TP1: Computational Statistics Course

Yassine ZAOUI
yassine.zaoui@ensta-paris.fr

Exercise 1: Box-Muller and Marsaglia-Bray Algorithm

(1) Let h be a bounded function. Since R and Θ are independent, one has:

$$\begin{aligned}\mathbb{E}[h(X,Y)] &= \mathbb{E}[h(R\cos(\Theta), R\sin(\Theta))] \\ &= \int_0^{2\pi}\int_{\mathbb{R}_+} h(r\cos(\theta), r\sin(\theta)) f_R(r) f_\Theta(\theta)\, d\theta\, dr \\ &= \int_0^{2\pi}\int_{\mathbb{R}_+} h(r\cos(\theta), r\sin(\theta)) \frac{re^{-r^2/2}}{2\pi}\, d\theta\, dr. \end{aligned} \tag{1}$$

Using the change of variables $x = r\cos(\theta)$ and $y = r\sin(\theta)$, the Jacobian of this transformation is the matrix:

$$\begin{pmatrix} \cos(\theta) & -r\sin(\theta) \\ \sin(\theta) & r\cos(\theta) \end{pmatrix},$$

and its determinant is r, i.e., $r\, dr\, d\theta = dx\, dy$. Therefore, we obtain:

$$\begin{aligned}\mathbb{E}[h(X,Y)] &= \int_{\mathbb{R}^2} h(x,y) \frac{e^{-(x^2+y^2)/2}}{2\pi}\, dx\, dy \\ &= \int_{\mathbb{R}^2} h(x,y) \frac{e^{-x^2/2}}{\sqrt{2\pi}} \frac{e^{-y^2/2}}{\sqrt{2\pi}}\, dx\, dy \\ &= \int_{\mathbb{R}^2} h(x,y) f_{N(0,1)}(x) f_{N(0,1)}(y)\, dx\, dy. \end{aligned} \tag{2}$$

Thus, we conclude that X and Y follow $N(0,1)$ distributions and are independent.

(2)The cumulative distribution function of the Rayleigh distribution is:

$$F(r) = P(R \le r) = \int_0^r ue^{-u^2/2}\, du = 1 - e^{-r^2/2}.$$

Trivially, we can derive the inverse:

$$F^{-1}(y) = \sqrt{-2\log(1-y)}, \quad \forall y \in [0,1).$$

We know that if $U \sim \text{Unif}([0,1])$ and T is a random variable with its distribution function F, where we know its inverse explicitly F^{-1}, then we can simulate the law of T by:

$$F^{-1}(U) \sim T.$$

As a result, we can simulate $R \sim \text{Rayleigh}(1)$ as follows:

- Simulate $U \sim \text{Unif}([0,1])$,
- Set $R = \sqrt{-2\log(1-U)}$ (so $R \sim \text{Rayleigh}(1)$).

Then, thanks to the result in part (1), we can simulate two independent standard Gaussian variables by simulating $\Theta \sim \text{Unif}([0, 2\pi])$ and taking:

$$X = R\cos(\Theta), \quad Y = R\sin(\Theta).$$

Thus, $X, Y \sim N(0,1)$ and X and Y are independent.

(3)(a) Let $R \in [0,1]$. We have:

$$P((V_1,V_2) \in D(O,R)) = P(V_1^2+V_2^2 \leq R^2) = P(V_1^2+V_2^2 \leq R^2, (2U_1-1)^2+(2U_2-1)^2 \leq 1)$$
$$+P(V_1^2+V_2^2 \leq R^2, (2U_1-1)^2+(2U_2-1)^2 > 1).$$

However, the second term is null since if $(2U_1-1)^2+(2U_2-1)^2 > 1$, then $V_1^2+V_2^2 > 1 \geq R^2$. Thus:

$$P((V_1,V_2) \in D(O,R)) = P(V_1^2+V_2^2 \leq R^2, (2U_1-1)^2+(2U_2-1)^2 \leq 1)$$
$$= P(V_1^2+V_2^2 \leq R^2 \,|\, (2U_1-1)^2+(2U_2-1)^2 \leq 1) P((2U_1-1)^2+(2U_2-1)^2 \leq 1).$$

