Computational Statistics

Master MVA

TP n°1

Exercise 1: Box-Muller and Marsaglia-Bray algorithm

1. Let $g : \mathbb{R}^2 \to \mathbb{R}^2$ be a bounded, measurable function, and let $\Phi : \left| \begin{array}{lcl} \mathbb{R}_+^* \times (0, 2\pi) & \to & \mathbb{R}^2 \backslash \{(x,0) \mid x \geq 0\} \\ (r, \theta) & \mapsto & (r\cos\theta, r\sin\theta) \end{array} \right.$, which is a $\mathcal{C}^1$-diffeomorphism, with Jacobian determinant at (r,θ) equal to $r = \sqrt{x^2+y^2}$ if $(x,y) = \Phi(r,\theta)$. Then we have:

$$
\begin{aligned}
\mathbb{E}\left[g(X,Y)\right] &= \mathbb{E}\left[g(R\cos\Theta, R\sin\Theta)\right] \\
&= \frac{1}{2\pi}\int_{\mathbb{R}_+^* \times (0,2\pi)} g(r\cos\theta, r\sin\theta) f_R(r)\, dr d\theta \\
&= \frac{1}{2\pi}\int_{\mathbb{R}^2 \backslash \{(x,0) | x \geq 0\}} g(x,y) f_R(\sqrt{x^2+y^2}) \frac{1}{\sqrt{x^2+y^2}}\, dxdy \text{ (change of variable using } \Phi) \\
&= \frac{1}{2\pi}\int_{\mathbb{R}^2 \backslash \{(x,0) | x \geq 0\}} g(x,y) \exp\left(-\frac{x^2+y^2}{2}\right) dxdy \\
&= \frac{1}{2\pi}\int_{\mathbb{R}^2} g(x,y) \exp\left(-\frac{x^2+y^2}{2}\right) dxdy.
\end{aligned}
$$

Therefore (X,Y) has for density $(x,y) \mapsto \left(\frac{1}{\sqrt{2\pi}}\exp(-x^2/2)\right)\left(\frac{1}{\sqrt{2\pi}}\exp(-y^2/2)\right)$, product of the densities of two $\mathcal{N}(0,1)$, from which we deduce:

$$\boxed{X \text{ and } Y \text{ have distribution } \mathcal{N}(0,1) \text{ and are independent}}.$$

2. To sample two independent Gaussian distributions $\mathcal{N}(0,1)$, we sample R with Rayleigh distribution with parameter 1 and Θ with uniform distribution on $[0, 2\pi]$ independent, and use question 1.

For R, we use inverse transform sampling. The cumulative function of a Rayleigh random variable is given by $F_R(r) \overset{\text{def}}{=} \mathbb{P}(R \leq r) = \mathbb{1}_{r\geq 0}\int_0^r t\exp(-t^2/2)\, dt = \mathbb{1}_{r \geq 0}(1 - \exp(-r^2/2))$, and we then have, for $u \in [0,1]$, $\mathbb{P}(R \leq r) \geq u \iff \mathbb{1}_{r\geq 0}(1-\exp(-r^2/2)) \geq u \iff r \geq \sqrt{-2\ln(1-u)}$, so $F_R^{-1}(u) = \sqrt{-2\ln(1-u)}$.
So if U has uniform distribution on $[0,1]$, then $F_R^{-1}(U) = \sqrt{-2\ln(1-U)}$ follows Rayleigh distribution, and so does $\sqrt{-2\ln(U)}$ (because U and $1-U$ have same law).

From this, we deduce Algorithm 1.

Algorithm 1 Question 2

$U, V \leftarrow \mathcal{U}([0,1])$ independent
$R \leftarrow \sqrt{-2\ln U}$
$\Theta \leftarrow 2\pi V$
return $(R\cos\Theta, R\sin\Theta)$

3.

a) The loop corresponds to a rejection sampling. At the end of the loop, the law of (V_1, V_2) is the law of a random vector uniformly distributed on $[-1,1]^2$ conditionally to the fact that its L^2 norm is at most 1, so it is the uniform distribution of the closed unit disk $\overline{D}(0,1)$.
Let us re-prove the correctness of rejection sampling for this specific case.

