Il reste donc à déterminer le minimum en y et z. On est donc amener à résoudre le problème suivant :
$$(P) \qquad \min_{t \in K} f(t)$$
où f est une fonction de $\mathbb{R}^2$ dans $\mathbb{R}$ donnée par $f(y,z) = yz + y^2 + z^2$ et
$$K = \{(y,z)^\top \in \mathbb{R}^2 : g(y,z) \leq 0\}$$
avec $g(y,z) = z + 1$ convexe de classe $\mathbb{C}^1$
2. On a $f(y,z) = (y + z/2)^2 + 3z^2/4$. C'est une somme de carré de formes linéaires donc c'est une fonction convexe.
3. Soit $(y,z)^\top \in K$. On a $\nabla g(y,z) = (0,1)^\top$ alors pour $v = (0,-1)^\top$, on a bien $\langle \nabla g(y,z), v \rangle < 0$. Donc K est qualifiée en $(y,z)^\top \in K$ et ceci étant vrai pour tout point $(y,z)^\top$ de K, K est qualifiée.
4. Le Lagrangien de (P) est donné pour tout $(y,z)^\top \in \mathbb{R}^2$ et $\lambda \geq 0$ par
$$L((y,z),\lambda) = f(y,z) + \lambda g(y,z) = yz + y^2 + z^2 + \lambda(z+1).$$
5. Les conditions KKT du problème sont : si $(y^*,z^*)^\top \in \mathbb{R}^2$ est solution de (P) alors il existe $\lambda^* \geq 0$ tel que :
   (a) $(y^*,z^*)^\top \in K$,
   (b) $\lambda^* g(y^*,z^*) = 0$,
   (c) $\nabla_{(y,z)} L((y^*,z^*),\lambda^*) = 0$.

   On obtient finalement que $y^* = 1/2$ et $z^* = -1$. Donc la valeur du problème d'optimisation est 1.75 atteint en $(-1, 1/2, -1)^\top$.
6. Le théorème de dualité dit que
$$\min_{t \in K} f(t) = \min_{t \in \mathbb{R}^2} \sup_{\lambda \geq 0} L(t,\lambda) = \sup_{\lambda \geq 0} \min_{t \in \mathbb{R}^2} L(t,\lambda).$$
   De plus, si $\lambda^* \geq 0$ est solution du problème dual
$$\sup_{\lambda \geq 0} d(\lambda) \text{ où } d(\lambda) = \min_{t \in \mathbb{R}^2} L(t,\lambda) = \frac{-\lambda^2}{3} + \lambda$$

alors une solution t^* de (P) est aussi solution de

$$\min_{t\in\mathbb{R}^2} L(t,\lambda^*).$$

On obtient que $\lambda^* = 3/2$ est solution du problème dual et alors $t^* = (1/2,-1)^\top$ est solution du problème primal.

****

Exercice 4.2 (Système d'équations linéaire et optimisation)

Soit A une matrice symétrique définie positive de $\mathbb{R}^{N\times N}$ et $y\in\mathbb{R}^N$. Montrer que les deux assertions suivantes sont équivalentes :

1. **$\hat{x}$ est solution de $Ax = y$**
2. **$\hat{x}$ minimise la fonction $x \to F(x) = \frac{1}{2}x^\top Ax - x^\top y$.**

****

Correction de l'exercice 4.2

Première solution : La Hessienne de F est A qui est symétrique définie positive. Donc $\hat{x}$ minimise F si et seulement si $\nabla F(\hat{x}) = 0$. Or $\nabla F(x) = Ax - y$.

Deuxième solution : Si $Ax_0 = y$ alors pour tout $x\in\mathbb{R}^N$, on a

$$F(x) - F(x_0) = \tfrac{1}{2}x^\top Ax - x^\top y - \tfrac{1}{2}x_0^\top Ax_0 + x_0^\top y = \tfrac{1}{2}\Big(x^\top Ax + x_0^\top Ax_0 - 2x^\top Ax_0\Big). \tag{9}$$

Or par Cauchy-Schwartz, on a

$$2|x^\top Ax_0| = 2|\langle A^{1/2}x, A^{1/2}x_0\rangle| \le 2\left\|A^{1/2}x\right\|_2\left\|A^{1/2}x_0\right\|_2 \le \left\|A^{1/2}x\right\|_2^2 + \left\|A^{1/2}x_0\right\|_2^2 = x^\top Ax + x_0^\top Ax_0.$$

On en déduit que $F(x) - F(x_0) \ge 0$. Donc x_0 est bien le minimum de F.

Réciproquement, si pour tout x, $F(x) - F(x_0) \ge 0$ alors pour tout $h\in\mathbb{R}^N$, on a

$$F(x_0+h) = F(x_0) + \big\langle \nabla F(x_0), h\big\rangle + \frac{1}{2}h^\top \nabla^2 F(x_0)h = F(x_0) + \big\langle Ax_0 - y, h\big\rangle + \frac{\left\|A^{1/2}h\right\|_2^2}{2}$$

et comme $F(x_0+h) \ge F(x_0)$, on en déduit que pour tout h, $2\big\langle Ax_0 - y, h\big\rangle + \left\|A^{1/2}h\right\|_2^2 \ge 0$. Et donc, quitte à remplacer h par λh pour $\lambda > 0$, on voit que $\big\langle Ax_0 - y, h\big\rangle \ge 0$ pour tout h donc $Ax_0 = y$.