Now:

$$P((2U_1-1)^2+(2U_2-1)^2 \leq 1) = \int_{(2u_1-1)^2+(2u_2-1)^2 \leq 1} du_1\, du_2$$
$$= \int_{(u_1-\frac{1}{2})^2+(u_2-\frac{1}{2})^2 \leq \frac{1}{4}} du_1\, du_2 = \text{Area}(D(I, \frac{1}{2})), \quad I\left(\frac{1}{2}, \frac{1}{2}\right),$$
$$= \frac{\pi}{4}.$$

Similarly, we have:

$$P(V_1^2+V_2^2 \leq R^2 \,|\, (2U_1-1)^2+(2U_2-1)^2 \leq 1) = \text{Area}(D(I, \frac{R}{2})) = \frac{\pi R^2}{4}.$$

Thus:

$$P((V_1,V_2) \in D(O,R)) = R^2.$$

Hence, $(V_1,V_2) \sim \text{Unif}(D(O,1))$.

(3)(b)As we saw in (3)(a), the probability to exit the loop is:

$$P((2U_1-1)^2+(2U_2-1)^2 \leq 1) = \frac{\pi}{4}.$$

Let N be the number of steps in the "while" loop. Clearly, $N \sim \text{Geom}\left(\frac{\pi}{4}\right)$. Thus, the expected number of steps in the "while" loop is:

$$\mathbb{E}[N] = \frac{4}{\pi}.$$

(3)(c) Let φ be a bounded operator. We have

$$\mathbb{E}[\varphi(T_1,V)] = \int_{D(O,1)} \varphi\left(\frac{v_1}{\sqrt{v_1^2+v_2^2}}, v_1^2+v_2^2\right) \frac{dv_1 dv_2}{\pi}.$$

We apply the expected change of variables $t_1 = \frac{v_1}{\sqrt{v_1^2+v_2^2}}$ and $v = v_1^2+v_2^2$.

Calculating the Jacobian of this transformation, we find:

$$J = \begin{pmatrix} \frac{v_2^2}{\sqrt{v_1^2+v_2^2}} & -\frac{v_1 v_2}{\sqrt{v_1^2+v_2^2}} \\ 2v_1 & 2v_2 \end{pmatrix},$$

and thus,

$$|\det(J)| = \frac{2|v_2|}{\sqrt{v}} = 2\sqrt{1-t_1^2},$$

since $v_2 = \pm\sqrt{v - t_1^2 v}$.

However, this transformation is not bijective over $D(O,1)$, but it is bijective over $D_1 = D(O,1) \cap \{v_2 > 0\}$ and $D_2 = D(O,1) \cap \{v_2 < 0\}$. Splitting our integral into two parts, we obtain

$$\mathbb{E}[\varphi(T_1, V)] = \int_{-1}^{1}\int_{0}^{1} \varphi(t_1, v)\frac{1}{2\pi\sqrt{1-t_1^2}}\, dv\, dt_1 + \int_{-1}^{1}\int_{0}^{1} \varphi(t_1, v)\frac{1}{2\pi\sqrt{1-t_1^2}}\, dv\, dt_1.$$

Thus,

$$\mathbb{E}[\varphi(T_1, V)] = \int_{-1}^{1}\int_{0}^{1} \varphi(t_1, v)\frac{1}{\pi\sqrt{1-t_1^2}}\, dv\, dt_1.$$

Therefore, we conclude that T_1 and V are independent random variables, with

$$q_{T_1}(t_1) = \frac{1}{\pi\sqrt{1-t_1^2}}\mathbb{1}_{[-1,1]}(t_1) \quad \text{and} \quad q_V(v) = \mathbb{1}_{[0,1]}(v),$$

so indeed $V \sim \text{Unif}([0,1])$.