Let $(V_1^{(n)})_{n\geq 1}, (V_2^{(n)})_{n\geq 1}$ be two independent sequences of i.i.d. random variables with distribution $\mathcal{U}([-1,1])$ and $T := \min\left\{n \geq 1 \mid (V_1^{(n)})^2 + (V_2^{(n)})^2 \leq 1\right\}$.
Then T follows a geometric distribution on $\mathbb{N}^*$ with probability of success given by

$$\mathbb{P}((V_1^{(1)})^2 + (V_2^{(1)})^2 \leq 1) = \mathbb{P}\left(\mathcal{U}([-1,1]^2) \in \overline{D}(0,1)\right) = \frac{|\overline{D}(0,1)|}{|[-1,1]^2|} = \frac{\pi}{4},$$

where $|\cdot|$ denotes Lebesgue measure.
For $f : [-1,1]^2 \to \mathbb{R}$ measurable and bounded, we have

$$\begin{aligned}
\mathbb{E}\left[f\left(V_1^{(T)}, V_2^{(T)}\right)\right] &= \mathbb{E}\left[\sum_{n=1}^{+\infty} f\left(V_1^{(T)}, V_2^{(T)}\right) \mathbb{1}_{\{T=n\}}\right] \quad \text{because } T < +\infty \text{ a.s.} \\
&= \mathbb{E}\left[\sum_{n=1}^{+\infty} f\left(V_1^{(n)}, V_2^{(n)}\right) \prod_{k=1}^{n-1} \mathbb{1}_{\left\{(V_1^{(k)})^2+(V_2^{(k)})^2>1\right\}} \mathbb{1}_{\left\{(V_1^{(n)})^2+(V_2^{(n)})^2\leq 1\right\}}\right] \\
&= \sum_{n=1}^{+\infty} \mathbb{E}\left[f\left(V_1^{(n)}, V_2^{(n)}\right) \mathbb{1}_{\left\{(V_1^{(n)})^2+(V_2^{(n)})^2\leq 1\right\}}\right] \prod_{k=1}^{n-1} \underbrace{\mathbb{E}\left[\mathbb{1}_{\left\{(V_1^{(k)})^2+(V_2^{(k)})^2>1\right\}}\right]}_{=1-\pi/4} \quad \text{(independence)} \\
&= \frac{1}{4}\left(\int_{[-1,1]^2} f(v_1, v_2) \mathbb{1}_{v_1^2+v_2^2\leq 1}\, dv_1 dv_2\right) \underbrace{\sum_{n=1}^{+\infty}\left(1-\frac{\pi}{4}\right)^{n-1}}_{=4/\pi} \quad \text{(using the law of } (V_1, V_2))\\
&= \frac{1}{\pi}\int_{\overline{D}(0,1)} f(v_1, v_2)\, dv_1 dv_2.
\end{aligned}$$

This shows that the distribution of $\left(V_1^{(T)}, V_2^{(T)}\right)$ is the uniform distribution on $\overline{D}(0,1)$, or:
$\boxed{\text{the distribution of } (V_1, V_2) \text{ at the end of the loop is the uniform distribution on } \overline{D}(0,1)}$, with density $(x,y) \mapsto \frac{1}{\pi}\mathbb{1}_{x^2+y^2\leq 1}$.

b) The number of steps in the "while" loop is given by T defined in the previous question.
We have seen that T follows a geometric distribution on $\mathbb{N}^*$ with probability of success $\frac{\pi}{4}$, so $\mathbb{E}[T] = \frac{4}{\pi}$.
Therefore $\boxed{\text{the expected number of steps in the "while" loop is } \frac{4}{\pi}}$.

c) Let us also define $T_2 = \dfrac{V_2}{\sqrt{V_1^2+V_2^2}}$.
Consider the $\mathcal{C}^1$-diffeomorphism $\Psi : \left|\begin{array}{lll} (0,1)\times(0,2\pi) & \to & D(0,1)\backslash\{(x,0) \mid x \in [0,1[\} \\ (r,\theta) & \mapsto & (\sqrt{r}\cos\theta, \sqrt{r}\sin\theta)\end{array}\right.$ with Jacobian determinant at (r,θ) equal to $\dfrac{1}{2}$.

