Master 2 MVA - Computational Statistics

TP1 : Reminder on Markov Chains – Stochastic gradient descent

---

Mathis LE BAIL

October 2023

Exercise 1 : Box-Muller and Marsaglia-Bray algorithm

Let R a random variable with Rayleigh distribution with parameter 1 , whose probability density function f_R is given below, and Θ with uniform distribution on $[0, 2\pi]$. We also assume that R and Θ are independent. We have

$$\forall r \in \mathbb{R}, \quad f_R(r) = r \exp\left(-\frac{r^2}{2}\right) \mathbb{1}_{\mathbb{R}^+}(r).$$

1. Let X and Y such that

$$X = R\cos(\theta) \qquad \text{and} \qquad Y = R\sin(\theta) \tag{1}$$

Let h be a continuous and bounded function,

$$\begin{aligned}\mathbb{E}[h(X,Y)] &= \mathbb{E}[h(R\cos(\theta), R\sin(\theta))] \\ &= \int_{\mathbb{R}_+\times[0,2\pi)} h(r\cos(\theta), r\sin(\theta)) f_{R,\Theta}(r,\theta)\, dr\, d\theta \\ &= \int_{\mathbb{R}_+\times[0,2\pi)} h(r\cos(\theta), r\sin(\theta)) f_R(r) f_\Theta(\theta)\, dr\, d\theta \quad R \text{ and } \Theta \text{ are independents}\end{aligned} \tag{2}$$

We apply the following change of variables in (2). Let's consider :

$$\Phi : \left\{ \begin{array}{ccc} \mathbb{R}_+^* \times]0, 2\pi[& \to & \mathbb{R}^2 \backslash (\mathbb{R}_- \times \{0\}) \\ (r,\theta) & \mapsto & (x,y) = (r\cos(\theta), r\sin(\theta)) \end{array} \right. \tag{3}$$

Φ is a C^1-diffeomorphism. Indeed, Φ is a bijection with these restrictions on the spaces and for every $(r,\theta) \in \mathbb{R}_+^* \times]0, 2\pi[$, we have :

$$|\operatorname{Jac}\Phi(r,\theta)| = \left| \begin{array}{cc} \cos(\theta) & -r\sin(\theta) \\ \sin(\theta) & r\cos(\theta) \end{array} \right| = r \neq 0 \tag{4}$$

So if we come back to (2) using that $|\operatorname{Jac}\Phi^{-1}(r,\theta)| = |\operatorname{Jac}\Phi(r,\theta)|^{-1} = \frac{1}{r}$

$$\begin{aligned}\mathbb{E}[h(X,Y)] &= \int_{\mathbb{R}_+^*}\int_0^{2\pi} h(\Phi(r,\theta)) f_R(r) f_\Theta(\theta)\, dr\, d\theta \\ &= \int_{\mathbb{R}_+^*}\int_0^{2\pi} h(\Phi(r,\theta)) r \exp\left(-\frac{r^2}{2}\right)\frac{1}{2\pi}\, dr\, d\theta \\ &= \int_{\mathbb{R}^2\backslash(\mathbb{R}_-\times\{0\})} h(x,y)\frac{1}{2\pi}\exp(-\frac{x^2+y^2}{2})\, dx\, dy \qquad r \text{ and } |\operatorname{Jac}\Phi(r,\theta)|^{-1} \text{ cancel out} \\ &= \int_{\mathbb{R}^2\backslash(\mathbb{R}_-\times\{0\})} h(x,y)\frac{1}{\sqrt{2\pi}}\exp(-\frac{x^2}{2})\frac{1}{\sqrt{2\pi}}\exp(-\frac{y^2}{2})\, dx\, dy \\ &= \int_{\mathbb{R}^2} h(x,y)\, f_{\mathcal{N}(0,1)}(x)\, f_{\mathcal{N}(0,1)}(y)\, dx\, dy \qquad (\mathbb{R}_-\times\{0\}) \text{ est de mesure nulle}\end{aligned} \tag{5}$$

So by identification theorem, we have $f_{(X,Y)}(x,y) = f_X(x)f_Y(y) = f_{\mathcal{N}(0,1)}(x)\, f_{\mathcal{N}(0,1)}(y)$. X and Y have $\mathcal{N}(0,1)$ distribution and are independent.

