HW2 Molecular Programming

MPRI 2.11.1

06.01.2022 - Due on Thu. 20/01 before 08:45

You are asked to complete the exercise marked with a [★] and to send me your solutions to:
nicolas.schabanel@ens-lyon.fr
as a PDF file named **HW2-Lastname.pdf** on **Thu. 20/01 before 08:45**.

‼ Pick *one of the two* exercises marked with a [★] ‼ You can work on the origami in team of 2-4 people

■ **Exercise 1 (Exponential random variables & kTAM implementation).** Recall that an *exponential random variable* X with parameter $\lambda > 0$ is defined by: $(\forall x \geqslant 0)\ \Pr\{X \geqslant x\} = e^{-\lambda x}$.

▶ **Question 1.1)** *Compute* $\mathbb{E}[X]$.
▷ *Hint. Recall that if X is a non-negative random variable, then $\mathbb{E}[X] = \int_0^\infty \Pr\{X \geqslant x\}dx$.*
Answer. ▷ Indeed, by Fubini's theorem:

$$\begin{aligned}\mathbb{E}[X] &= \int_{x=0}^{\infty} x \Pr\{X \in [x, x+dx]\} = \int_{x=0}^{\infty} \left(\int_0^x 1\, dy\right) \Pr\{X \in [x, x+dx]\}\\ &= \int_0^\infty \int_{x=y}^{\infty} \Pr\{X \in [x, x+dx]\}\, dy = \int_0^\infty \Pr\{X \geqslant y\}\, dy\\ &= \int_0^\infty e^{-\lambda x} dx = 1/\lambda.\end{aligned}$$

◁

▶ **Question 1.2)** *Show that the exponential distribution is memoryless, i.e. if X is exponentially distributed with parameter λ, then $(\forall t, u \geqslant 0)\ \Pr\{X \geqslant t+u \,|\, X \geqslant t\} = \Pr\{X \geqslant u\}$.*
Answer. ▷ $\Pr\{X \geqslant t+u \,|\, X \geqslant t\} = \dfrac{\Pr\{X \geqslant t+u \wedge X \geqslant t\}}{\Pr\{X \geqslant t\}} = \dfrac{\Pr\{X \geqslant t+u\}}{\Pr\{X \geqslant t\}} = e^{-\lambda(t+u)+\lambda t} = e^{-\lambda u} = \Pr\{X \geqslant u\}$. ◁

Let X and Y be two independent exponentially distributed random variables with respective parameters λ and μ.

▶ **Question 1.3)** *Show that $\min(X, Y)$ is also exponentially distributed. What is its parameter?*
Answer. ▷ Let $M = \min(X, Y)$, then:

$$\begin{aligned}\Pr\{M \geqslant t\} &= \Pr\{X \geqslant t \wedge Y \geqslant t\}\\ &= \Pr\{X \geqslant t\} \cdot \Pr\{Y \geqslant t\}, \quad \text{by independence of } X \text{ and } Y\\ &= e^{-(\lambda+\mu)t}.\end{aligned}$$

M is thus exponentially distributed with parameter $\lambda + \mu$. ◁

▶ **Question 1.4)** *What is the probability that* $\min(X, Y) = X$*?*
Answer. ▷ Let $M = \min(X, Y)$, then:

$$\begin{aligned}
\Pr\{M = X\} &= \int_{t=0}^{\infty} \Pr\{Y > t \wedge X \in [t, t + dt]\} \\
&= \int_{t=0}^{\infty} \Pr\{Y > t\} \cdot \Pr\{X \in [t, t + dt]\}, \quad \text{by independence of } X \text{ and } Y \\
&= \int_{t=0}^{\infty} e^{-\mu t} \cdot \lambda e^{-\lambda t} dt = \frac{\lambda}{\lambda + \mu}.
\end{aligned}$$