Now, let ψ be a bounded operator. Then,

$$\mathbb{E}[\psi(T_1)] = \int_{-1}^{1} \psi(t_1)\frac{1}{\pi\sqrt{1-t_1^2}}\, dt_1 = \int_{-1}^{1} \psi(t_1)\frac{1}{2\pi\sqrt{1-t_1^2}}\, dt_1 + \int_{-1}^{1} \psi(t_1)\frac{1}{2\pi\sqrt{1-t_1^2}}\, dt_1.$$

For the first integral, we make the substitution $t_1 = \cos(\theta)$ with $\theta \in (0,\pi)$, so $dt_1 = -\sin(\theta)\, d\theta = -\sqrt{1-t_1^2}\, d\theta$. For the second integral, we use the same change of variables but with $\theta \in (\pi, 2\pi)$, yielding $dt_1 = -\sin(\theta)\, d\theta = \sqrt{1-t_1^2}\, d\theta$. Thus,

$$\mathbb{E}[\psi(T_1)] = \int_{0}^{\pi} \psi(\cos(\theta))\frac{1}{2\pi}\, d\theta + \int_{\pi}^{2\pi} \psi(\cos(\theta))\frac{1}{2\pi}\, d\theta = \int_{0}^{2\pi} \psi(\cos(\theta))\frac{1}{2\pi}\, d\theta.$$

Hence, we conclude that T_1 has the same distribution as $\cos(\Theta)$, where $\Theta \sim \text{Unif}([0,2\pi])$. (3)(d) Similarly to (3)(c), one can prove that if $T_2 = \frac{V_2}{\sqrt{V_1^2+V_2^2}}$, then T_2 has the same distribution as $\sin(\Theta)$, where $\Theta \sim \text{Unif}([0,2\pi])$ and T_2 is independent of V.

Let φ be a bounded operator. We have

$$\mathbb{E}[\varphi(X,Y)] = \mathbb{E}[\psi(\Theta, V)] = \int_{0}^{2\pi}\int_{0}^{1} \varphi\left(\cos(\theta)\sqrt{-2\log(v)}, \sin(\theta)\sqrt{-2\log(v)}\right)\frac{dv\, d\theta}{2\pi}.$$

Let us consider the change of variables $x = \cos(\theta)\sqrt{-2\log(v)}$ and $y = \sin(\theta)\sqrt{-2\log(v)}$. The Jacobian of this bijective transformation is given by

$$J = \begin{pmatrix} -y & -\frac{\cos(\theta)}{v\sqrt{-2\log(v)}} \\ x & -\frac{\sin(\theta)}{v\sqrt{-2\log(v)}} \end{pmatrix},$$

and

$$|\det(J)| = \frac{1}{v} = e^{\frac{1}{2}(x^2+y^2)}.$$

Therefore,

$$\mathbb{E}[\varphi(X,Y)] = \int_{\mathbb{R}^2} \varphi(x,y)e^{-\frac{1}{2}(x^2+y^2)}\, dx\, dy.$$

Thus, we conclude that $(X,Y) \sim \mathcal{N}(0, I_2)$.

Exercise 2: Invariant Distribution

(1)

We are given the following distributions for any $n \in \mathbb{N}$, where $k \in \mathbb{N}^*$ is fixed:

$$q_{X_{n+1}|X_n=\frac{1}{k}}(x) = \left(1 - \frac{1}{k^2}\right)\delta\left(x - \frac{1}{k+1}\right) + \frac{1}{k^2}\mathbb{1}_{[0,1]}(x),$$

$$q_{X_{n+1}|X_n\neq\frac{1}{k}}(x) = \mathbb{1}_{[0,1]}(x).$$

Thus, for any $A \subset \mathbb{R}$, we have:

$$P\left(\frac{1}{k}, A\right) = \int_A q_{X_{n+1}|X_n=\frac{1}{k}}(x)\,dx = \left(1 - \frac{1}{k^2}\right)\mathbb{1}_A\left(\frac{1}{k+1}\right) + \frac{1}{k^2}\int_{A\cap[0,1]} dx.$$

Now, if $x \neq \frac{1}{k}$, then:

$$P(x, A) = \int_A q_{X_{n+1}|X_n\neq\frac{1}{k}}(x)\,dx = \int_{A\cap[0,1]} dx.$$

(2)

By analogy with how we calculate the transition matrix applied to a probability measure in a discrete state space (i.e., $(\Pi M)(x) = \sum_y \Pi(y)M_{yx}$), we use the formula:

$$(\Pi P)(A) = \int_{\mathbb{R}} \Pi(dy)P(y, A) = \int_{\mathbb{R}} \Pi(y)P(y, A)\,dy = \int_{[0,1]} P(y, A)\,dy = \int_{[0,1]\setminus\{1/k\}} P(y, A)\,dy.$$