2. To sample 2 independent Gaussian distributions $\mathcal{N}(0,1)$, we will use the writing of X and Y in 1. In order to do that, we need to sample according the Rayleigh distribution. For this, we will use the Inverse transform sampling.
We compute the cumulative distribution function of a random variable R of density $f_R(r) = r\exp\left(-\frac{r^2}{2}\right)\mathbb{1}_{\mathbb{R}^+}(r)$, $r \in \mathbb{R}$:

$$\begin{aligned} F_R(r) &= \int_{-\infty}^{r} f_R(s)\,ds \\ &= \int_{0}^{r} s\exp\left(-\frac{s^2}{2}\right)\,ds \\ &= 1-\exp\left(-\frac{r^2}{2}\right) \end{aligned} \tag{6}$$

We compute its inverse, for $u \in [0,1)$:

$$\begin{aligned} & F_R(r) = u \\ \Leftrightarrow & \exp\left(-\frac{r^2}{2}\right) = 1-u \\ \Leftrightarrow & r^2 = -2\log(1-u) \end{aligned} \tag{7}$$

Thus, for $u \in [0,1)$, $F_R^{-1}(u) = \sqrt{-2\log(1-u)}$. If $U \sim \mathcal{U}([0,1])$ then $1-U \sim \mathcal{U}([0,1])$. So to obtain a random variable R following a Rayleigh distribution, we can just take $R = \sqrt{-2\log(U)}$. By applying the Inverse transform sampling, we obtain the following algorithm :

Sampling 2 independent Gaussian distributions $\mathcal{N}(0,1)$

```
Sample $U_1$ with distribution $\mathcal{U}([0,1])$;
Sample $U_2$ with distribution $\mathcal{U}([0,2\pi])$;
Compute $R = \sqrt{-2\log(U_1)}$;
return $(X,Y) = (R\cos(U_2), Y = R\sin(U_2))$
```

3. a) If $U \sim \mathcal{U}([0,1])$ then $V = 2U-1 \sim \mathcal{U}([-1,1])$. After one iteration, (V_1,V_2) follows a uniform distribution on the square $[-1,1]^2$. Given the exit condition of the loop $V_1^2+V_2^2 > 1$, at the end (V_1,V_2) follows a uniform distribution on the disk with center (0,0) and radius 1 : $D = \{(v_1,v_2) | v_1^2+v_2^2 \leq 1\}$.

The probability density function associated to (V_1,V_2) is :

$$f_{(V_1,V_2)}(v_1,v_2) = \frac{1}{|D|}\mathbb{1}_{(v_1,v_2)\in[-1,1]^2}(v_1,v_2)\mathbb{1}_{v_1^2+v_2^2\leq 1}(v_1,v_2) = \frac{1}{\pi}\mathbb{1}_{v_1^2+v_2^2\leq 1}(v_1,v_2)$$

3. b) We set

$$T_1 = \frac{V_1}{\sqrt{V_1^2+V_2^2}}, \quad T_2 = \frac{V_2}{\sqrt{V_1^2+V_2^2}} \quad \text{and} \quad V = V_1^2+V_2^2.$$

We once again apply the identification theorem to identify the distributions of (V_1,V_2) and V. Let h be a continuous and bounded function,

$$\begin{aligned} \mathbb{E}[h((T_1,T_2),V)] &= \mathbb{E}[h(\frac{V_1}{\sqrt{V_1^2+V_2^2}}, \frac{V_2}{\sqrt{V_1^2+V_2^2}}, V_1^2+V_2^2)] \\ &= \int_{\mathbb{R}^2} h(\frac{v_1}{\sqrt{v_1^2+v_2^2}}, \frac{v_2}{\sqrt{v_1^2+v_2^2}}, v_1^2+v_2^2) f_{V_1,V_2}(v_1,v_2)\,dv_1\,dv_2 \\ &= \int_{D} h(\frac{v_1}{\sqrt{v_1^2+v_2^2}}, \frac{v_2}{\sqrt{v_1^2+v_2^2}}, v_1^2+v_2^2)\frac{1}{\pi}\,dr\,d\theta \qquad \text{a) : } (V_1,V_2)\sim\mathcal{U}(D) \end{aligned} \tag{8}$$

We apply the same C^1-diffeomorphism Φ as in 1. with a slight modification to the spaces :

$$\Phi : \left\{ \begin{array}{ccc} (0,1]\times(0,2\pi) & \to & D \\ (r,\theta) & \mapsto & (v_1,v_2) = (r\cos(\theta), r\sin(\theta)) \end{array} \right. \tag{9}$$

We have $(r, \theta) = \Phi^{-1}(v_1, v_2)$ on the defined spaces. So if we come back to (8) :

$$\begin{aligned}
\mathbb{E}[h((T_1, T_2), V)] &= \int_{(0,1]\times(0,2\pi)} h(\cos(\theta), \sin(\theta), r^2)\frac{1}{\pi} r\, dr\, d\theta && \text{with } |\operatorname{Jac}\Phi(r,\theta)| = r \\
&= \int_0^{2\pi}\int_0^1 h((\cos(\theta), \sin(\theta)), r^2)\frac{1}{2\pi} 2r\, dr\, d\theta && \text{with Fubini} \\
&= \int_0^{2\pi}\int_0^1 h((\cos(\theta), \sin(\theta)), v)\frac{1}{2\pi}\, dv\, d\theta && \text{with the bijective and monotone change of variable} \\
&= \int_{\mathbb{R}^2} h((\cos(\theta), \sin(\theta)), v)\frac{1}{2\pi}\mathbb{1}_{[0,2\pi]}(\theta)\,\mathbb{1}_{[0,1]}(v)\, dv\, d\theta \\
&= \int_{\mathbb{R}^2} h((\cos(\theta), \sin(\theta)), v)\, f_{\mathcal{U}([0,2\pi])}(\theta)\, f_{\mathcal{U}([0,1])}(v)\, dv\, d\theta
\end{aligned} \tag{10}$$

By identification theorem, we can deduce that (T_1, T_2) and V are independent, $V \sim \mathcal{U}([0,1])$ and (T_1, T_2) has the same distribution as $(\cos(\Theta), \sin(\Theta))$ with $\Theta \sim \mathcal{U}([0, 2\pi])$.