For $f : [-1,1]^2 \times [0,1] \to \mathbb{R}$ measurable and bounded, using the law of (V_1, V_2) found in question a):

$$\begin{aligned}
\mathbb{E}\left[f\left((T_1,T_2),V\right)\right] &= \mathbb{E}\left[f\left(\left(\frac{V_1}{\sqrt{V_1^2+V_2^2}}, \frac{V_2}{\sqrt{V_1^2+V_2^2}}\right), V_1^2+V_2^2\right)\right] \\
&= \frac{1}{\pi}\int\int_{D(0,1)} f\left(\left(\frac{v_1}{\sqrt{v_1^2+v_2^2}}, \frac{v_2}{\sqrt{v_1^2+v_2^2}}\right), v_1^2+v_2^2\right) dv_1 dv_2 \\
&= \frac{1}{2\pi}\int\int_{(0,1)\times(0,2\pi)} f\left((\cos\theta, \sin\theta), r\right) dr d\theta \text{ (change of variable using } \Psi).
\end{aligned}$$

So if $U \sim \mathcal{U}([0,1])$ and $\Theta \sim \mathcal{U}([0,2\pi])$ are independent, we have shown that

$$\mathbb{E}\left[f((T_1,T_2),V)\right] = \mathbb{E}\left[f((\cos\Theta, \sin\Theta), U)\right]$$

i.e., (T_1,T_2) and V are independent, $V \sim \mathcal{U}([0,1])$, and (T_1,T_2) has the same distribution as $(\cos\Theta, \sin\Theta)$, so:

$$\boxed{T_1 \text{ and } V \text{ are independent, } V \sim \mathcal{U}([0,1]), \text{ and } T_1 \text{ has the same distribution as } \cos\Theta \text{ with } \Theta \sim \mathcal{U}([0,2\pi])}.$$

d) $S = \sqrt{-2\log\left(V_1^2+V_2^2\right)} = \sqrt{-2\log V}$ where $V \sim \mathcal{U}([0,1])$ using the previous question.

By the result established in question 2., we have that the distribution of S is Rayleigh with parameter 1.

And $(X,Y) = (ST_1, ST_2) \overset{\text{(distribution)}}{=} (S\cos\Theta, S\sin\Theta)$ with S and Θ independent, of laws given previously (using the distributions found in question c), and the fact that V and (T_1,T_2) are independent) so by question 1:

$$\boxed{X \text{ and } Y \text{ follow the distribution } \mathcal{N}(0,1) \text{ and are independent}}.$$

Exercise 2: Invariant distribution

1. $(X_n)_{n\geq 0}$ can also be defined as follows.
Let $(U_n)_{n\geq 1}$ be a sequence of i.i.d. random variables with uniform distribution on $[0,1]$, then define

$$X_{n+1} \overset{\text{(distribution)}}{=} \begin{cases} \dfrac{1}{k+1} & \text{if } X_n = \frac{1}{k} \text{ and } U_{n+1} \leq 1 - X_n^2 \\ \mathcal{U}([0,1]) & \text{if } X_n = \frac{1}{k} \text{ and } U_{n+1} > 1 - X_n^2 \\ \mathcal{U}([0,1]) & \text{otherwise.} \end{cases}$$

