Computational statistics
TP 1 : Reminder on Markov Chains – Stochastic gradient descent

Thibault de Surrel

Exercice 1

1.

Soit h une fonction mesurable et bornée. On note $f_{(R,\Theta)}$ la densité jointe du couple de variables aléatoires (R, Θ). Alors, on a

$$\begin{aligned}
\mathbb{E}[h(X,Y)] &= \iint_{\mathbb{R}^2} h(r\cos\theta, r\sin\theta) f_{(R,\Theta)}(r,\theta)\mathrm{d}r\mathrm{d}\theta \\
&= \iint_{\mathbb{R}^2} h(r\cos\theta, r\sin\theta) f_R(r) f_{\Theta)}(\theta)\mathrm{d}r\mathrm{d}\theta \quad \text{car les v.a. } R \text{ et } \Theta \text{ sont indépendantes} \\
&= \int_0^{2\pi}\int_0^{+\infty} h(r\cos\theta, r\sin\theta) r \exp\left(-\frac{r^2}{2}\right)\frac{1}{2\pi}\mathrm{d}r\mathrm{d}\theta
\end{aligned}$$

On effectue un changement de variables des coordonnées polaires aux coordonnées cartésiennes : $(x, y) = (r\cos\theta, r\sin\theta)$ de jacobien $\frac{1}{r}$. Alors, par théorème de changement de variables et comme $r^2 = x^2 + y^2$, on a :

$$\begin{aligned}
\mathbb{E}[h(X,Y)] &= \iint_{\mathbb{R}^2} h(x,y)\frac{1}{2\pi}\exp\left(-\frac{x^2+y^2}{2}\right)\mathrm{d}x\mathrm{d}y \\
&= \iint_{\mathbb{R}^2} h(x,y)\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{x^2}{2}\right)\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{y^2}{2}\right)\mathrm{d}x\mathrm{d}y
\end{aligned}$$

où l'on reconnait le produit de deux densités d'une gaussienne $\mathcal{N}(0, 1)$. Les deux variables aléatoires X et Y sont donc indépendantes et suivent toutes deux une $\mathcal{N}(0, 1)$.

2.

Pour simuler deux variables aléatoires X et Y qui suivent une $\mathcal{N}(0, 1)$, on va utiliser la question précédente et simuler R et Θ. On utilise la méthode de la transformée inverse afin de simuler R, qui consiste à inverser sa fonction de répartition. En notant F_R la fonction de répartition de R, on a :

$$\forall r \geq 0,\ F_R(r) = \int_0^r f_R(t)\mathrm{d}t = 1 - \exp\left(-\frac{r^2}{2}\right)$$

On peut alors l'inverser :

$$u = F_R(r) \Leftrightarrow u = 1 - \exp(-\frac{r^2}{2}) \Leftrightarrow r = \sqrt{-2\log(1-u)}$$

Ainsi, on a, pour tout $u \in [0, 1]$, $F_R^{-1}(u) = \sqrt{-2\log(1-u)}$. L'algorithme pour simuler X et Y est alors l'algorithme 1.

3.

a) L'algorithme de Marsaglia-Bray met en place une méthode de rejet. Ainsi, à la fin de la boucle "while", la variable (V_1, V_2) suit une loi uniforme sur le disque unité : $\mathcal{U}(D(0, 1))$.

Algorithme 1 : Simuler loi gaussienne

Simuler U_1 et U_2 indépendantes et de loi $\mathcal{U}([0,1])$;
Poser $R = \sqrt{-2\log(1-U_1)}$;
Poser $\Theta = 2\pi U_2$;
Renvoyer $(X,Y) = (R\cos\Theta, R\sin\Theta)$;

b) On commence par montrer que les variables aléatoires T_1 et V sont indépendantes. Pour cela, soit h une fonction mesurable bornée, on a, comme (V_1, V_2) suivent une loi uniforme sur le disque unité,

$$\begin{aligned}\mathbb{E}[h(T_1,V)] &= \frac{1}{\pi}\iint_{D(0,1)} h\left(\frac{v_1}{\sqrt{v_1^2+v_2^2}}, v_1^2+v_2^2\right)\mathrm{d}v_1\mathrm{d}v_2 \\ &= \frac{2}{\pi}\iint_{D(0,1)\cap\{v_2\geq 0\}} h\left(\frac{v_1}{\sqrt{v_1^2+v_2^2}}, v_1^2+v_2^2\right)\mathrm{d}v_1\mathrm{d}v_2 \quad \text{par parité de la fonction } h \text{ en } v_2\end{aligned}$$