****

Exercice 4.3 (Descente de gradient)

On considère la fonction

$$\varphi(t) = \frac{t}{\sqrt{1+t^2}}.$$

On remarque que $\varphi(x^*) = 0$ si et seulement si $x^* = 0$. Décrire la convergence de l'algorithme de Newton pour résoudre l'équation $\varphi(x) = 0$.

****

Correction de l'exercice 4.3 On a pour tout $t \in \mathbb{R}$,

$$\varphi'(t) = (1-t^2)^{-3/2}.$$

Alors l'algorithme de Newton est $(t_k)_k$ pour une certain $t_0 \in \mathbb{R}$ et

$$t_{k+1} = t_k - \frac{\varphi(t_k)}{\varphi'(t_k)} = -t_k^3.$$

On a donc

1. $(t_k)_k$ converge vers 0 quand $|t_0| < 1$
2. $(t_k)_k$ est constante en 1 ou oscille entre -1 et 1 quand $|t_0| = 1$
3. $(t_k)_k$ diverge quand $|t_0| > 1$.

****

Exercice 4.4 (Caractérisation de la convexité forte)

Soit $f : (a,b) \to \mathbb{R}$. On dit que f est fortement convexe de module de convexité $c > 0$ quand pour tout $x, y \in (a,b)$, on a pour tout $t \in [0,1]$,

$$f(tx + (1-t)y) \leq tf(x) + (1-t)f(y) - ct(1-t)(x-y)^2.$$

Montrer que

1. **f est fortement convexe de module de convexité $c > 0$ si et seulement si pour tout $x_0 \in (a,b)$, pour tout x,**

$$f(x) \geq f(x_0) + L(x - x_0) + c(x - x_0)^2$$

où $L \in [f'_-(x_0), f'_+(x_0)]$ sont les dérivées à gauche et droite de f en x_0.

2. **si f est différentiable, alors f est fortement convexe de module de convexité $c > 0$ si et seulement si pour tout $x, y \in (a,b)$**

$$(f'(x) - f'(y))(x - y) \geq 2c(x-y)^2.$$

3. **si f est deux fois différentiable, alors f est fortement convexe de module de convexité $c > 0$ si et seulement si pour tout $x \in (a,b)$**

$$f''(x) \geq 2c.$$

****

Correction de l'exercice 4.4 cf. "Some Characterizations of Strongly Convex Functions in Inner Product Spaces" by Teodoro Lara, Merentes, Rosales, Valear in Mathematica Aeterna

****

Exercice 4.5 (Sous-différentielle et problème de minimisation)
Soit $f : \mathbb{R} \to \mathbb{R}$ une fonction convexe. On rappelle que la sous-différentielle de f en x est

$$\partial^- f(x) = \big\{ g \in \mathbb{R} : f(y) \geq f(x) + \langle g, y - x \rangle \text{ pour tout } y \in \mathbb{R} \big\}.$$

Montrer que x^* est solution du problème $\min_x f(x)$ si et seulement si $0 \in \partial^- f(x^*)$.

****

Correction de l'exercice 4.5 On suppose que pour tout $x \in \mathbb{R}$, on a $f(x) \geq f(x^*)$. On a donc bien que pour tout y, $f(y) \geq f(x^*) + \langle 0, y - x \rangle$ donc $0 \in \partial^- f(x^*)$.

Réciproquement, si $0 \in \partial^- f(x^*)$ alors pour tout y, $f(y) \geq f(x^*) + \langle 0, y - x \rangle = f(x^*)$. Donc x^* est bien solution au problème de minimisation $\min_x f(x)$.

****

Exercice 4.6 (La méthode du simplexe)
Résoudre par la méthode du simplexe le système suivant :

$$\min \left(-5x_1 - 2x_2 - 3x_3 : \begin{array}{rl} x_1 + x_2 & \leq 5 \\ 2x_1 - x_2 + x_3 & \leq 10 \\ x_2 - x_3 & \leq 4 \\ x_1, x_2, x_3 \geq 0 & \end{array} \right)$$

****

Correction de l'exercice 4.6 Le minimum vaut -55, il est atteint par $x_1 = 0$, $x_2 = 5$ et $x_3 = 15$.

z	x_1	x_2	x_3	x_4	x_5	x_6	b
1	5	2	3	0	0	0	0
0	(1)	1	0	1	0	0	5
0	2	-1	1	0	1	0	10
0	0	1	-1	0	0	1	4
1	0	-3	3	-5	0	0	-25
0	1	1	0	1	0	0	5
0	0	-3	(1)	-2	1	0	0
0	0	1	-1	0	0	1	4
1	0	6	0	1	-3	0	-25
0	1	1	0	(1)	0	0	5
0	0	-3	1	-2	1	0	0
0	0	-2	0	-2	1	1	4
1	-1	5	0	0	-3	0	-30
0	1	(1)	0	1	0	0	5
0	2	-1	1	0	1	0	10
0	2	0	0	0	1	1	14
1	-6	0	0	-5	-3	0	-55
0	1	1	0	1	0	0	5
0	3	0	1	1	1	0	15
0	2	0	0	0	1	1	14