This simplifies to:

$$= \int_{[0,1]\setminus S}\int_{A\cap[0,1]} du\,dy = \int_{[0,1]}\int_{A\cap[0,1]} du\,dy = \int_{A\cap[0,1]} du = \Pi(A),$$

where $\Pi \sim \text{Unif}([0,1])$ and $S = \left\{\frac{1}{k} \,\middle|\, k \in \mathbb{N}^*\right\}$ (note that $\text{mes}(S) = 0$). Thus, indeed, Π is invariant for P.

(3)

Let $x \notin S$. Then

$$Pf(x) = \mathbb{E}[f(X_1)|X_0 = x] = \int_{\mathbb{R}} f(x_1)q_{X_1|X_0=x}(x_1)\,dx_1 = \int_{\mathbb{R}} f(x_1)\mathbb{1}_{[0,1]}(x_1)\,dx_1.$$

Thus,

$$= \int_{[0,1]} f(x_1)\,dx_1 = \mathbb{E}_\Pi[f(X)],$$

which is constant and independent of x. Therefore,

$$\forall n \in \mathbb{N}^*, \quad P^n f(x) = \mathbb{E}_\Pi[f(X)] = \int_{[0,1]} f(x_1)\,dx_1.$$

(4)(a)

Let $n \in \mathbb{N}^*$. We have

$$P^n\left(x, \frac{1}{k+n}\right) = \int_{\mathbb{R}\times\{\frac{1}{k+n}\}} q^{(n-1)}(x,y)q(y,z)\,dy\,dz$$

$$= \int_{\{\frac{1}{k+n-1}\}\times\{\frac{1}{k+n}\}} q^{(n-1)}(x,y)q(y,z)\,dy\,dz + \int_{\mathbb{R}\setminus\{\frac{1}{k+n-1}\}\times\{\frac{1}{k+n}\}} q^{(n-1)}(x,y)q(y,z)\,dy\,dz,$$

where $q^{(n)}(x,y)$ is the n-step transition density given by

$$q^{(n)}(x,y) = \int\cdots\int_{\mathbb{R}^{n-1}} q(x,x_1)\cdots q(x_{n-1},y)\,dx_1\ldots dx_{n-1},$$

and $q(x,y) = q_{X_1|X_0=x}(y)$.

For each term, the integral with respect to z is computed over a negligible set, so if the density $q(y,z)$ does not have any Dirac measure $\delta\left(z-\frac{1}{k+n}\right)$ in it, the integral is zero. Since we know there exists a path of non-zero probability from $X_0 = x = \frac{1}{k}$ to $X_n = \frac{1}{k+n}$ (specifically, the path $X_0 = \frac{1}{k}, X_1 = \frac{1}{k+1}, \ldots, X_{n-1} = \frac{1}{k+n-1}, X_n = \frac{1}{k+n}$), it follows that there must be a $\delta\left(z-\frac{1}{k+n}\right)$ term. According to (1), the only case where this occurs is when $y = \frac{1}{k+n-1}$; therefore, the first term is non-zero, while the second term is zero. Thus, necessarily, we use the expression of the transition density with a Dirac measure and obtain:

$$\int_{\{\frac{1}{k+n}\}} q(y,z)\,dz = 1 - \frac{1}{(k+n-1)^2}.$$

As a result, we can write the following recursion:

$$P^n\left(x, \frac{1}{k+n}\right) = \left(1 - \frac{1}{(k+n-1)^2}\right)\int_{\{\frac{1}{k+n-1}\}} q^{(n-1)}(x,y)\,dy$$

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

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

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

$$= \prod_{i=0}^{n-1} \frac{k+i-1}{k+i} \prod_{i=0}^{n-1} \frac{k+i+1}{k+i}$$

$$= \prod_{i=0}^{n-1} \frac{k+i-1}{k+i} \prod_{i=1}^{n} \frac{k+i}{k+i-1}$$

$$= \frac{k-1}{k}\frac{k+n}{k+n+1}.$$

Thus,

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

(4)(b)

According to (4)(a), and since $\frac{1}{n+k} \in A$, we have

$$1 \geq \lim_{n\to+\infty} P^n(x, A) \geq \lim_{n\to+\infty} P^n\left(x, \frac{1}{k+n}\right) = 1.$$

So, $\lim_{n\to+\infty} P^n(x, A) = 1$.