3. c) We have $(X, Y) = (ST_1, ST_2) = \left(\sqrt{-2\log(V)}T_1, \sqrt{-2\log(V)T_2}\right)$.

- With b), $V \sim \mathcal{U}([0,1])$ so with 2., we have $S = \sqrt{-2\log(V)} \sim$ Rayleigh(1).
- $(T_1, T_2) = (\cos(\Theta), \sin(\Theta))$ with $\Theta \sim \mathcal{U}([0, 2\pi])$
- (T_1, T_2) independent of V so S is independent of Θ

Thus according to 1., X and Y follow independent $\mathcal{N}(0,1)$ distributions.

3. d) $\mathbb{P}\left(V_1^2 + V_2^2 \leq 1 | V_1 \sim \mathcal{U}([-1,1]), V_2 \sim \mathcal{U}([-1,1])\right)$ is equal to the ratio of the area of the disk D to the area of the square $[-1,1]^2$, which is $\frac{\pi}{4}$. Thus, the expected number of steps in the "while" loop is the expectation of a geometric distribution with parameter $p = \frac{\pi}{4}$ (probability of being in the disk D).

For $X \sim Geo(\frac{\pi}{4})$, $\mathbb{E}[X] = \frac{4}{\pi}$. The expected number of steps is $\frac{4}{\pi}$.

Exercise 2 : Invariant distribution

We define a Markov chain $(X_n)_{n\geq 0}$ with values in $[0,1]$ as follows : given the current value $X_n (n \in \mathbb{N})$ of the chain,

- if $X_n = \frac{1}{m}$ (for some positive integer m), we let :

$$\begin{cases} X_{n+1} = \frac{1}{m+1} & \text{with probability } 1 - X_n^2 \\ X_{n+1} \sim \mathcal{U}([0,1]) & \text{with probability } X_n^2. \end{cases}$$

- if not, $X_{n+1} \sim \mathcal{U}([0,1])$.

1. We will determine the distribution of $X_{n+1}|X_n = x$ using identification theorem. Let h be a continuous and bounded function on $[0,1]$,

$$\mathbb{E}[h(X_{n+1})|X_n = x] = \int h(y)P(x,dy) \tag{11}$$

where $P(x,dy)$ is the probability density function of the random variable $X_{n+1}|X_n = x$.

We denote $Q = \{\frac{1}{m}|m \in \mathbb{N}^*\}$

If $x \notin Q$,

$$\begin{aligned} \mathbb{E}[h(X_{n+1})|X_n = x] &= \int h(y)P(x,dy) \\ &= \int_{\mathbb{R}} h(y)\mathbb{1}_{[0,1]}(y)dy \end{aligned} \tag{12}$$

Thus,

$$P(x,A) = \int_A P(x,dy) = \int_{A\cap[0,1]} dy \tag{13}$$

If $x \in Q$ i.e. there exists $m \in \mathbb{N}^*$ such that $x = \frac{1}{m}$,

We introduce the random variable B_m following a Bernoulli distribution of parameter $\frac{1}{m^2}$.

$$\begin{aligned} \mathbb{E}[h(X_{n+1})|X_n = x] &= \mathbb{E}[h(X_{n+1})|X_n = x, B_m = 0]P(B_m = 0) + \mathbb{E}[h(X_{n+1})|X_n = x, B_m = 1]P(B_m = 1) \\ &= h(\frac{1}{m+1})(1 - \frac{1}{m^2}) + \frac{1}{m^2}\int_{[0,1]} h(y)dy \qquad \text{with the law of total probability} \end{aligned} \tag{14}$$

Thus,

$$P(x,A) = (1-x^2)\delta_{\frac{1}{m+1}}(A) + x^2\int_{A\cap[0,1]} dy \qquad \text{if } x = \frac{1}{m} \tag{15}$$

2. We denote $\pi = \mathcal{U}([0,1])$, we want to show that π is invariant for P i.e. for every $A \subseteq [0,1]$, we have $\pi P(A) = \pi(A)$.