Let $x = \dfrac{1}{k}$. Then for A a borelian set:

$$\begin{aligned} P(x,A) &= \mathbb{P}(X_{n+1} \in A \mid X_n = x) \\ &= \mathbb{P}(X_{n+1} \in A, U_{n+1} \leq 1 - X_n^2 \mid X_n = x) + \mathbb{P}(X_{n+1} \in A, U_{n+1} > 1 - X_n^2 \mid X_n = x) \\ &= \mathbb{P}(U_{n+1} \leq 1 - X_n^2 \mid X_n = x)\mathbb{P}(X_{n+1} \in A \mid U_{n+1} \leq 1 - X_n^2, X_n = x) \\ &+ \mathbb{P}(U_{n+1} > 1 - X_n^2 \mid X_n = x)\mathbb{P}(X_{n+1} \in A \mid U_{n+1} > 1 - X_n^2, X_n = x) \\ &= (1-x^2)\delta_{\frac{1}{k+1}}(A) + x^2\mathbb{P}(\mathcal{U}([0,1] \in A) \\ &= (1-x^2)\delta_{\frac{1}{k+1}}(A) + x^2 \int_{A\cap[0,1]} dt. \end{aligned}$$

If $x \notin \left\{\dfrac{1}{k}, k \in \mathbb{N}^*\right\}$, then:

$$\begin{aligned} P(x,A) &= \mathbb{P}(X_{n+1} \in A \mid X_n = x) \\ &= \mathbb{P}\left(\mathcal{U}([0,1]) \in A\right) \\ &= \int_{A\cap[0,1]} dt. \end{aligned}$$

We have shown: $\boxed{P(x,A) = \begin{cases} (1-x^2)\delta_{\frac{1}{k+1}}(A) + x^2 \displaystyle\int_{A\cap[0,1]} dt & \text{if } x = \dfrac{1}{k} \\ \displaystyle\int_{A\cap[0,1]} dt & \text{otherwise.} \end{cases}}$

2. For any borelian set $A \subset [0,1]$, $\pi P(A) = \displaystyle\int \pi(dx)P(x,A) = \int P(x,A)\pi(x)\,dx$ (where we have $\pi(dx) = \pi(x)\,dx$, because π is used to denote both the measure and the density).

And using question 1., for π-almost every $x \in [0,1]$ (the set $\left\{\frac{1}{k} \mid k \geq 1\right\}$ has Lebesgue measure 0), we have $P(x,A) = \displaystyle\int_{A\cap[0,1]} dt = \pi(A)$, so $\pi P(A) = \displaystyle\int_0^1 \pi(A)\,dx = \pi(A)$.
We have proven $\pi P = \pi$, meaning that $\boxed{\pi \text{ is invariant for } P}$.

3. For $x \notin \left\{\dfrac{1}{k}, k \in \mathbb{N}^*\right\}$, $Pf(x) = \mathbb{E}\left[f(X_1) \mid X_0 = x\right] = \mathbb{E}\left[f(\mathcal{U}([0,1]))\right] = \displaystyle\int f(x)\pi(dx)$.
And for any n, $P^{n+1}f(x) = P(P^nf)(x) = \displaystyle\int P(x,dy)(P^nf)(y)$.
But the probability measure $P(x,\cdot)$ is actually π using question 1, so $P^{n+1}f(x) \overset{(*)}{=} \displaystyle\int (P^nf)(y)\,\pi(y)\,dy$.
By induction, we can then show that $\boxed{\forall n \in \mathbb{N}^*, \forall x \notin \left\{\dfrac{1}{k}, k \in \mathbb{N}^*\right\}, P^nf(x) = \displaystyle\int f(y)\pi(y)\,dy}$.

- It was shown above for $n = 1$.
- If it is true for some $n \geq 1$, then for π-almost every $y \in [0,1]$ $(P^n f)(y) = \int f(z)\pi(z)\,dz$, so with $(*)$, for all $x \notin \left\{\frac{1}{k}, k \in \mathbb{N}^*\right\}$, $P^{n+1}f(x) = \int f(z)\pi(z)\,dz$.

