

# Computational Statistics - TD1

Louis Martinez

[louis.martinez@telecom-paris.fr](mailto:louis.martinez@telecom-paris.fr)

## Exercise 1: Box-Muller and Marsaglia-Bray algorithm

1)

Let  $h : \mathbb{R}^2 \rightarrow \mathbb{R}$  a measurable function:

$$\mathbb{E}(h(R \cos(\Theta), R \sin(\Theta))) = \int_{\mathbb{R}^+ \times [0, 2\pi]} h(r \cos(\theta), r \sin(\theta)) \frac{1}{2\pi} r \exp\left(-\frac{r^2}{2}\right) dr d\theta$$

We perform the change of variable:  $x = r \cos(\theta), y = r \sin(\theta)$

$$\det(J) = \begin{vmatrix} \cos(\theta) & -r \sin(\theta) \\ \sin(\theta) & r \cos(\theta) \end{vmatrix} = r$$

Hence:

$$\begin{aligned} \mathbb{E}(h(X, Y)) &= \frac{1}{2\pi} \int_{\mathbb{R} \times \mathbb{R}} h(x, y) \exp\left(-\frac{x^2 + y^2}{2}\right) dx dy \\ &= \int_{\mathbb{R} \times \mathbb{R}} h(x, y) \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{y^2}{2}\right) dx dy \end{aligned}$$

So

$$\begin{aligned} f_{XY}(x, y) &= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{y^2}{2}\right) \\ &= f_{X(x)} f_{Y(y)} \end{aligned}$$

Hence,  $X$  and  $Y$  have  $\mathcal{N}(0, 1)$  distributions and are independent.

2)

Given  $R$  with Rayleigh distribution of parameter 1 and  $\Theta \sim \mathcal{U}[0, 2\pi]$ , two independent variables, we're able to sample two random variables  $X$  and  $Y$  following a  $\mathcal{N}(0, 1)$  distribution.

Let  $U \sim \mathcal{U}[0, 1]$  and  $F_R$  the cdf of  $R$ .

- We know that  $F_R^{-1}(U)$  has the same distribution as  $R$ . Indeed, let  $T : [0, 1] \mapsto \mathbb{R}$  a strictly monotone function such that  $T(U)$  has the same distribution as  $R$ :

$$\begin{aligned} \forall r \in \mathbb{R}^+, F_R(r) &= \mathbb{P}(R \leq r) && \text{Definition of a s cdf} \\ &= \mathbb{P}(T(U) \leq r) && \text{Definition of } T \\ &= \mathbb{P}(U \leq T^{-1}(r)) && T \text{ is strictly increasing, hence } T^{-1} \text{ as well} \\ &= T^{-1}(r) && \text{Because } U \text{ is uniform on } [0, 1] \end{aligned}$$

And  $F_R(r) = T^{-1}(r) \Leftrightarrow T(r) = F_R^{-1}(r)$ .

- We compute  $F_R^{-1}$ :

$$\begin{aligned}
F_R(r) &= \int_0^r r \exp\left(-\frac{r^2}{2}\right) dr \\
&= \left[-\exp\left(-\frac{r^2}{2}\right)\right]_0^r \\
&= 1 - \exp\left(-\frac{r^2}{2}\right)
\end{aligned}$$

So  $F_R^{-1}(u) = \sqrt{-2 \ln(1 - u)}$  ( $u \in ]0, 1[$ )

We finally deduce the algorithm:

```

1 sample  $\Theta \sim \mathcal{U}[0, 2\pi]$ 
2 sample  $U \sim \mathcal{U}[0, 1]$ 
3  $R \leftarrow \sqrt{-2 \ln(1 - U)}$ 
4  $X \leftarrow R \cos(\Theta)$ 
5  $Y \leftarrow R \sin(\Theta)$ 
6 return  $X, Y$ 

```

3)

a)