Let $A \subseteq [0,1]$,

$$\pi P(A) = \int_Q P(x,A)\pi(dx) + \int_{\bar{Q}} P(x,A)\pi(dx) \tag{16}$$

where $\bar{Q} = [0,1] \setminus Q$

Now, we have

$$\int_Q P(x,A)\pi(dx) = 0 \tag{17}$$

because Q is countable, so we integrate over a countable number of singletons with Lebesgue measure zero. Thus,

$$\begin{aligned}
\pi P(A) &= \int_{\bar{Q}} P(x,A)\pi(dx) \\
&= \int_{\bar{Q}} \int_{A\cap[0,1]} dy\, \pi(dx) \\
&= \int_{A\cap[0,1]} dy \int_{\bar{Q}} \pi(dx) \quad \text{with Fubini} \\
&= \pi(A) \times 1 = \pi(A)
\end{aligned} \tag{18}$$

3. Let $x \notin Q = \{\frac{1}{m} | m \in \mathbb{N}^*\}$ and f a bounded measurable function,

$$\begin{aligned}
Pf(x) &= \mathbb{E}[f(X_1)|X_0 = x] \\
&= \int P(x,dy)f(y) \\
&= \int_{[0,1]} f(y)dy \\
&= \int f(y)\pi(y)dy
\end{aligned} \tag{19}$$

For $n \geq 1$,

$$\begin{aligned}
P^n f(x) =& \mathbb{E}\left[f(X_n)|X_0 = x\right] \\
=& \mathbb{E}\left[f(X_n)|X_0 = x, X_1 = s \in \bar{Q}\right] P\left(X_1 = s \in \bar{Q}|X_0 = x\right) + \\
& \mathbb{E}\left[f(X_n)|X_0 = x, X_1 = \frac{1}{m}, m \in \mathbb{N}^*\right] P\left(X_1 = \frac{1}{m}, m \in \mathbb{N}^*|X_0 = x\right)
\end{aligned} \tag{20}$$

And

$$\begin{cases}
P\left(X_1 = s \in \bar{Q}|X_0 = x\right) = 1 \\
P\left(X_1 = \frac{1}{m}, m \in \mathbb{N}^*|X_0 = x\right) = 0
\end{cases} \tag{21}$$

because the law of $X_1|X_0 = x \sim \mathcal{U}([0,1])$ and Q is countable whereas the interval $[0,1]$ is uncountably infinite.

We obtain then,

$$\begin{aligned}
P^n f(x) =& \mathbb{E}\left[f(X_n)|X_0 = x, X_1 = s \in \bar{Q}\right] \\
=& \mathbb{E}\left[f(X_{n-1})|X_1 = s \in \bar{Q}\right] \qquad \text{with Markov property}
\end{aligned} \tag{22}$$

Thus, we can show by induction that,

$$P^n f(x) = \mathbb{E}[f(X_1)|X_0 = x] = \int f(y)\pi(y)dy$$

We conclude

$$\lim_{n\to+\infty} P^n f(x) = \int f(y)\pi(y)dy$$

4. a) Let $x = \frac{1}{m}$ with $m \geq 2$ and $n \in \mathbb{N}^*$,

We want to compute $P^n\left(\frac{1}{m}, \frac{1}{n+m}\right)$. With the same argument as in 3., at the moment X_n quits the countable space Q, it can't come back to Q with probability 1 as its law becomes π on the uncountable space $[0,1]$. So one can intuitively grasp that the only way to reach state $\frac{1}{n+m}$ at time n is to never leave Q and so the probability would be $P^n\left(\frac{1}{m}, \frac{1}{n+m}\right) = \prod_{l=0}^{n-1}\left(1 - \frac{1}{(m+l)^2}\right)$. One prove this by induction on $n \in \mathbb{N}^*$ for every $m \geq 2$.

For $n = 1$, $P(\frac{1}{m}, \frac{1}{1+m}) = (1 - \frac{1}{m^2})$ by definition of the transition kernel of $(X_n)_n$ for every integer $m \geq 2$.

Inductive Step : Let $n > 1$, let's assume that $P^n\left(\frac{1}{m}, \frac{1}{n+m}\right) = \prod_{l=0}^{n-1}\left(1 - \frac{1}{(m+l)^2}\right)$ stands until n for every integer $m \geq 2$, we want to prove it still holds for $n+1$. Let $m \geq 2$,

$$\begin{aligned}
P^{n+1}\left(\frac{1}{m}, \frac{1}{n+m+1}\right) &= P\left(P^n\left(\frac{1}{m}, \frac{1}{m+n+1}\right)\right) \\
&= \int P\left(\frac{1}{m}, dy\right) P^n\left(y, \frac{1}{m+n+1}\right) \\
&= \int P^n\left(y, \frac{1}{m+n+1}\right)\left[\frac{1}{m^2}\int_{dy\cap[0,1]} dt + \left(1 - \frac{1}{m^2}\right)\delta_{\frac{1}{m+1}}(dy)\right] \\
&= P^n\left(\frac{1}{m+1}, \frac{1}{m+n+1}\right)\left(1 - \frac{1}{m^2}\right) \qquad \text{because } \int_{dy\cap[0,1]} dt = 0
\end{aligned} \tag{23}$$

