INSA Rennes, cours ARO07–MSSD

Random Models of Dynamical Systems
Introduction to SDE's

TP 2 : Numerical simulation of stochastic differential equations
Wright–Fisher diffusion approximation

December 8, 2021

The objective of this problem is to show the efficiency of the diffusion approximation principle. As an example to illustrate this claim, consider the following simplified model for the reproduction of individuals through the transmissions of alleles (alternative types of the same gene). Consider here the case of one gene, with two alleles A and a. The population size N is assumed finite and constant at each generation. At generation k, each individual inherits the allele of its parent, a randomly (uniformly) selected (with replacement) individual present in the population at generation $(k-1)$.

Define the random variable Y_k^N, taking values in the finite set $\{0, 1, \cdots, N\}$, to be the number of allele of type A present at generation k. Recall [see Exercise 3 in TD 3] that the sequence $(Y_k^N\ ,\ k \geq 0)$ forms a Markov chain with transition probability matrix

$$\pi_{i,j}^N = \mathbb{P}[Y_k^N = j \mid Y_{k-1}^N = i] = \binom{N}{j} \left(\frac{i}{N}\right)^j \left(1 - \frac{i}{N}\right)^{N-j} ,$$

for any $i, j \in \{0, 1, \cdots, N\}$. Thinking more in term of proportions, rather than in terms of number of individuals, introduce the normalized random variable $X_k^N = Y_k^N/N$, taking values in the finite subset $\{0, 1/N, \cdots, 1 - 1/N, 1\} \subset [0, 1]$, and the piecewise linear continuous–time Markov process

$$X^N(t) = X_{k-1}^N + (X_k^N - X_{k-1}^N)\,(N\,t - (k-1)) \qquad \text{for any } k-1 \leq N\,t \leq k$$

interpolating points X_k^N at time instants $t_k^N = k/N$.

(i) **For a given population size N, simulate a few realizations of the Markov process $X^N(t)$ on the time interval $[0, T]$.**

For each realization, check out whether the sample–path reaches one of the two absorbing boundaries $x = 0$ or $x = 1$ before time T, and plot the sample–path with the hitting time highlighted.

Take $T = 10$, try different initial proportions at time $t = 0$, say $x = 0.1$, $x = 0.2$, $x = 0.3$, $x = 0.4$ and $x = 0.5$ for instance, and different population sizes, $N = 1000$, $N = 2000$, $N = 5000$ and $N = 10000$ for instance.

[Hint: simulate realizations of the Markov chain Y_k^N for $k = 0, 1, \cdots, NT$, then compress space and time simultaneously by a factor $1/N$.]

Recall [see Exercise 3 in TD 3] that, as N increases to infinity, the piecewise linear continuous–time Markov process $X^N(t)$ interpolating points X_k^N at time instants $t_k^N = k/N$, converges in distribution to the solution $X(t)$ of the SDE

$$X(t) = X(0) + \int_0^t \sqrt{X(s)\,(1 - X(s))}\, dB(s) \ .$$

taking values in the interval $[0, 1]$. Clearly, the diffusion coefficient of this SDE does not satisfy the known sufficient conditions for existence and uniqueness of a solution. However [see Written Exam, Academic Year 20/21], there does exist a unique solution to this particular SDE.

(ii) **Solve numerically the SDE on the time interval $[0, T]$, using a Euler scheme with time step h and using simulated Brownian motion increments. Repeat the same experiment a few times (using different Brownian motion increments in each different experiment).**

For each experiment, check out whether the sample–path reaches one of the two absorbing boundaries $x = 0$ or $x = 1$ before time T, and plot the sample–path with the hitting time highlighted.

Take $T = 10$, try different initial proportions at time $t = 0$, say $x = 0.1$, $x = 0.2$, $x = 0.3$, $x = 0.4$ and $x = 0.5$ for instance, and different time steps, from $h = 2^{-6}$ down to $h = 2^{-10}$ for instance.

Heuristically, the diffusion approximation principle provides an approximate value for the mean hitting time of the Markov chain Y_k^N (or the normalized Markov chain X_k^N) when the population size N is large: indeed

$$\begin{aligned}
\tau^N &= \inf\{k \geq 0 \,:\, Y_k^N = 0 \text{ or } Y_k^N = N\}/N \\
&= \inf\{k \geq 0 \,:\, X_k^N = 0 \text{ or } X_k^N = 1\}/N \\
&= \inf\{t \geq 0 \,:\, X^N(t) = 0 \text{ or } X^N(t) = 1\} \\
&\approx \inf\{t \geq 0 \,:\, X(t) = 0 \text{ or } X(t) = 1\} = \tau \ ,
\end{aligned}$$

hence $\mathbb{E}_{0,x}[\tau^N] \approx \mathbb{E}_{0,x}[\tau]$. Recall [see Exercise 4 in TD 3] that the explicit expression

$$\mathbb{E}_{0,x}[\tau] = -2\,[x \log x + (1 - x) \log(1 - x)] \ ,$$

holds for any $0 \leq x \leq 1$.