$V_1$  and  $V_2$  both have a uniform distribution on  $[-1, 1]$  so **without** the while loop,  $(V_1, V_2) \sim \mathcal{U}[-1, 1]^2$ . The while loop ensures that  $V_1^2 + V_2^2 \leq 1$ , i.e that  $(V_1, V_2)$  lies in the unit disk. As any part of the disk is as likely to be sampled, we conclude that

after the while loop  $(V_1, V_2)$  has a uniform distribution on the unit disk.

b)

The while loop simulates independent Bernoulli trials of getting  $(V_1, V_2)$  such that  $V_1^2 + V_2^2 \leq 1$ . For one trial, the probability  $p$  of sampling inside the unit disk is the ratio between the area of the square  $[-1, 1]^2$  and the area of the disk, hence  $p = \frac{\pi}{4}$ .

Let  $T$  the random variable counting the number of trials needed to get one success. We denote by  $V_i^{(k)}$ ,  $i \in \{1, 2\}$ ,  $k \in \mathbb{N}^*$  the  $k$ -th sampling of  $V_i$ :

$$\begin{aligned}
\forall n \in \mathbb{N}^*, \mathbb{P}(T = n) &= \mathbb{P}\left(V_1^{(n)2} + V_2^{(n)2} > 1\right) \prod_{k=1}^{n-1} \mathbb{P}\left(V_1^{(k)2} + V_2^{(k)2} \leq 1\right) \\
&= p(1 - p)^{n-1}
\end{aligned}$$

So  $T \sim \mathcal{G}(p)$ , and the expected number of trials is  $\mathbb{E}(T) = \frac{4}{\pi}$ .

The expected number of steps is  $\frac{4}{4-\pi}$

c)

The joint pdf of  $(V_1, V_2)$  after the while loop is given by:

$$f_{V_1, V_2}(v_1, v_2) = \frac{1}{\pi} \mathbb{1}_{\{v_1^2 + v_2^2 \leq 1\}}$$

We switch to polar coordinates for convenience:  $V_1 = R \cos(\Theta)$ ,  $V_2 = R \sin(\Theta)$ ,  $R \in [0, 1]$ ,  $\Theta \in [0, 2\pi[$ . So we can redefine  $T_1 = \frac{V_1}{R} = \cos(\Theta)$  and  $V = R^2$ .

1. We first show that  $V \sim \mathcal{U}[0, 1]$ :

$$F_V(v) = \mathbb{P}(V \leq v) = \mathbb{P}(R^2 \leq v) = \mathbb{P}(R \leq \sqrt{v}) = \int_0^{\sqrt{v}} \frac{1}{\pi} 2\pi r dr = (\sqrt{v})^2 = v$$

thus:  $f_V(v) = F_V'(v) = 1, \forall v \in [0, 1]$

Hence  $V$  has a uniform distribution on  $[0, 1]$

2. Since  $\Theta \sim \mathcal{U}[0, 2\pi[$ , its pdf is  $f_\Theta = \frac{1}{2\pi} \mathbb{1}_{[0, 2\pi[}$ , continuous on  $[0, 2\pi[$ . As  $T_1 = \cos \Theta$  (hence  $\Theta = \arccos(T_1)$ ), we deduce that:

$$\begin{aligned} \forall t \in [-1, 1], f_{T_1}(t) &= f_\Theta(\theta) \left| \frac{d\theta}{dt} \right| \\ &= \frac{1}{2\pi} \left| -\frac{1}{\sqrt{1-t^2}} \right| \quad \text{Derivative of } \arccos \\ &= \frac{1}{2\pi} \frac{1}{\sqrt{1-t^2}} \end{aligned}$$

So with  $\Theta \sim \mathcal{U}[0, 2\pi]$ ,  $T_1$  has the same distribution as  $\cos \Theta$  and its pdf is  $t \mapsto \frac{1}{2\pi} \frac{1}{\sqrt{1-t^2}}$ .