The assumption holds for every $m \geq 2$ so

$$\begin{aligned}
P^n\left(\frac{1}{m+1}, \frac{1}{m+n+1}\right)\left(1 - \frac{1}{m^2}\right) &= \left(1 - \frac{1}{m^2}\right)\prod_{l=0}^{n-1}\left(1 - \frac{1}{(m+1+l)^2}\right) \\
&= \left(1 - \frac{1}{m^2}\right)\prod_{l=1}^{n}\left(1 - \frac{1}{(m+l)^2}\right) \\
&= \prod_{l=0}^{n}\left(1 - \frac{1}{(m+l)^2}\right)
\end{aligned} \tag{24}$$

We have shown the property still holds for $n+1$.

This completes the proof by induction. $P^n\left(\frac{1}{m}, \frac{1}{n+m}\right) = \prod_{l=0}^{n-1}\left(1 - \frac{1}{(m+l)^2}\right)$ for every $n \in \mathbb{N}^*$ and integer $m \geq 2$.

4. b) Let $A = \bigcup_{q\in\mathbb{N}}\left\{\frac{1}{m+1+q}\right\}$ and again $x = \frac{1}{m}$ with $m \geq 2$,

We recall

$$\pi(A) = \int_{A\cap[0,1]} dy = 0 \tag{25}$$

because we integrate over a countable number of singletons $\left\{\frac{1}{m+1+q}\right\}_{q\in\mathbb{N}}$ of Lebesgue measure null.

And now,

$$\begin{aligned} P^n(x,A) &= \sum_{q\in\mathbb{N}} P^n\left(\frac{1}{m}, \frac{1}{m+1+q}\right) \\ &= P^n\left(\frac{1}{m}, \frac{1}{m+n}\right) + \sum_{q\in\mathbb{N}, q\neq n-1} P^n\left(\frac{1}{m}, \frac{1}{m+1+q}\right) \\ &= \prod_{l=0}^{n-1}\left(1-\frac{1}{(m+l)^2}\right) + \sum_{q\in\mathbb{N}, q\neq n-1} P^n\left(\frac{1}{m}, \frac{1}{m+1+q}\right) \end{aligned} \tag{26}$$

We study $\prod_{l=0}^{n-1}\left(1-\frac{1}{(m+l)^2}\right)$ when $n \to +\infty$:

$$\begin{aligned} \prod_{l=0}^{n-1}\left(1-\frac{1}{(m+l)^2}\right) &= \left(1-\frac{1}{m^2}\right)\left(1-\frac{1}{(m+1)^2}\right)\dots\left(1-\frac{1}{(m+n-1)^2}\right) \\ &= \frac{m^2-1}{m^2}\frac{(m+1)^2-1}{(m+1)^2}\dots\frac{(m+n+1)^2-1}{(m+n+1)^2} \end{aligned} \tag{27}$$

Using the remarkable identity $\frac{(m+k)^2-1}{(m+k)^2} = \frac{(m+k-1)(m+k+1)}{(m+k)^2}$ with integer $k \geq 0$, we obtain :

$$\begin{aligned} \prod_{l=0}^{n-1}\left(1-\frac{1}{(m+l)^2}\right) &= \frac{m^2-1}{m^2}\frac{(m+1)^2-1}{(m+1)^2}\dots\frac{(m+n+1)^2-1}{(m+n+1)^2} \\ &= \frac{(m-1)(m+1)}{m^2}\frac{m(m+2)}{(m+1)^2}\frac{(m+1)(m+3)}{(m+2)^2}\dots\frac{(m+n-2)(m+n)}{(m+n-1)^2} \\ &= \frac{m-1}{m}\frac{m+n}{m+n-1} \underset{n\to+\infty}{\sim} \frac{m-1}{m} > 0 \end{aligned} \tag{28}$$

Thus,

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

We have $\lim_{n\to+\infty} P^n(x,A) \geq \frac{m-1}{m} > 0 = \pi(A)$

Exercise 3 : Stochastic Gradient Learning in Neural Networks

1. The stochastic gradient descent is an iterative optimization algorithm that aims to find in the search space the parameter w that minimizes the empirical risk $R_n(w)$. To do that, it updates the model parameter w iteratively using the opposite of the gradient of the empirical risk with respect to w.