On pose alors $\phi\colon (v_1,v_2)\mapsto \left(\frac{v_1}{\sqrt{v_1^2+v_2^2}}, v_1^2+v_2^2\right)$ qui est un $\mathcal{C}^1-$difféomorphisme de $D(0,1)\cap\{v_2\geq 0\}\setminus\{(0,0)\}$ vers $]-1,1[\times]0,1]$. En effet, ϕ est de classe $\mathcal{C}^1$ sur $D(0,1)\cap\{v_2\geq 0\}\setminus\{(0,0)\}$ et on a

$$(t,v) = \phi(v_1,v_2) \Leftrightarrow \begin{cases} t = \dfrac{v_1}{\sqrt{v_1^2+v_2^2}} \\ v = v_1^2+v_2^2 \end{cases} \Leftrightarrow \begin{cases} v_1 = \sqrt{v}t \\ v_2 = \sqrt{v(1-t^2)} \end{cases}$$

Ainsi, $\phi^{-1}\colon (t,v)\mapsto (\sqrt{v}t, \sqrt{v(1-t^2)})$ qui est bien de classe $\mathcal{C}^1$ sur $]-1,1[\times]0,1]$. On a de plus

$$J\phi^{-1}(t,v) = \begin{pmatrix} \sqrt{v} & \frac{t}{2\sqrt{v}} \\ \frac{\sqrt{v}t}{\sqrt{1-t^2}} & \frac{\sqrt{1-t^2}}{2\sqrt{v}} \end{pmatrix}$$

Et ainsi, $\mid J\phi^{-1}(t,v)\mid = \frac{1}{2\sqrt{1-t^2}}$. Donc, par théorème de changement de variables, on a :

$$\mathbb{E}[h(T_1,V)] = \frac{1}{\pi}\iint_{]-1,1[\times]0,1]}\frac{h(t,v)}{\sqrt{1-t^2}}\mathrm{d}t\mathrm{d}v = \iint_{\mathbb{R}^2} h(t,v)f_{T_1}(t)f_V(v)\mathrm{d}t\mathrm{d}v$$

où $f_{T_1}: t\mapsto \frac{1}{\pi\sqrt{1-t^2}}\mathbb{1}_{]-1,1[}(t)$ et $f_V: v\mapsto \mathbb{1}_{]0,1]}(v)$. Ainsi, T_1 et V sont indépendantes et V suit une $\mathcal{U}([0,1])$.

Montrons maintenant que T_1 suit la même loi que $cos\Theta$ où $\Theta\sim\mathcal{U}([0,2\pi])$. Pour cela, soit h une fonction mesurable et bornée,

$$\mathbb{E}[h(\cos\Theta)] = \int_0^{2\pi} h(cos(\theta))\frac{1}{2\pi}\mathrm{d}\theta = \frac{1}{2\pi}\int_{-\pi}^{\pi} h(\cos(u+\pi))\mathrm{d}u = \frac{1}{\pi}\int_0^{\pi} h(-cos u)\mathrm{d}u = \frac{1}{\pi}\int_{-1}^{1}\frac{h(t)}{\sqrt{1-t^2}}\mathrm{d}t$$

On a utilisé ci-dessus le changement de variables $u = \theta - \pi$ avec le fait que $\cos(\theta+\pi) = -\cos(\theta)$ puis la parité de cos et enfin le changement de variable $t = \cos u$. Ainsi, en comparant l'expression de f_{T_1} et celle de la densité de $\cos\Theta$, on voit que T_1 et $\cos\Theta$ suivent la même loi. On fait de même pour montrer que T_2 et V sont indépendantes et que T_2 et $\sin\Theta$ suivent la même loi.