3.  $R$  and  $\Theta$  are independent random variables so any  $f(R)$  and  $g(\Theta)$  are also independent random variables, for any function  $f$  and  $g$  defined on  $R(\Omega)$  and  $\Theta(\Omega)$ . By choosing  $f : x \mapsto x^2$  and  $g : x \mapsto \cos(x)$  we conclude that:

$T_1 = g(\Theta)$  and  $V = f(R)$  are independent.

d)

Based on question 1. and 2., we can show that  $S$  has a Rayleigh distribution:

$$\begin{aligned} \forall s \geq 0, F_S(s) &= \mathbb{P}(S \leq s) && \text{Definition} \\ &= \mathbb{P}(\sqrt{-2 \log(V)} \leq s) && \text{Using } V = V_1^2 + V_2^2 \\ &= \mathbb{P}\left(V \geq \exp\left(-\frac{s^2}{2}\right)\right) \\ &= 1 - \mathbb{P}\left(V \leq \exp\left(-\frac{s^2}{2}\right)\right) \\ &= 1 - \exp\left(-\frac{s^2}{2}\right) && \text{Because } V \sim \mathcal{U}[0, 1] \end{aligned}$$

This shows that the cdf of  $S$  is the cdf of a Rayleigh distribution.

So  $S$  has a Rayleigh distribution.

What's more we showed in question 3.c) that  $T_1$  has the same distribution as  $\cos(\Theta)$ . We can show with the same reasoning that  $\frac{V_2}{\sqrt{V_1^2 + V_2^2}}$  has the same distribution as  $\sin(\Theta)$ .

With question 1. we conclude that both  $X$  and  $Y$  have a  $\mathcal{N}(0, 1)$  distribution.

Still with question 1., we know that  $X$  and  $Y$  are independent.

$(X, X)$  follow a  $\mathcal{N}(0, I_2)$  distribution.

## Exercise 2: Invariant distribution

1)

Let  $n \geq 0$ :

- We suppose that  $X_n$  can't be written as  $\frac{1}{k}$ , ( $k \in \mathbb{N}^*$ )

$$\begin{aligned}\mathbb{P}(X_{n+1} \in A | X_n \notin \{\frac{1}{k}, k \in \mathbb{N}^*\}) &= \int_A \mathbb{1}_{[0,1]}(t) dt \\ &= \int_{A \cap [0,1]} dt\end{aligned}$$

- We suppose that it exists a positive integer  $k$  such that  $X_n = \frac{1}{k}$ .  
Let's define the random variable  $Y_n \sim \mathcal{B}(X_n)$ . We can reformulate the transition from  $X_n$  to  $X_{n+1}$  as follows:

$$\begin{cases} X_{n+1} = \frac{1}{k+1} & \text{if } Y_n = 0 \\ X_{n+1} \sim \mathcal{U}[0, 1] & \text{if } Y_n = 1 \end{cases}$$

With the law of total probability we deduce that:

$$\begin{aligned}\mathbb{P}\left(X_{n+1} \in A | X_n = \frac{1}{k}\right) &= \mathbb{P}(Y_n = 0) \mathbb{P}\left(X_{n+1} | X_n = \frac{1}{k}, Y_n = 0\right) + \mathbb{P}(Y_n = 1) \mathbb{P}\left(X_{n+1} | X_n = \frac{1}{k}, Y_n = 1\right) \\ &= \left(1 - \frac{1}{k^2}\right) \delta_{\frac{1}{k+1}}(A) + \frac{1}{k^2} \int_{A \cap [0,1]} dt\end{aligned}$$

Hence we finally have:

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

2)

Let  $\pi$  the pdf of the uniform distribution on  $[0, 1]$ . We have  $\pi(dx) = dx$  :

- if  $x \neq \frac{1}{k}$  ( $k \in \mathbb{N}^*$ ):

$$\int_{[0,1]} P(x, A) dx = \int_{[0,1]} \int_{A \cap [0,1]} dt dx = \pi(A)$$

- if  $x = \frac{1}{k}$ , ( $k \in \mathbb{N}^*$ ), the set  $\{\frac{1}{k}, k \in \mathbb{N}^*\}$  is a countable set of real numbers, so it's Lebesgue measure is 0, so:

$$\int_{[0,1]} P(x, A) = \int_{[0,1]} \int_{A \cap [0,1]} dt dx = \pi(A)$$

We finally get the equality  $\int_{[0,1]} P(x, A) \pi(dx) = \pi(A)$  for any measurable subset  $A \subseteq [0, 1]$ .