The empirical risk is $R_n(w) = \frac{1}{n}\sum_{i=1}^{n}(y_i - w^t x_i)^2$ so we have :

$$\nabla_w R_n(w) = \frac{1}{n}\sum_{i=1}^{n}\nabla_w (y_i - w^t x_i)^2 = \frac{1}{n}\sum_{i=1}^{n} -2(y_i - w^t x_i)x_i$$

As it is expensive to compute the gradient at each point $i \in \{1, .., n\}$ of the train set, the stochastic gradient algorithm evaluates at each update the gradient at only one point (x_i, y_i) sampled at random. Thus, $\nabla_w R_n(w)$ is approximated by $-2(y_i - w^t x_i)x_i$. Here are the steps of the algorithm :

1. Initialize the model parameter w with random value w^0 or some other initialization method.
2. While (step $k \leq n_{max}$) and $(\nabla_w R_p(w^k) > \epsilon)$:
   (a) We choose at random a point (x_i, y_i) in the train set and a learning rate step size $\epsilon_k > 0$
   (b) We update the parameter with the estimation of the gradient at point (x_i, y_i) :

$$w^{k+1} = w^k - \epsilon_k \nabla_w R_i(w^k) = w^k + 2\epsilon_k (y_i - {w^k}^t x_i)x_i$$

where $\nabla_w R_p(w^k)$ is the gradient this time estimated on a greater number of samples p. In order not to reduce the speed of the SGD, this gradient is not computed at each step but after a certain number of steps.

For the implementation we choose the sequence of steps defined in the course $(\epsilon_k)_k = (\frac{1}{k^\alpha})_k$ with $\alpha \in (\frac{1}{2}, 1)$

2. 3. In our toy example, we choose a normal vector $\bar{w} = (-1, 1)$. For the vector w^* estimated by our implementation of the SGD, we obtain the following results :

```
The vector w estimated :  [-1.93461048  1.98428947]
The ratio between the slopes of the two vectors : 1.026
```

One compared the slopes of the two straight lines playing the role of hyperplane, the ratio is close to 1 which indicates that the estimation is quite close of the exact value.

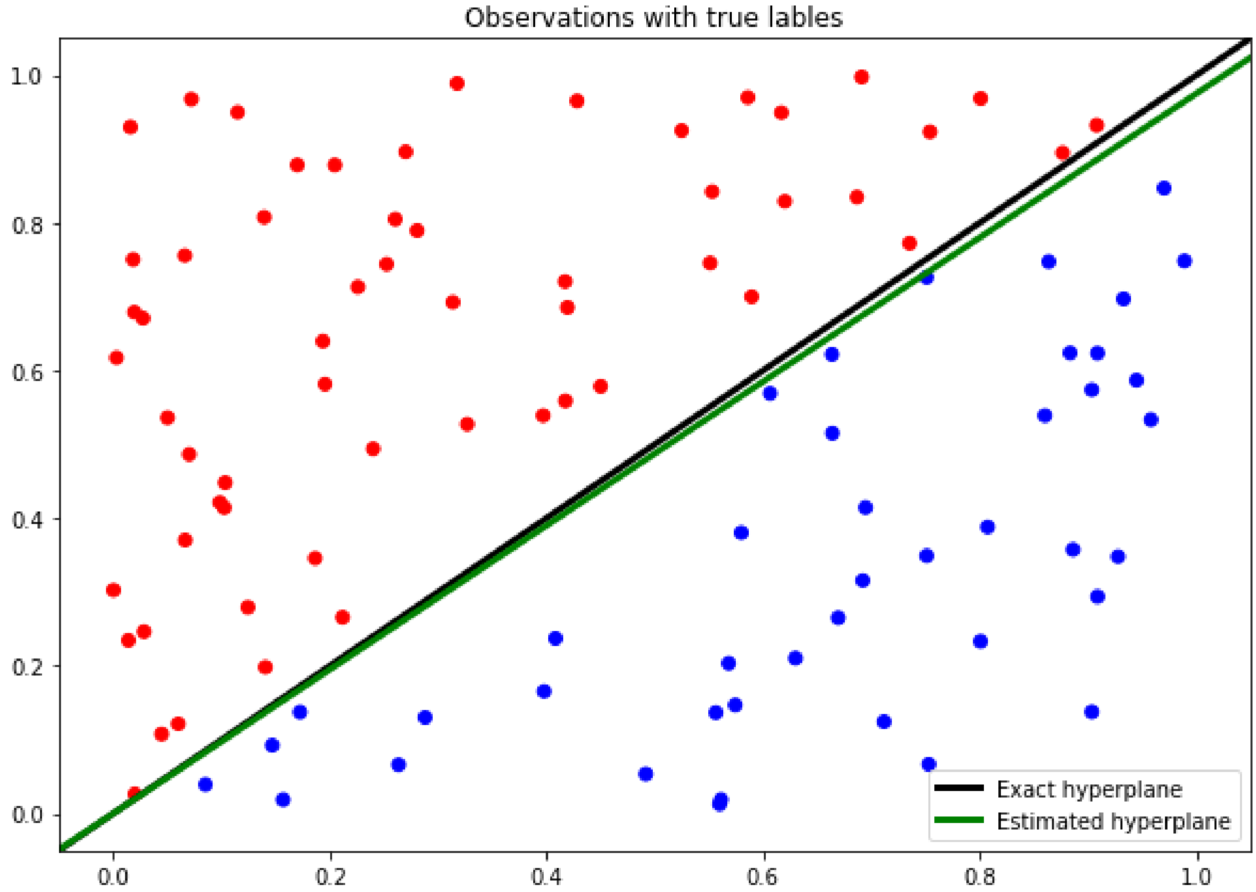

FIGURE 1 – Graphical representation of the set of fake observations defined on the square $[0,1]^2$ with their label. The exact hyperplane is represented in black whereas the estimated hyperplane is in green, one can see graphically that the two straight lines are quite closed.