So $\forall n \geq 1, P^n f(x) = \int f(z)\pi(z)\,dz$, and $\boxed{\lim_{n\to+\infty} P^n f(x) = \int f(y)\pi(y)\,dy}$.

4. $x = \frac{1}{k}, k \geq 2$.

a)

$$P^{n+1}\left(\frac{1}{k}, \frac{1}{n+1+k}\right) = \int P\left(\frac{1}{k}, dy\right) P^n\left(y, \frac{1}{n+1+k}\right).$$

With question 1., we have $P\left(\frac{1}{k}, \cdot\right) = \frac{1}{k^2}\pi + \left(1 - \frac{1}{k^2}\right)\delta_{\frac{1}{k+1}}$, so

$$\begin{aligned}P^{n+1}\left(\frac{1}{k}, \frac{1}{n+1+k}\right) &= \frac{1}{k^2}\int \underbrace{P^n\left(y, \frac{1}{n+1+k}\right)}_{=0\ \pi\text{-a.e.}}\pi(y)\,dy + \left(1 - \frac{1}{k^2}\right)\int P^n\left(y, \frac{1}{n+1+k}\right)\delta_{\frac{1}{k+1}}(dy)\\ &= \left(1 - \frac{1}{k^2}\right)P^n\left(\frac{1}{k+1}, \frac{1}{n+1+k}\right).\end{aligned}$$

Therefore, by induction, we get

$$\begin{aligned}P^{n+1}\left(\frac{1}{k}, \frac{1}{n+1+k}\right) &= \prod_{i=0}^{n-1}\left(1 - \frac{1}{(k+i)^2}\right) \times P\left(\frac{1}{k+n}, \frac{1}{n+k+1}\right)\\ &= \prod_{i=0}^{n-1}\left(1 - \frac{1}{(k+i)^2}\right) \times \left(1 - \frac{1}{(k+n)^2}\right)\\ &= \prod_{i=0}^{n}\left(1 - \frac{1}{(k+i)^2}\right),\end{aligned}$$

so

$$\boxed{P^n\left(\frac{1}{k}, \frac{1}{n+k}\right) = \prod_{i=0}^{n-1}\left(1 - \frac{1}{(k+i)^2}\right)}.$$

b)

- On the first hand, since A is countable, we have:

$$\pi(A) = \pi\left(\bigcup_{q\in\mathbb{N}}\left\{\frac{1}{k+1+q}\right\}\right) = \sum_{q\in\mathbb{N}}\underbrace{\pi\left(\left\{\frac{1}{k+1+q}\right\}\right)}_{=0} = 0.$$

- On the other hand,

$$P^n(x, A) = \sum_{q\geq 0}\underbrace{P^n\left(\frac{1}{k}, \frac{1}{k+1+q}\right)}_{=0 \text{ if } q\neq n-1} \overset{(*)}{=} P^n\left(\frac{1}{k}, \frac{1}{k+n}\right) \overset{4)a)}{=} \prod_{i=0}^{n-1}\left(1 - \frac{1}{(k+i)^2}\right).$$

Here, we do not prove that $P^n\left(\frac{1}{k}, \frac{1}{k+1+q}\right)$ if $q \neq n-1$, because for what comes next, one can replace the equality $(*)$ with an inequality $\geq$ (inequality which is obvious), which is enough to conclude.

And

$$\prod_{i=0}^{n-1}\left(1-\frac{1}{(k+i)^2}\right)=\prod_{i=0}^{n-1}\left(1-\frac{1}{k+i}\right)\left(1+\frac{1}{k+i}\right)=\prod_{i=0}^{n-1}\left(\frac{k+i-1}{k+i}\frac{k+i+1}{k+i}\right)=\frac{k-1}{k}\frac{k+n}{k+n-1},$$

so $\lim_{n\to+\infty} P^n(x,A) \overset{(\geq)}{=} \frac{k-1}{k} > 0$ (because $k \geq 2$).

We therefore have $\boxed{\lim P^n(x,A) \neq \pi(A)}$.