The objective is now to obtain an accurate numerical approximation of the mean hitting time of the boundary $\{0, 1\}$ by the process $X(t)$ starting from the proportion $x \in [0, 1]$ at time $t = 0$, using empirical Monte Carlo approximation based on Euler scheme. In practice, for any $x \in [0, 1]$, given a step–size h and a sample size M (the number of independent Monte Carlo experiments), the procedure consists of

- for any $i = 1, \cdots, M$, solve numerically the SDE starting from x at time $t = 0$, using a Euler scheme with step–size h, and store the value of the first hitting time $\bar{\tau}_i^h$ of the boundary $\{0, 1\}$ by the approximate solution $\bar{X}_i^h(t)$,

- compute the empirical mean

$$\mathrm{MHT}_x^{h,M} = \frac{1}{M} \sum_{i=1}^{M} \bar{\tau}_i^h \ ,$$

as an estimator (depending on h and M) of the mean hitting time $\mathbb{E}_{0,x}[\tau]$.

(iii) **Evaluate and plot the empirical mean $\mathrm{MHT}_x^{h,M}$ as a function of the initial proportion x, for different values of h and M. Compare this estimator with the available explicit expression for the mean hitting time $\mathbb{E}_{0,x}[\tau]$.**

Try different initial conditions at time $t = 0$, say $x = 0.1$, $x = 0.2$, $x = 0.3$, $x = 0.4$ and $x = 0.5$ for instance, different time steps, from $h = 2^{-6}$ down to $h = 2^{-10}$ for instance, and different sample sizes, $M = 100$, $M = 1000$ up to $M = 10000$ for instance.

[Hint: to avoid any loop in the code, that would slow down the execution, do not simulate the M approximate solutions *sequentially*, and rather simulate these M approximate solutions *simultaneously*, using the vector facilities offered by MATLAB or Python.]

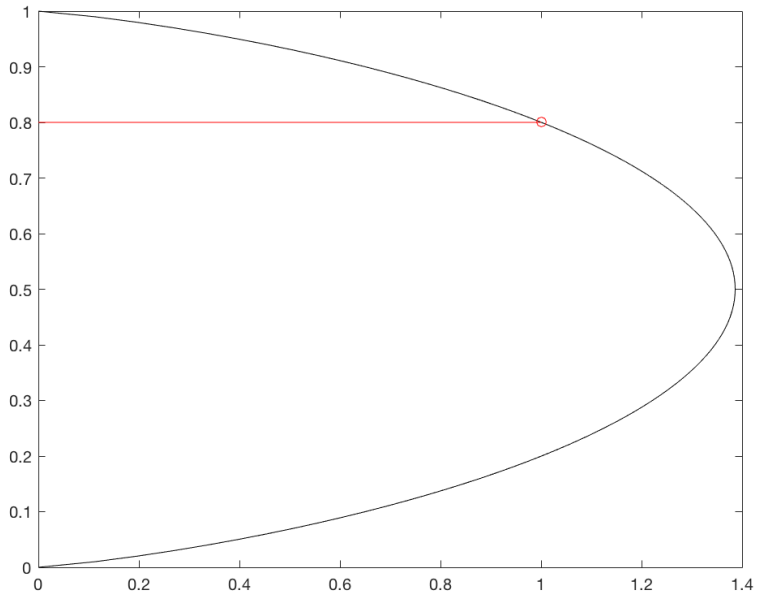

Figure 1: Mean hitting time $\mathbb{E}_{0,x}[\tau]$ — Exact expression vs. starting point $x \in [0, 1]$ (black solid line) — Empirical estimation for starting point $x = 0.8$ (red line and circle)

At this stage, it could be argued that a numerical approximation of the mean hitting time $\mathbb{E}_{0,x}[\tau^N]$ could be obtained directly, using a similar empirical Monte Carlo approximation for the Wright–Fisher model directly, i.e. without using the Wright–Fisher diffusion. Indeed, on top

of Monte Carlo approximation, there are two additional levels of approximation in the diffusion approximation principle: obviously, one is letting the population size N go to infinity, and the other is using a Euler scheme to solve numerically the Wright–Fisher diffusion obtained in the limit. A direct approach would skip these two levels of approximation, and only Monte Carlo approximation would be needed. In practice, for any initial proportion x at time 0, given a sample size M (the number of independent Monte Carlo experiments), the procedure consists of

- for any $i = 1, \cdots, M$, simulate a realization of the Markov process starting from x at time $t = 0$, and store the value of the first hitting time τ_i^N of the boundary $\{0, 1\}$ by the realization $X_i^N(t)$,
- compute the empirical mean

$$\mathrm{MHT}_x^{N,M} = \frac{1}{M} \sum_{i=1}^{M} \tau_i^N \ ,$$

  as an estimator (depending on M only, since the population size N is fixed here) of the mean hitting time $\mathbb{E}_{0,x}[\tau^N]$.

(iv) **For a given initial proportion at time $t = 0$, say $x = 0.3$ for instance, evaluate the empirical mean $\mathrm{MHT}_x^{N,M}$ for different values of M. Compare this estimator with the available explicit expression for the mean hitting time $\mathbb{E}_{0,x}[\tau]$ of the Wright–Fisher diffusion.**

Try different sample sizes, $M = 100$, $M = 1000$ up to $M = 10000$ for instance.

(v) **Compare and discuss the computation time needed to obtain**

- **the empirical mean $\mathrm{MHT}_x^{N,M}$ in the Wright–Fisher model with a finite (but large) population size N,**
- **the empirical mean $\mathrm{MHT}_x^{h,M}$ in the limiting Wright–Fisher diffusion.**

To make the comparison fair, take the same sample size M (number of independent Monte Carlo experiments) in both cases, recall that $1/N$ is interpreted as a time step in the diffusion approximation, and take h of the same order as $1/N$.