c) D'après la question précédente, la variable aléatoire $V = V_1^2 + V_2^2$ suit une loi uniforme sur $[0,1]$. Ainsi, 1_V suit aussi une loi uniforme sur $[0,1]$. Donc, d'après la question 2, S suit la même loi de R, c'est-à-dire une loi de Rayleigh. Toujours d'après la question précédente, (T_1,T_2) suit la même loi que $(\cos\Theta,\sin\Theta)$. Donc, (X,Y) suit la même loi que $(R\cos\Theta, R\sin\Theta)$ et par la question 2, on en déduit que X et Y suivent une loi normale centrée et réduite $\mathcal{N}(0,1)$.

d) La probabilité d'accepter un tirage de (V_1,V_2) est de $\frac{\pi}{4}$. En effet, on veut que le point de coordonnées (V_1,V_2), tiré dans le carré $[-1,1]^2$, tombe dans le cercle unité. Ainsi, la probabilité d'acceptation est égale à l'aire du cercle sur celle du carré, d'où $\frac{\pi}{4}$. Ainsi, le nombre d'étape suit une loi géométrique de paramètre $\frac{\pi}{4}$ car à chaque essai, on fait un tirage de Bernoulli de paramètre $\frac{\pi}{4}$. Le nombre moyen d'itérations est donc $\frac{4}{\pi}$.

Exercice 2

1.

On note $Q = \{\frac{1}{n},\ n \in \mathbb{N}^*\}$ et $(\mathcal{F}_n)_{n\in\mathbb{N}}$ la filtration canoniquement associée à la chaine de Markov $(X_n)_{n\in\mathbb{N}}$, c'est-à-dire que pour tout $n \in \mathbb{N},\ \mathcal{F}_n = \sigma(X_0, ..., X_n)$.

Soit $n \in \mathbb{N}$ tel que $X_n \notin Q$, alors $X_{n+1} \sim \mathcal{U}([0,1])$ donc, par définition de $P(x, A)$,

$$\forall x \in Q^C \cap [0,1],\ \forall A \in \mathcal{B}([0,1]),\ P(x,A) = \int_{[0,1]} \mathbb{1}_A(t)\mathrm{d}t = \int_{[0,1]\cap A} \mathrm{d}t$$

Soit $n \in \mathbb{N}$ tel que $X_n \in Q$, alors pour h une fonction mesurable on a

$$\mathbb{E}[h(X_{n+1} \mid \mathcal{F}_n] = \mathbb{E}\left[h\left(\frac{1}{m+1}\right)(1 - X_n^2) + h(U)X_n^2 \mid \mathcal{F}_n\right]$$

où U suit une loi uniforme sur $[0,1]$ indépendante de X_n. Alors, comme $X_n \in \mathcal{F}_n$ et U est indépendante de $\mathcal{F}_n$, on a que

$$\mathbb{E}[h(X_{n+1} \mid \mathcal{F}_n] = h\left(\frac{1}{m+1}\right)(1 - X_n^2) + \mathbb{E}[h(U)]X_n^2 = h\left(\frac{1}{m+1}\right)(1 - X_n^2) + \left(\int_{[0,1]} h(u)\mathrm{d}u\right)X_n^2 \qquad (1)$$

Ainsi, en prenant $h = \mathbb{1}_A$ où A est un borélien de $[0,1]$, on a

$$\forall x = \frac{1}{m} \in Q,\ \forall A \in \mathcal{B}([0,1]),\ P(x,A) = x^2 \int_{[0,1]\cap A} \mathrm{d}t + \delta_{\frac{1}{m+1}}(A)(1 - x^2)$$

D'où l'expression demandée du noyau de transition P.

2.