$\pi$  is invariant to the transition kernel  $P$ .

3)

Let  $x \notin \{\frac{1}{k}, k \in \mathbb{N}^*\}$ , we have:

$$\begin{aligned}
Pf(x) &= \mathbb{E}[f(X_1)|X_0 = x] \\
&= \int f(t)P(x, dt) \quad \text{Definition} \\
&= \int_{[0,1]} f(t)\pi(t)dt \quad x \notin \left\{ \frac{1}{k}, k \in \mathbb{N}^* \right\} \Rightarrow P(x, \cdot) = \pi(\cdot)
\end{aligned}$$

Let  $n \geq 1$ :

$$\begin{aligned}
P^n f(x) &= \mathbb{E}[f(X_n)|X_0 = x] \\
&= \mathbb{E}\left[f(X_n)|X_0 = x, X_1 \neq \frac{1}{k}, k \in \mathbb{N}^*\right] \mathbb{P}\left(X_1 \neq \frac{1}{k}, k \in \mathbb{N}^* \mid X_0 = x\right) \\
&\quad + \mathbb{E}\left[f(X_n)|X_0 = x, X_1 = \frac{1}{k}, k \in \mathbb{N}^*\right] \mathbb{P}\left(X_1 = \frac{1}{k}, k \in \mathbb{N}^* \mid X_0 = x\right) \quad \text{Law of total probabilities}
\end{aligned}$$

As  $X_1 \mid X_0 = x \sim U[0, 1]$ , we deduce that:

$$\begin{cases} \mathbb{P}(X_1 \neq \frac{1}{k}, k \in \mathbb{N}^* \mid X_0 = x) = 1 \\ \mathbb{P}(X_1 = \frac{1}{k}, k \in \mathbb{N}^* \mid X_0 = x) = 0 \end{cases}$$

Because  $\{\frac{1}{k}, k \in \mathbb{N}^*\}$  is countable, hence of null measure.

So

$$\begin{aligned}
P^n f(x) &= \mathbb{E}[f(X_n) \mid X_0 = x, X_1 \neq \frac{1}{k}, k \in \mathbb{N}^*] \\
&= \mathbb{E}[f(X_n) \mid X_1 \neq \frac{1}{k}, k \in \mathbb{N}^*] \quad \text{Markov property} \\
&= \mathbb{E}[\mathbb{E}[f(X_n) \mid X_{n-1}] \mid X_1 \neq \frac{1}{k}, k \in \mathbb{N}^*] \\
&= \mathbb{E}[f(X_{n-1}) \mid X_1 \neq \frac{1}{k}, k \in \mathbb{N}^*]
\end{aligned}$$

By iteratively repeating the process we get:

$$P^n f(x) = \mathbb{E}[f(X_1) \mid X_0 = x] = \int_{[0,1]} f(t)\pi(t)dt$$

We conclude that

$$\boxed{\lim_{n \rightarrow +\infty} P^n f(x) = \int_{[0,1]} f(t)\pi(t)dt}$$

4)

Let  $x = \frac{1}{k}, k \geq 2$ :

a)

Now we have  $P(x, A) = x^2 \int_{A \cap [0,1]} dt + (1 - x^2) \delta_{\frac{1}{k+1}}(A)$ .

If  $x = \frac{1}{k}$ , the chain can reach:

$$\begin{cases} \frac{1}{k+1} \text{ with a probability } 1 - \frac{1}{k^2} \\ \text{be uniformly distributed on } [0, 1] \text{ with probability } \frac{1}{k^2} \end{cases}$$

So to reach  $\frac{1}{k+n}$  after  $n$  steps, the chain should always reach  $\frac{1}{k+i}$ , ( $0 \leq i \leq n$ ), as it's very unlikely to reach  $\frac{1}{k+n}$  if it's resampled at least once uniformly on  $[0, 1]$ .