◁

▶ **Question 1.5)** *Same questions as the two above for* n *indenpendent exponentially dsitributed variables* $X_1, \ldots, X_n$ *with parameters* $\lambda_1, \ldots, \lambda_n$*.*
Answer. ▷ By immediate recurrence, let $Z = \min(X_1, \ldots, X_n)$ and $\Lambda = \lambda_1 + \cdots + \lambda_n$:

$$Z \sim \text{Exp}(\Lambda), \quad \mathbb{E}[Z] = \frac{1}{\Lambda}, \quad \text{and } \Pr\{Z = X_i\} = \frac{\lambda_i}{\Lambda}.$$

◁

▶ **Question 1.6)** *Assume that a non-negative random variable* X *is given by its tail distribution* $F(x) = \Pr\{X \geqslant x\}$*. Show that* X *is identically distributed as* $F^{-1}(U)$ *where* U *is a uniform random variable in* $[0, 1]$*.*

Describe how to sample an exponential random variable of rate λ*.*
Answer. ▷ First, note that F is a non-increasing function going from 1 at $x = 0$, to 0 at ∞, its inverse $F^{-1} : [0, 1] \to \mathbb{R}_+$ is thus defined almost everywhere. Now, for all $x \geqslant 0$, $\Pr\{F^{-1}(U) \geqslant x\} = \Pr\{U \leqslant F(x)\} = F(x) = \Pr\{X \geqslant x\}$ as F is non-increasing. Thus X and $F^{-1}(U)$ are indentically distributed.

Now, if $X \sim \text{Exp}(\lambda)$, we have $F(x) = \Pr\{X \geqslant x\} = e^{-\lambda x}$. As $F^{-1}(u) = -\ln(u)/\lambda$, we can sample X by picking a uniform random value u in $[0, 1]$ and outputing $x = -\ln(u)/\lambda$. ◁

▶ **Question 1.7)** *Propose an algorithm together with a data structure to implement the kTAM model with attachement rate* $r_f = k_f[Strand] = k_f\, e^{-G_{mc}}$ *and detachment rate* $r_{s,b} = k_f\, e^{-b \cdot G_{se}}$ *where* b *is the number of bonds made by the strand with the current agregate.*

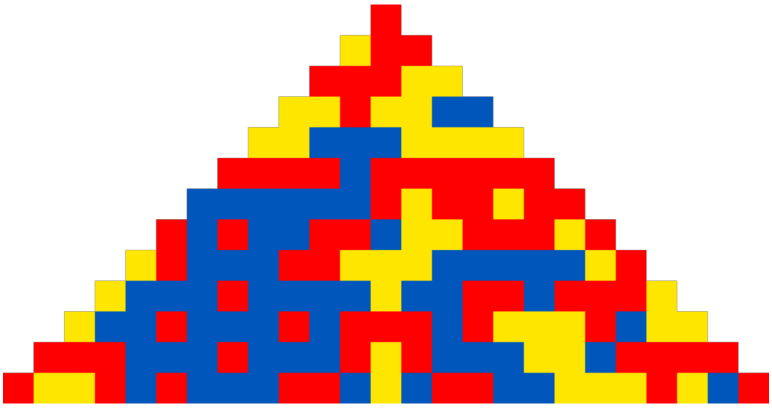
Use parameters $k_f = 10^6/M/sec$*,* $G_{mc} = 12.9$ *and* $G_{se} = 6.5$ *for the algorithmic phase.*
Answer. ▷ We consider cells of $\mathbb{Z}^2$ where the seed is placed forever (the seed must be marked as undetachable). We determine the date of the next attachement (resp. detachment) event in an empty (resp. full) cell of $\mathbb{Z}^2$, using an independent exponential random variable $\text{Exp}(r_f)$ (resp. $\text{Exp}(r_{s,b})$) so that these events happen every $1/r_f$ (resp. $1/r_{s,b}$) on average.

Each time a cell is updated, we draw new random dates for the upcoming events of its neighbors (recall that exponential variable are memoryless). We store all the upcoming events in a priority queue (a heap) from which we pop the next event (the one with the minimum date).