On note π la loi uniforme. On veut donc montrer que $\pi = \pi P$. Pour montrer cela, prenons un borélien A de $[0,1]$,

$$\pi P(A) = \int_{\mathbb{R}} \pi(x)P(x,A)\mathrm{d}x = \int_{[0,1]} P(x,A)\mathrm{d}x$$

Or, l'ensemble Q est négligeable pour la mesure de Lebesgue. En effet, en notant μ cette mesure de Lebesgue, on a $\mu(Q) \leq \sum_{n\in\mathbb{N}^*} \mu(\{\frac{1}{n}\}) = 0$ par sous-additivité de la mesure et parce que la mesure de Lebesgue d'un singleton est nulle. Ainsi, on a

$$\pi P(A) = \int_{[0,1]\cap Q^C} P(x,A)\mathrm{d}x = \int_{[0,1]\cap Q^C}\left(\int_{A\cap[0,1]} \mathrm{d}t\right)\mathrm{d}x = \int_{A\cap[0,1]}\left(\int_{[0,1]\cap Q^C} \mathrm{d}x\right)\mathrm{d}t = \int_{A\cap[0,1]} \mathrm{d}t = \pi(A)$$

d'après l'expression de P obtenue à la question précédente et par le théorème de Fubini-Tonelli, les fonctions considérées étant toutes positives, qui nous permet d'échanger les deux intégrales.

La loi uniforme est donc bien invariante pour P.

3.

Soit f une fonction mesurable et bornée. Conditionnellement à $(X_0 = x)$, on a $X_1 \sim \mathcal{U}([0,1])$ car $x \in \int_{[0,1]\cap Q^C}$. Donc, par la formule de transfert on a :

$$Pf(x) := \mathbb{E}[f(X_1) \mid X_0 = x] = \int_{\mathbb{R}} f(t)\pi(t)\mathrm{d}t$$

Par récurrence sur $n \in \mathbb{N}^*$, montrons que $P^n f(x) = \int_{\mathbb{R}} f(t)\pi(t)\mathrm{d}t$. On a déjà initialisé la récurrence. Soit $n \in \mathbb{N}$. Supposons que $P^n f(x) = \int_{\mathbb{R}} f(t)\pi(t)\mathrm{d}t$. Alors, on a

$$\begin{aligned}
P^{n+1}f(x) &= P(P^n f)(x) \\
&= \mathbb{E}[P^n f(X_1) \mid X_0 = x] \\
&= \int_{\mathbb{R}} P^n f(t)\pi(t)\mathrm{d}t \quad \text{par la formule de transfert (comme précédemment)} \\
&= \int_{\mathbb{R}} \left(\int_{\mathbb{R}} f(s)\pi(s)\mathrm{d}s \right) \pi(t)\mathrm{d}t \quad \text{par hypothèse de récurrence} \\
&= \int_{\mathbb{R}} \left(\int_{\mathbb{R}} \pi(t)\mathrm{d}t \right) f(s)\pi(s)\mathrm{d}s \quad \text{d'après le théorème de Fubini} \\
&= \int_{\mathbb{R}} f(s)\pi(s)\mathrm{d}s \quad \text{car } \pi \text{ est un densité de probabilité}
\end{aligned}$$

Le théorème de Fubini s'applique bien car

$$\int_{\mathbb{R}} \int_{\mathbb{R}} |\pi(t)f(s)\pi(s)|\mathrm{d}t\mathrm{d}s = \int_{\mathbb{R}} \left(\int_{\mathbb{R}} \pi(t)\mathrm{d}t \right) |f(s)|\pi(s)\mathrm{d}s = \int_{\mathbb{R}} |f(s)|\pi(s)\mathrm{d}s < +\infty$$

la première égalité étant le théorème de Fubini-Tonelli et l'inégalité stricte finale venant du faite que f est bornée.

Ainsi, on a bien que pour tout $n \in \mathbb{N}^*$, $P^n f(x) = \int_{\mathbb{R}} f(t)\pi(t)\mathrm{d}t$ et ainsi, $\lim_{n\to+\infty} P^n f(x) = \int_{\mathbb{R}} f(t)\pi(t)\mathrm{d}t$.

4.

a) Montrons par récurrence sur $n \in \mathbb{N}^*$ que $P^n(x, \frac{1}{n+m}) = \prod_{k=1}^{n-1} \left(1 - \frac{1}{(m+k)^2}\right)(1-x^2)$.