Hence we expect the probability to reach  $\frac{1}{k+n}$  starting from  $\frac{1}{k}$  after  $n$  steps with a probability  $\prod_{i=1}^{n-1} \left(1 - \frac{1}{(k+i)^2}\right)$

So we show by induction on  $n \in \mathbb{N}^*$ ,  $\mathcal{P}_n : P^n\left(\frac{1}{k}, \frac{1}{k+n}\right) = \prod_{i=0}^{n-1} \left(1 - \frac{1}{(k+i)^2}\right)$

- For  $n = 1$  we have  $P\left(\frac{1}{k}, \frac{1}{k+1}\right) = \left(1 - \frac{1}{k^2}\right)$  by definition of the transition kernel.
- Let  $n \in \mathbb{N}^*$ , we suppose that we have  $\mathcal{P}_n$ .

$$\begin{aligned}
P^{n+1}\left(\frac{1}{k}, \frac{1}{k+n+1}\right) &= P\left(P^n\left(\frac{1}{k}, \frac{1}{n+k+1}\right)\right) \\
&= \int_{[0,1]} P\left(t, \frac{1}{k+n+1}\right) P^n\left(\frac{1}{k}, dt\right) \\
&= P\left(\frac{1}{n+k}, \frac{1}{n+k+1}\right) P^n\left(\frac{1}{k}, \frac{1}{n+k}\right) \quad P\left(t, \frac{1}{n+k+1}\right) = 0 \text{ for } t \neq \frac{1}{n+k} \\
&= \left(1 - \frac{1}{(k+n)^2}\right) \prod_{i=1}^{n-1} \left(1 - \frac{1}{(k+i)^2}\right) \quad \text{As we have } \mathcal{P}_n
\end{aligned}$$

So

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

**b)**

Let  $A = \bigcup_{q \in \mathbb{N}} \left\{ \frac{1}{k+1+q} \right\}$ . We have:

- $P^n(x, A) \neq 0 \Leftrightarrow \delta_{\frac{1}{k+n}}(A) \neq 0 \Leftrightarrow q = n-1$
- $\pi(A) = 0$  because  $A$  is a countable set, hence for null Lebesgue measure.

$$P^n(x, A) = \sum_{q \in \mathbb{N}} P^n\left(x, \frac{1}{k+1+q}\right) = P^n\left(x, \frac{1}{k+n}\right) = \prod_{i=1}^{n-1} \left(1 - \frac{1}{(k+i)^2}\right) \quad (\text{Previous question})$$

$$\begin{aligned}
\prod_{i=1}^{n-1} \left(1 - \frac{1}{(k+i)^2}\right) &= \prod_{i=1}^{n-1} \left(1 - \frac{1}{k+i}\right) \prod_{i=1}^{n-1} \left(1 + \frac{1}{k+i}\right) \\
&= \prod_{i=1}^{n-1} \left(\frac{k+i-1}{k+i}\right) \prod_{i=1}^{n-1} \left(\frac{k+i+1}{k+i}\right) \\
&= \frac{k}{k+n-1} \times \frac{k+n}{k+1} \quad \text{Telescopic product} \\
&= \frac{k}{k+1} \times \frac{k+n}{k+n-1} \xrightarrow{n \rightarrow +\infty} \frac{k}{k+1}
\end{aligned}$$

$$\lim_{n \rightarrow +\infty} P^n\left(\frac{1}{k}, A\right) = \frac{k}{k+1} \neq 0 = \pi(A)$$

### Exercise 3: Stochastic Gradient Learning in Neural Networks

1)

1 **Input:**

$\left\{ (x_i, y_i)_{i \in \llbracket 1, n \rrbracket} \right\}$  a set of input-output samples ( $n \in \mathbb{N}$ )  
 $(\eta_k)_{k \in \mathbb{N}}$  a sequence of learning rates  
 $\varepsilon$  a tolerance