Recall to delete obsolete upcoming events each time one of its neighbors is updated to avoid flooding the memory. Recall also to avoid flooding the heap with attachement events in an isolated cell.

Note that an attachment event is determined as follows: Pick the event time as $t \sim \text{Exp}(k_f \cdot \sum_j [Strand_j])$ and select $Strand_i$ with probability $[Strand_i]/\sum_j [Strand_j]$ as it is distributed as the first time of a strand among $Strand_1, \ldots, Strand_n$ attaches (recall the answers to questions above). ◁

■ **Exercise 2 (Tileset for simulating cellular automata).** A cellular automaton consists of a finite set of *states* Q, a function $f : Q^3 \to Q$, called the *rule*, and an initial configuration $c^0 \in Q^*$. The configuration at time $t+1$ is obtained from the configuration at time t as follows: $c_i^{t+1} = f(c_i^t, c_{i+1}^t, c_{i+2}^t)$ for $0 \leqslant i < |c^t| - 2$. The calculation stops at the first time T such that $|c^T| < 3$ and the result of the computation is c_0^T. A classic visualization of the computation of a cellular automaton consists of a pyramid where the bottom line is the initial configuration and time goes upwards. Here is an example:



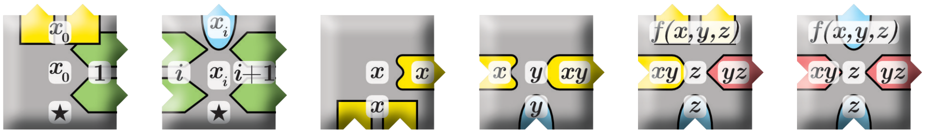
for the rule
$$f(x,y,z) = \begin{cases} \text{🟨} & \text{if } \{x,y,z\} = \{\text{🟥}\} \text{ or } \{\text{🟨},\text{🟦}\} \\ \text{🟦} & \text{if } \{x,y,z\} = \{\text{🟨}\} \text{ or } \{\text{🟦},\text{🟥}\} \\ \text{🟥} & \text{if } \{x,y,z\} = \{\text{🟦}\} \text{ or } \{\text{🟨},\text{🟥}\} \text{ or } \{\text{🟨},\text{🟦},\text{🟥}\} \end{cases}$$

► **Question 2.1)** *Propose a finite tileset whose self-assembly at temperature $T^\circ = 2$ simulates the computation of any Q-state cellular automata from any initial configuration and whose size is independent of the initial configuration length. Give a generic example of the execution of your assembly for generic computation steps. Give the number of variants of each tile type as a function of $|Q|$. Provide the procedure which selects the tiles used to simulate a given Q-state cellular automaton.*

▷ *Hint. Do you need upscaling? Consider reshaping the pyramid to simplify your design.*

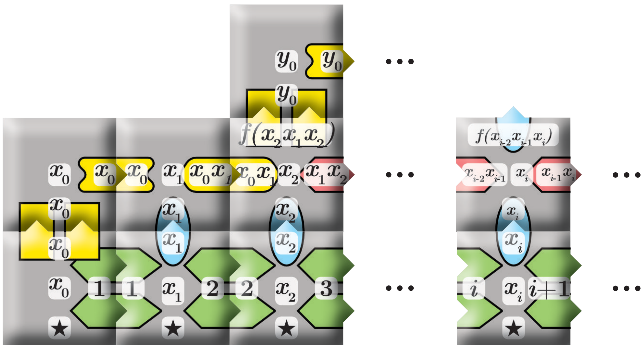
Answer. ▷ As we need to know the states of both neighbors to compute the next step, we choose to slant the pyramid by shifting every row by 2 positions to the right. The first two tiles of each row will collect the first two states of the row and the following tiles on that row will compute the new state and propagate the values of the two new preceding states to the right.

Here is the tileset. The tiles marked with a star are used to encode the initial configuration. There are $O(|Q|^3)$ tiles in total.