Initialisation : Pour $n = 1$. D'après la formule de la question 1, appliquée avec $A = \{\frac{1}{m+1}\}$, on a

$$P(x, \frac{1}{1+m}) = 1 - x^2$$

car $\int_{\{\frac{1}{m+1}\}\cap[0,1]} \mathrm{d}t = 0$ et $\delta_{\frac{1}{m+1}}(\{\frac{1}{m+1}\}) = 1$. D'où l'initialisation.

Hérédité : Soit $n \in \mathbb{N}^*$. Supposons la propriété vraie au rang n. Montrons qu'elle est vraie au rang $n+1$.

$$\begin{aligned}
P^{n+1}\left(x, \frac{1}{n+1+m}\right) &= \mathbb{P}(X_{n+1} = \frac{1}{n+m+1} \mid X_0 = x) \\
&= \mathbb{P}(X_{n+1} = \frac{1}{n+m+1} \cap X_n = \frac{1}{n+m} \mid X_0 = x) + \mathbb{P}(X_{n+1} = \frac{1}{n+m+1} \cap X_n \neq \frac{1}{n+m} \mid X_0 = x)
\end{aligned}$$

Calculons ces deux termes séparément :

$$\begin{aligned}
\mathbb{P}(X_{n+1} = \frac{1}{n+m+1} \cap X_n = \frac{1}{n+m} \mid X_0 = x) &= \mathbb{P}(X_{n+1} = \frac{1}{n+m+1} \mid X_n = \frac{1}{n+m} \cap X_0 = x)\mathbb{P}(X_n = \frac{1}{n+m} \mid X_0 = x) \\
&\quad \text{par définition de la probabilité conditionnelle} \\
&= \mathbb{P}(X_{n+1} = \frac{1}{n+m+1} \mid X_n = \frac{1}{n+m})P^n(x, \frac{1}{m+n}) \\
&\quad X_n \text{ étant une chaîne de Markov} \\
&= \left(1 - \frac{1}{(m+n)^2}\right) P^n(x, \frac{1}{m+n}) \\
&\quad \text{par définition de } X_n \\
&= \prod_{k=1}^{n} \left(1 - \frac{1}{(m+k)^2}\right)(1-x^2) \quad \text{par hyp. de rec.}
\end{aligned}$$

Pour le deuxième terme, on a :

$$\begin{aligned}\mathbb{P}(X_{n+1} = \frac{1}{n+m+1} \cap X_n \neq \frac{1}{n+m} \mid X_0 = x) &= \mathbb{P}(X_{n+1} = \frac{1}{n+m+1} \mid X_n \neq \frac{1}{n+m} \cap X_0 = x)\mathbb{P}(X_n \neq \frac{1}{n+m} \mid X_0 = x) \\ &= \mathbb{P}(X_{n+1} = \frac{1}{n+m+1} \mid X_n \neq \frac{1}{n+m})\mathbb{P}(X_n \neq \frac{1}{n+m} \mid X_0 = x)\end{aligned}$$

Or on a :

$$\begin{aligned}\mathbb{P}(X_{n+1} = \frac{1}{n+m+1} \mid X_n \neq \frac{1}{n+m}) = \mathbb{P}(X_{n+1} = \frac{1}{m+n+1} \mid X_n \in Q \setminus \{\frac{1}{m+n}\})\mathbb{P}(X_n \in Q \setminus \{\frac{1}{m+n}\}) + \\ \mathbb{P}(X_{n+1} = \frac{1}{m+n+1} \mid X_n \notin Q)\mathbb{P}(X_n \notin Q)\end{aligned}$$

Enfin, conditionnellement à $X_n \in Q \setminus \{\frac{1}{m+n}\}$, on a que $X_{n+1} \sim \mathcal{U}([0,1])$ et de même, conditionnellement à $X_n \notin Q$, on a $X_{n+1} \sim \mathcal{U}([0,1])$. Ainsi, on a

$$\mathbb{P}(X_{n+1} = \frac{1}{n+m+1} \cap X_n \neq \frac{1}{n+m} \mid X_0 = x) = 0$$

D'où la récurrence.