2  $w \leftarrow w_0$

3  $k \leftarrow 1$

4 **while** not stopping criterion :

5   **sample**  $(x_{i_k}, y_{i_k}) \in \left\{ (x_i, y_i)_{i \in \llbracket 1, n \rrbracket} \right\}$

6    $w \leftarrow w + \eta_k \nabla_w R(w_k, x_{i_k}, y_{i_k})$

7    $k \leftarrow k + 1$

8 **output**  $w$

Where:

- $\nabla_w R(w_k, x_{i_k}, y_{i_k}) = -2(y_{i_k} - w^T x_{i_k}) x_{i_k}$

- $\eta_k = \eta_0 d^{\lfloor \frac{k+1}{r} \rfloor}, (\eta_0 < 1)$

Defining the learning rate series that way, it's updated every  $r$  steps of the descent, and decreased by a ratio  $d$ . There are many other ways to define this sequence, but this one is intuitive.

**Here is an implementation of the algorithm that can be found in [this notebook](#):**

```
def SGD(X, y, max_iter, init_lr, d=0.7, r=10):
    """
    init_lr: initial learning rate
    d: how much the learning rate should change at each drop
    r: drop rate (how often the rate should be dropped)
    """

    w = np.random.randn(X.shape[1])
    losses = []
    accuracies = []

    for i in range(max_iter):
        lr = init_lr * d**np.floor((i+1) / r)
        k = np.random.randint(0, len(y))
        x_k, y_k = X[k, :], y[k]

        # Weights update
        w += lr * 2 * (y_k - w @ x_k.T) * x_k
        w /= norm(w) # set norm to 1

        y_pred = predict(X, w)
        accuracies.append(accuracy(y_pred, y))
        loss = np.mean((y - w @ X.T)**2)
        losses.append(loss)

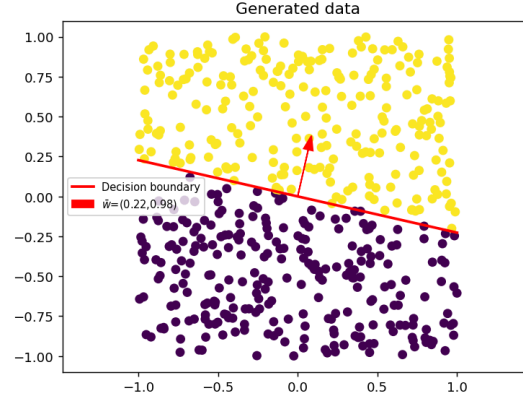
    return w, losses, accuracies
```

*Note about the code:* The accuracy function is coded in the notebook and implements the following formula:

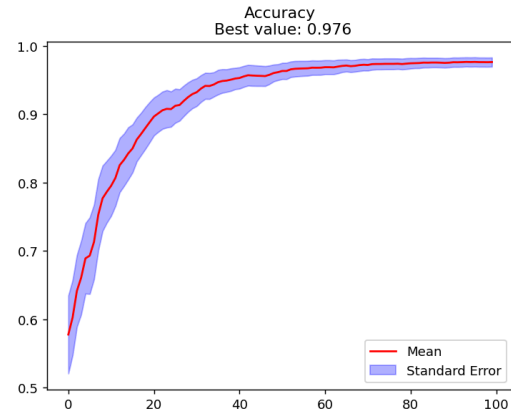
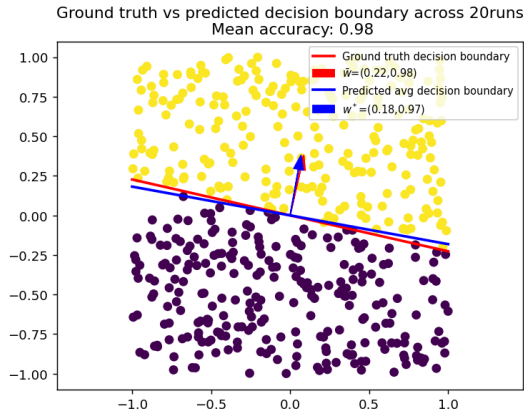
$$\text{Accuracy}(\hat{y}, y) = \frac{1}{n} \sum_{i=1}^n |\hat{y}_i - y_i| \quad (\hat{y}, y) \in \mathbb{R}^n$$

2)

The code used to generate the following figures can be found on [this notebook](#).