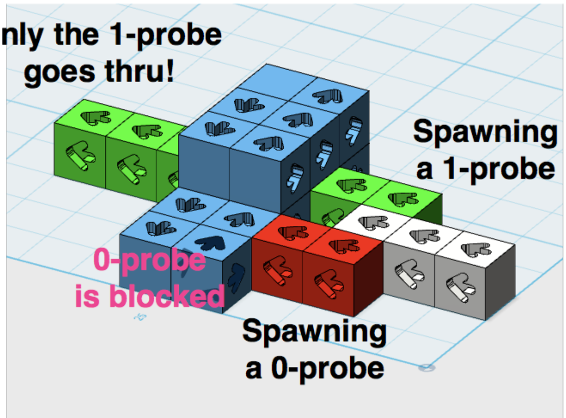
$|c^0|$ tiles $\quad$ $|Q|$ tiles $\quad$ $|Q|^2$ tiles $\quad$ $|Q|^3$ tiles $\quad$ $|Q|^3$ tiles

Here is a generic assembly:



◁

■ **Exercise 3 (Probabilistic simulation Turing Machine at $T^\circ = 1$ in 2D).** Recall that in 3D, for any single-tape binary-alphabet Turing machine M, there is a tile set which simulates M using a clever trick to encode 0s and 1s. These are encoded with bridges and read using two probes where only one go through the bridge:



▶ **Question 3.1)** *By adjusting the concentrations (and thus the rate at which the different tiles attached), describe a tile set together with concentrations for each tile type, that simulates a given single-tape binary-alphabet Turing machine M with an arbitrary small error ε for each symbol read in 2D at temperature $T^\circ = 1$.*

Answer. ▷ No more bridges are possible, so we must give up blocking the read-1 signals. We will then just build a wall to block the read-0 signal whenever we want to write a 1 on the tape. We then increase the concentration of the read-0 signal tiles, i.e. so that the read-0 signal is α more likely to success than the read-1 signal, i.e. it has a probability $1 - 1/\alpha$ to block the read-1 signal if it's free to go, that is, if we intended to write a 0 on the tape (there is never a mistake in reading 1s as in this case the read-0 signal is always blocked). Then, a Turing machine runing in time T and space S, is simulated without mistake with probability at least $(1 - 1/\alpha)^{TS} \geqslant 1/2$ for $\alpha \geqslant TS/\ln 2$. ◁

[★] Exercise 4 (Scale the wall). Recall that a tile assembly system $\mathcal{T} = (T, \sigma, \tau)$ consists of a tile set T, a seed tile $\sigma \in T$ and a temperature $\tau \in \mathbb{N}$. Consider the situation in Fig. 1 consisting of a wall of height h.

▶ **Question 4.1)** *Can you find a tile assembly system $\mathcal{T}$ for the abstract Tile Assembly Model (aTAM) where the rules are as follows?*

- *The seed tile is placed at position $S = (0, 0)$*
- *For all $h \in \mathbb{N}$, every terminal assembly of $\mathcal{T}$ should place a tile at the target position $T = (10, h)$ and be of finite size*

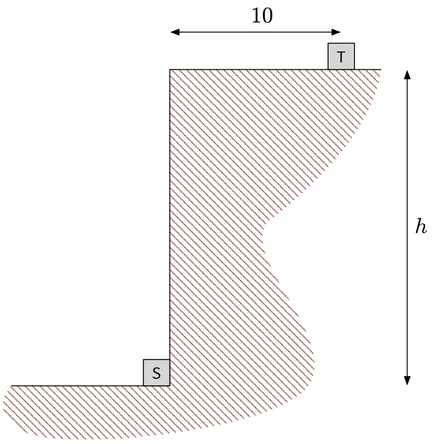


Figure 1: A wall of height h.

- *$\mathcal{T}$ may not place tiles to the right and below the cut of the plane shown in Figure 1.*
- *You may give an infinite sequence of glues such that the h-prefix of that sequence will appear on the wall, to help the tiles 'climb up'.*

Answer. ▷ First notice that there is a simple probabilistic solution to this problem that will always place the tile at position T but will be finite only almost surely (if the concentrations of B and C tiles are at least as large as the concentration of A), see Fig. 2.