On a donc le résultat :

$$P^n\left(x, \frac{1}{n+m}\right) = \prod_{k=1}^{n-1}\left(1 - \frac{1}{(m+k)^2}\right)(1 - \frac{1}{m^2})$$

car $x = \frac{1}{m}$.

b) On a $A = \bigcup_{q\in\mathbb{N}}\{\frac{1}{m+q+1}\}$. Ainsi, A est un ensemble discret et dénombrable donc $\pi(A) = 0$ où π est la mesure uniforme de la question 2. Montrons que $\lim_{n\to+\infty} P^n(x, A)$ est non nulle.

On a, pour $n \in \mathbb{N}$,

$$\begin{aligned}P^n(x, A) &= \mathbb{P}(X_n \in A \mid X_0 = x) \\ &= \sum_{q\in\mathbb{N}} \mathbb{P}(X_n = \frac{1}{m+1+q} \mid X_0 = x) \quad \text{car l'union de A est disjointe} \\ &= \mathbb{P}(X_n = \frac{1}{m+1+(n-1)} \mid X_0 = x) + \underbrace{\sum_{q\in\mathbb{N}\setminus\{n-1\}} \mathbb{P}(X_n = \frac{1}{m+1+q} \mid X_0 = x)}_{\geq 0} \qquad (2) \\ &\geq \mathbb{P}(X_n = \frac{1}{m+1+(n-1)} \mid X_0 = x)\end{aligned}$$

Or, d'après la question précédente, on sait que

$$\mathbb{P}(X_n = \frac{1}{m+n} \mid X_0 = x) = P^n\left(x, \frac{1}{n+m}\right) = \prod_{k=1}^{n-1}\left(1 - \frac{1}{(m+k)^2}\right)(1 - \frac{1}{m^2})$$

On cherche donc à déterminer la limite de $\prod_{k=1}^{n-1} \left(1 - \frac{1}{(m+k)^2}\right)$ quand $n \to +\infty$. Or on a, en passant au log,

$$\begin{aligned}
\log\left(P^n\left(x, \frac{1}{n+m}\right)\right) &= \sum_{k=1}^{n-1} \log\left(1 - \frac{1}{(m+k)^2}\right) + \log(1 - \frac{1}{m^2}) \\
&= \sum_{k=1}^{n-1} \log\left(\frac{(m+k-1)(m+k+1)}{(m+k)^2}\right) + \log(1 - \frac{1}{m^2}) \\
&= \sum_{k=1}^{n-1} \log(m+k-1) + \sum_{k=1}^{n-1} \log(m+k+1) - 2\sum_{k=1}^{n-1} \log(m+k) + \log(1 - \frac{1}{m^2}) \\
&= \sum_{i=m}^{m+n-2} \log(i) + \sum_{j=m+2}^{m+n} \log(j) - 2\sum_{k=m+1}^{m+n-1} \log(k) + \log(1 - \frac{1}{m^2}) \\
&= \log(m) + \log(m+n) - \log(m+1) - \log(m+n-1) + \log(1 - \frac{1}{m^2}) \\
&= \log\left(\frac{m(m+n)}{(m+1)(m+n-1)}\right) + \log(1 - \frac{1}{m^2})
\end{aligned}$$

Ainsi, on a que

$$\lim_{n\to+\infty} \log\left(P^n\left(x, \frac{1}{n+m}\right)\right) = \log(\frac{m}{m+1}) + \log(1 - \frac{1}{m^2}) = \log\left(\frac{m}{m+1}(1 - \frac{1}{m^2})\right)$$

Ainsi,

$$\lim_{n\to+\infty} P^n\left(x, \frac{1}{n+m}\right) = \frac{m}{m+1}\left(1 - \frac{1}{m^2}\right)$$

Ainsi, en passant à la limite dans la dernière inégalité de 2 quand $n \to +\infty$, on a que

$$\lim_{n\to+\infty} P^n(x, A) \geq \frac{m}{m+1}\left(1 - \frac{1}{m^2}\right) > 0$$

Ainsi, on a

$$\lim_{n\to+\infty} P^n(x, A) \neq \pi(A)$$

Exercice 3

1.