4. We add an additive Gaussian noise of mean 0 and standard deviation 0.2 to our observations :

$$\{z_i\}_{i=1}^n = \{z_i\}_{i=1}^n + \mathcal{N}(0, 0.2^2)$$

We look again at our estimated vector w^* in this configuration. We obtain the following results :

```
The vector w estimated :  [-1.45839226  1.61091093]
The ratio between the slopes of the two vectors : 1.105
```

One see that the ratio between the slopes of the two vectors deviates further from 1 compared to the same metric in the configuration of 3. However, it doesn't seem to deteriorate too drastically, it can be explained by the fact that the Gaussian noise is not yet too significant in comparison to the coordinates of the observations (standard deviation equals to 0.2).

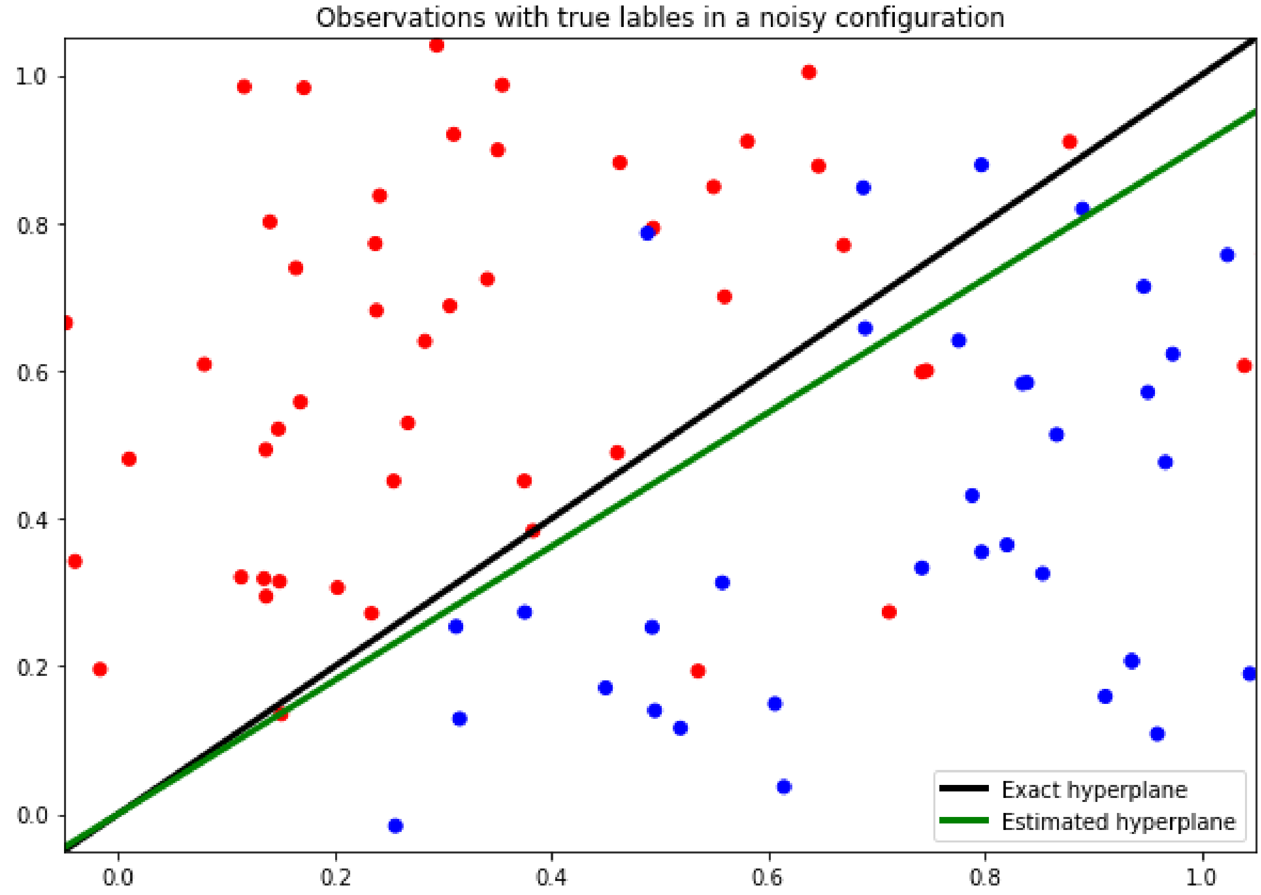

FIGURE 2 – Graphical representation of the set of fake observations defined on the square $[0,1]^2$ with their label in the case of a Gaussian additive noise of standard deviation 0.2. The exact hyperplane is represented in black whereas the estimated hyperplane is in green, one can see graphically that the two straight lines remain relatively closed.