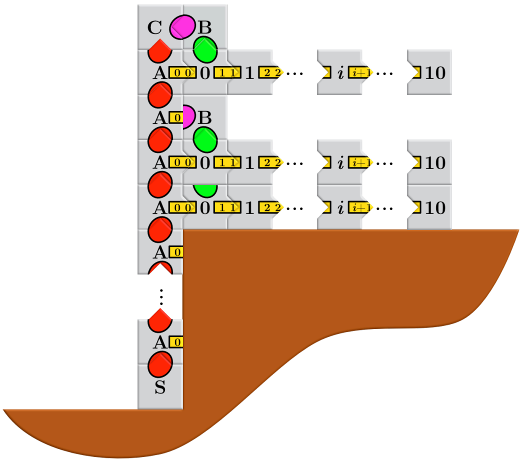


Figure 2: A probabilistic solution.

However, there is no solution for the deterministic case. Indeed, let us assume that there is a deterministic tileset that solves the problem for all $h > 0$. Then, let us show that it can assemble an infinite shape. We will proceed incrementally by building a sequence of shapes of strictly increasing heights as illustrated in Fig 3. The key is consider the assembly just before a tile is attached to the right of the vertical line L of the wall. Let A_1 be the assembly just before that time. Let h_1 be the height of the highest tile just to the left of L in A_1. A_1 is also a valid assembly for the case where $h = h1$. Thus, as the tileset is a solution for $h = h_1$, any extension of A_1 must ultimately solve the problem for $h = h_1$. It must then indeed at some point place a tile to the right of L, strictly above h_1. Let us consider A_2 the assembly just before that and $h_2 > h_1$ the height of the highest tile just to the left of L in A_2. Repeating the process we conclude that the tileset necessarily contains an infinite assembly that never crosses L and thus does not solve the problem.

◁

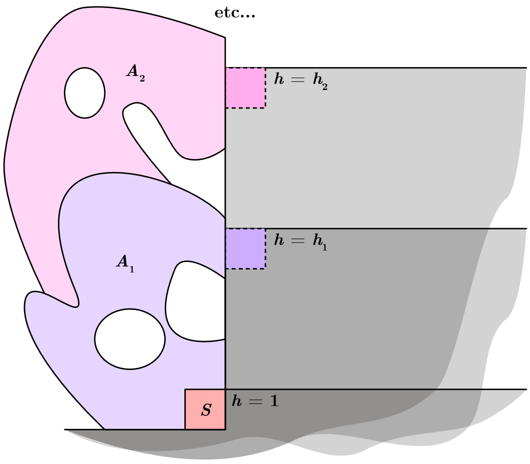


Figure 3: No deterministic solution.

[★] Exercise 5 (DNA origami). Use `ENSnano` to design a DNA origami.

Getting the software. First, download ENSnano (either compiled or source) from:

http://www.ens-lyon.fr/ensnano/

You will also find useful tutorials on ENSnano youtube channel:

https://www.youtube.com/channel/UC-5XMNJouMPsB9pW5Q9UVcA

▶ **Question 5.1)** *1. Pick a 3D shape to design it as a DNA Origami. You might take inspiration from:*

http://www.dna.caltech.edu/Papers/DNAorigami-nature.pdf
or https://www.dietzlab.org/research/

2. *Design your 3D DNA origami using* `ENSnano`
(the scaffold must be at most 7249 nucleotides long)
Compute its expected size in nanometers. Recall that parallel double helices are about 2.65 *nm away when next to each other; two nucleotides are about* 0.332 *nm distant along the helix axis; a complete turns correspond to* ~ 10.44 *nucleotides; the angle between the two helices in the narrow groove is* $\sim \frac{12}{34} \cdot 2\pi$.
3. *email me the* **.ens** *file and a one-page report presenting your origami.*