However, since A is countable, it is a set of null measure, and thus $\Pi(A) = 0$.

We conclude that $\lim_{n\to+\infty} P^n(x, A) \neq \Pi(A)$.

Exercise 3: Stochastic Gradient Learning in Neural Networks

(1)

In this framework, we define the objective function $R(w)$ as

$$R(w) = \mathbb{E}_z[J(w, z)]$$

where $J(w, z) = y - w^\top x$. We aim to minimize R over $\mathbb{R}^d$, which is a Hilbert space and can also be considered as a closed, convex, non-empty subset of itself.

Assumptions

- **($\mathbf{H}_0$)**: $J(\cdot, z)$ is convex and differentiable on $\mathbb{R}^d$. This implies that $R = \mathbb{E}_z[J(\cdot, z)]$ is also convex and differentiable on $\mathbb{R}^d$. Thus, for all $w \in \mathbb{R}^d$, we have

$$\nabla R(w) = \int -2(y - w^\top x)x \, dP(z).$$

- **($\mathbf{H}_1$)**: For $g(w, z) = -2(y - w^\top x)x$, we have

$$\nabla R(w) = \mathbb{E}_z[g(w, z)].$$

Gradient Stochastic Descent Algorithm

The algorithm for the Gradient Stochastic Descent (GSD) method to minimize empirical risk $R(w)$ is as follows:

Input: Choose an initial point $w^0 \in \mathbb{R}^d$ and let (ϵ_k) be a sequence of steps decreasing to 0.

Repeat:

1. Simulate $z^{k+1} \sim \mathscr{P}(z)$.
2. Update the weight as

$$w^{k+1} = \text{proj}_{\mathbb{R}^d}\left(w^k - \epsilon_{k+1} g(w^k, z^{k+1})\right) = w^k - \epsilon_{k+1} g(w^k, z^{k+1}).$$

3. Stop if

$$\left|\frac{R_n(w^{k+1}) - R_n(w^k)}{R_n(w^k)}\right| < \epsilon.$$

(2)

See the provided python script.

(3)+(4)

The following figures show the results of our algorithm applied to a normal standard distribution (1) and a uniform distribution over $[-1, 1]$ (2). In each plot, we display the true classification hyperplane along with the estimated hyperplane, both without noisy data and with added Gaussian noise.