As we can expect, as we increase the variance of the additive Gaussian noise, the estimated hyperplane diverges from the theoretical one. To visualize this, we represent in Figure (3), the evolution of the ratio between the slopes of the two lines as function of the standard deviation of the additive centered Gaussian noise.

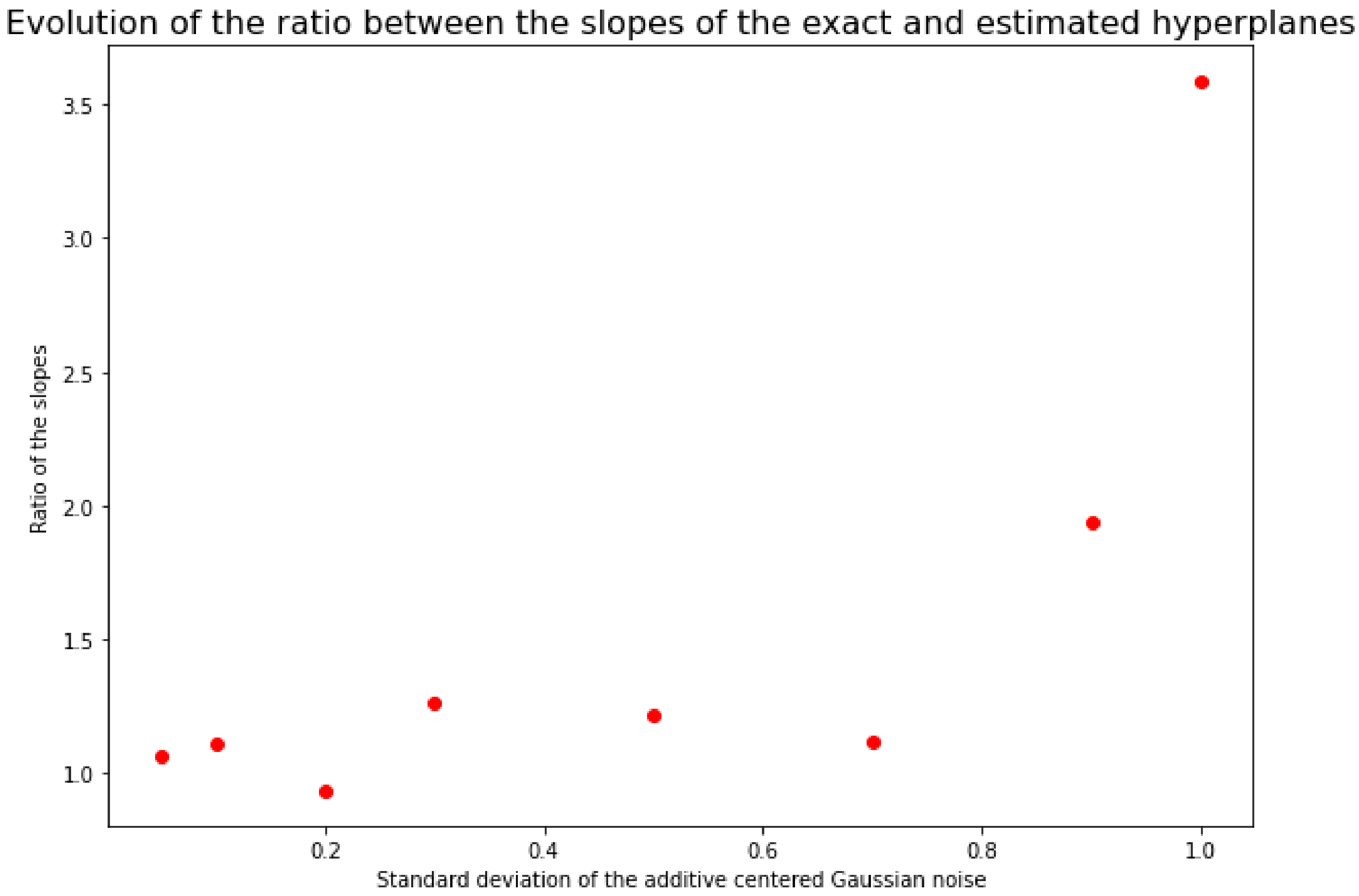

Figure 3 – Evolution of the ratio between the slopes of the two lines as function of the standard deviation of the additive centered Gaussian noise. The deterioration of the estimator remains bounded until we add a noise of standard deviation greater than 0.8. This is not surprising considering that our observations x_i are within the range $[0, 1]^2$

5. We tested our SGD algorithm on the *Breast Cancer Wisconsin (Diagnostic) Data Set*. It seems to work correctly, depending on the seed we fix we obtain an accuracy that oscillates between around 65% and 85% which is satisfactory for a simple model like a linear classifier.