3)



The training process was performed with the following hyper-parameters on  $n = 500$  data points. To be able to estimate the standard error, thus evaluating the robustness of the model, the training is repeated  $n_{\text{runs}} = 20$  times. The red curve on the accuracy plot correspond to the mean curve cross the  $n_{\text{runs}}$  attempts. Finally we use the classical standard error estimator  $\sigma_{\text{std}} = \frac{\hat{\sigma}}{\sqrt{n_{\text{runs}}}}$ .

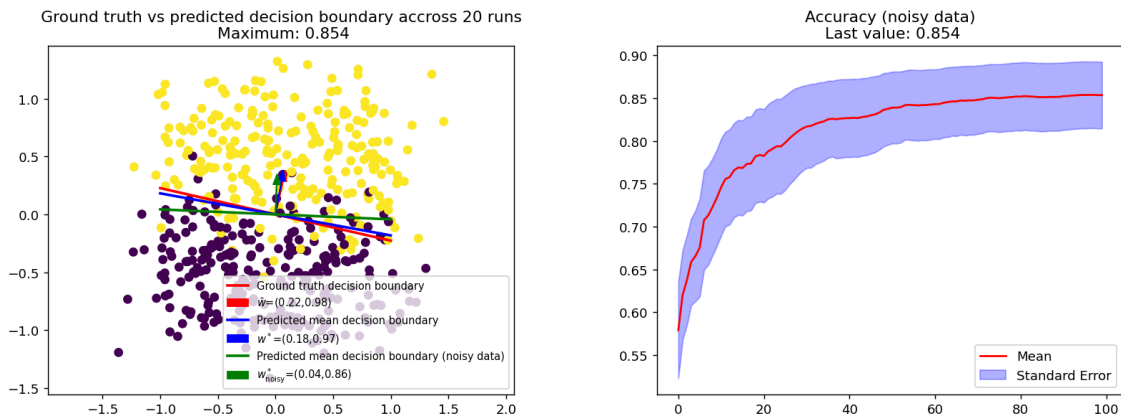
Here are specified the hyper-parameters used to generate the plots:

- `init_lr = 0.1`
- `max_iter = 100`
- `n_runs = 20`

We get a boundary close to the ground truth with a low standard error, as illustrated in the accuracy curve.



4)



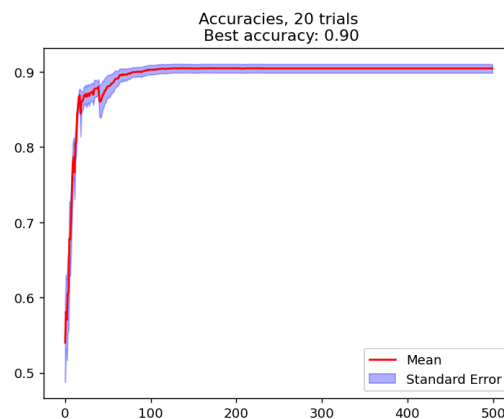
Noisy data was generated by adding a Gaussian noise of mean 0 and variance 0.04.

We notice that the decision boundary predicted on noisy is farther from the ground-truth than the one predicted on not noisy data. In addition, the accuracy curve shows that the standard error is larger than for the non-noised version of the dataset.

5)

#### Implementation details for the Breast Cancer Wisconsin dataset

- 30% of the dataset is used as a test set
- The training set is standardized so that we don't have to fit a bias
- The test set is standardized using the mean and the variance computed on the training set
- `lr_init = 0.008`
- `max_iter = 500`
- `n_runs = 20`



We get an accuracy of  $0.89 \pm 0.01$  on the test set.