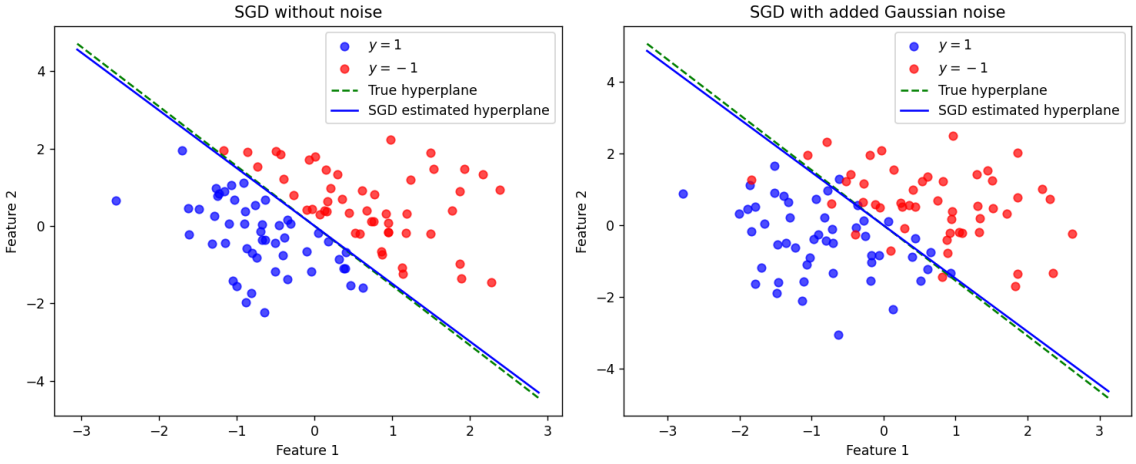

Figure 1: Data generated with $\mathcal{N}(0, 1)$

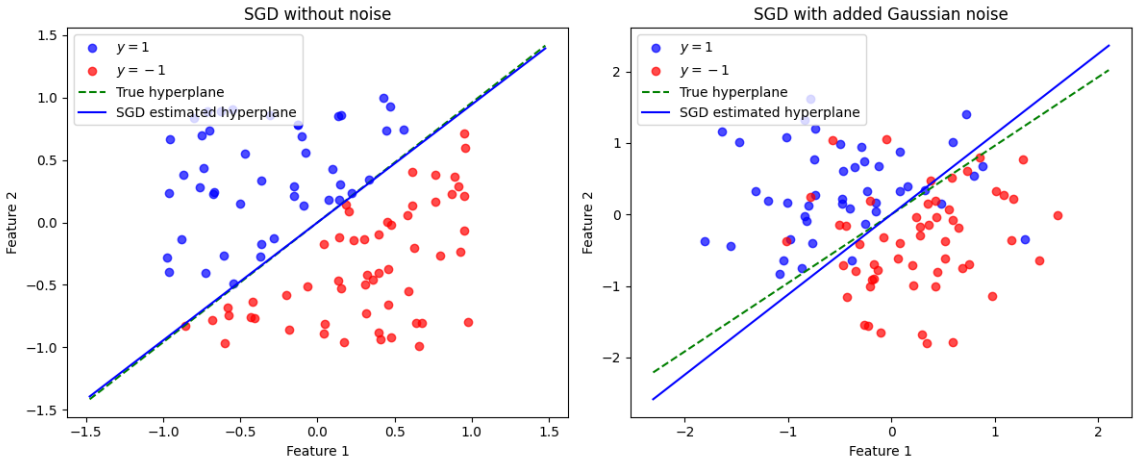

Figure 2: Data generated with $\mathcal{U}[-1, 1]$

Summary Table of Results

Used Distribution	$\mathbf{W}_{\text{bar}}$	$\mathbf{W}_{\text{estimated}}$	Risk	$\mathbf{W}_{\text{estimated, noisy}}$	$Risk_{noisy}$
$\mathcal{N}(0, 1)$	$[-0.36, -0.24]$	$[-0.66, -0.44]$	0.34	$[-0.54, -0.36]$	0.39
$\mathcal{U}[-1, 1]$	$[-0.37, 0.39]$	$[-0.97, 1.03]$	0.31	$[-0.63, 0.56]$	0.56

Table 1: Summary of the estimated weights and risks for clean and noisy data

Interpretation of Results

From the figures, we observe that the SGD algorithm performs better with clean data than with noisy data. This is in particular the case of the data generated with a uniform distribution. However, the difference is almost inexisting for the case with data generated with a normal distribution.

Maybe, the difference in performance is less significant for the normally distributed data, because the added Gaussian noise is uncorrelated with the initial Gaussian distribution, so it does not significantly affect the overall distribution. This is not the case for the uniformly distributed data, where the Gaussian noise has a more pronounced impact.
Moreover, although the risk in clean data cases is consistently lower than in noisy cases, we observe that $W_{\text{estimated, noisy}}$ is, in some cases, closer to W_{bar} than $W_{\text{estimated}}$ without noise.

(5)

For this last question, I tried to apply my algorithm to the data loaded thank to the link provided. However, I kept having a Runtime error and the computed risk is always nan. After some debugging, the issue turned out to be with the update of weights. In fact, at some points, all the weights become nan and introduce the Runtime error.
I opted for the first questions for a decreasing sequence as my learning rate in SGD algorithm. Nevertheless, when I made the learning rate constant namely $\epsilon = 0.001$, I obtained wonderful results and no Runtime errors occured. In fact, we obtain that:

- Training Accuracy: 0.97
- Testing Accuracy: 0.95
- Precision: 0.90
- Recall: 0.98