On définit la fonction de risque empirique R_n par

$$R_n(w) = \mathbb{E}_{z\sim\mu_n}[J(w,z)] = \frac{1}{n}\sum_{i=1}^{n}(y_i - {}^t w x_i)^2$$

où μ_n est la distribution uniforme sur mes $\{z_i\}_{i\in[\![1,n]\!]}$ et $J(w,z) = (y - {}^t wx)^2$ avec $z = (x,y) \in \{z_i\}_{i\in[\![1,n]\!]}$. On cherche le vecteur normal de l'hyperplan séparateur, on peut donc chercher un vecteur unitaire et ainsi reformuler le problème en

$$\min_{w\in\mathbb{S}^{d-1}} R_n(w) \qquad (P)$$

Ainsi, $\mathcal{U}_c = \mathbb{S}^{d-1}$. Afin d'implémenter l'algorithme du gradient stochastique, on vérifie les hypothèses :

- (H1) La variable aléatoire $J(w,\cdot)\colon z \in \Omega \mapsto (y - {}^t wx)^2$ est mesurable et admet une espérance finie pour tout $w \in \mathcal{U}_c$.
- (H2) La fonction $J(\cdot, z)\colon w \in \mathcal{U}_c \mapsto (y - {}^t wx)^2$ est convexe, continue, à valeur dans $\mathbb{R}$ et différentiable de gradient

$$\forall w \in \mathcal{U}_c,\ \forall z \in \{z_i\}_{i\in[\![1,n]\!]},\ \nabla_w J(w,z) = -2(y - {}^t wx)x$$

— (H3) Pour tout $z \in \{z_i\}_{i\in[\![1,n]\!]}$ et pour tout $w \in \mathcal{U}_c = \mathbb{S}^{d-1}$, on a le gradient de J par rapport à w est bornée car il est continue, $\mathcal{U}_c$ est compact et $\{z_i\}_{i\in[\![1,n]\!]}$ est fini.

— (H4) Le problème admet une solution car R_n est continue sur le compact $\mathcal{U}_c$.

— (H5) On définit une suite $(\varepsilon_k)_{k\in\mathbb{N}}$ vérifiant

$$\sum_{k\geq 0} \varepsilon_k = +\infty \text{ et } \sum_{k\geq 0} \varepsilon_k^2 < +\infty$$

On prend par exemple la suite $(1/k)_{k\in\mathbb{N}*}$.

Alors, sous ses hypothèses, l'algorithme du gradient stochastique décrit à l'algorithme 2 converge en moyenne quadratique vers une solution au problème (P).

Algorithme 2 : Algorithme du gradient stochastique

Data : Soit $w_0 \in \mathcal{U}_c$, une suite $(\varepsilon_k)_{k\in\mathbb{N}}$ vérifiant (H5) et N le nombre d'itérations
for *k = 0 : N-1* **do**
    $z_{k+1} = (x_{k+1}, y_{k+1}) \sim \mu_n$;
    $w_{k+1} = w_k + 2\varepsilon_k(y_{k+1} - {}^t w_k x_{k+1})x_{k+1}$;
    $w_{k+1} = w_{k+1}/\|w_{k+1}\|$;
end
return w_N

3.

On test l'algorithme précédent sur un nuage de point composé de 10000 points séparés par un hyperplan aléatoire de normale w . En regardant l'erreur sur 1000 nuages de points différents, on voit que notre algorithme fait une erreur moyenne de 2%. Ainsi, le vecteur normal estimé w^* est assez proche de w. On peut donc dire que l'algorithme fonctionne.

4.

En ajoutant un bruit gaussien à nos données (on ajout une $\mathcal{N}(0, 0.3)$ à nos points après leur avoir attribués un label), on voit que l'algorithme performe légèrement moins bien avec une erreur moyenne de 13%. Cette erreur reste faible, mais largement supérieure au cas où l'on avait des données non bruitées.

5.

Pour le jeu de données *Breast Cancer Wisconsin (Diagnostic) Data Set*, on commence par centrer les données car l'algorithme du gradient stochastique tel qu'implémenté force l'hyperplan à passer par 0. On applique ensuite l'algorithme du gradient stochastique afin de trouver l'hyperplan de $\mathbb{R}^9$ séparateur. Après calcul, l'hyperplan trouvé commet 10.5% d'erreur.