Exercices sur les modèles de Markov et les HMM

Correction

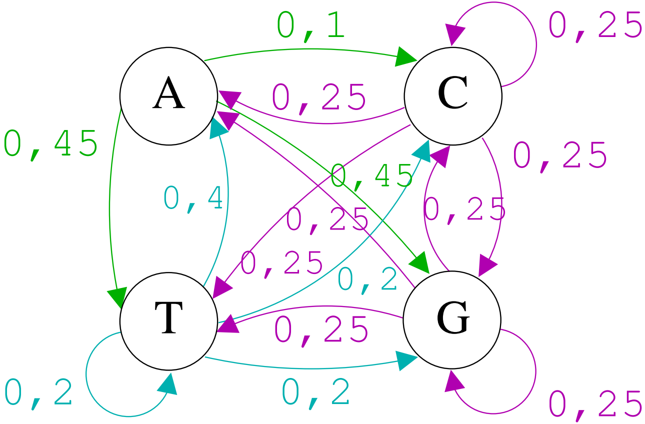
Hélène Touzet

Exercice 1: Le génome de *Simplicimum bestiolus*

On considère un organisme dont le génome suit les règles suivantes:

1. un `A` est suivi soit d'un `C` avec une probabilité de 0.1, soit d'un `G`, soit d'un `T` avec la même probabilité,
2. un `T` a deux fois plus de chance d'être suivi par un `A` que par tout autre nucléotide,
3. les `C` et les `G` sont suivis de n'importe quel nucléotide, avec la même probabilité.

Donnez un modèle de Markov qui décrive ce génome.


A
C
T
G
0,1
0,25
0,25
0,45
0,25
0,4
0,45
0,25
0,25
0,25
0,2
0,25
0,2
0,2
0,25

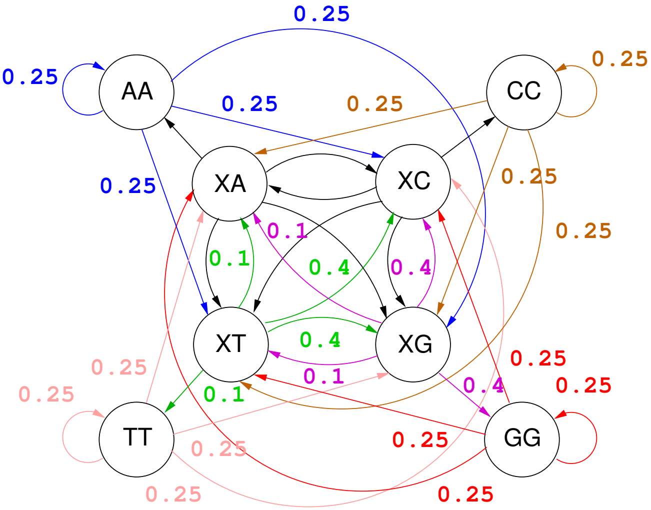
Exercice 2: Le génome de *Bizarrum organismus*

Construire un modèle de Markov pour modéliser le génome de cet organisme:

1. le nucléotide actuel est un `A`, un `C`, un `G` ou un `T` avec une probabilité de 25% si les deux nucléotides précédents sont identiques.
2. le nucléotide actuel a 40% de chance d'être un `C`, 40% de chance d'être un `G`, 10% de chance d'être un `A` et 10% de chance d'être un `T`, si les deux nucléotides précédents sont différents.

huit états

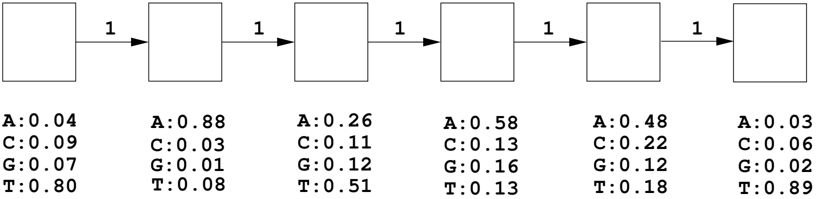
- `AA`: les deux derniers nucléotides lus sont AA
- `CC`: les deux derniers nucléotides lus sont CC
- `GG`: les deux derniers nucléotides lus sont GG
- `TT`: les deux derniers nucléotides lus sont TT
- `XA`: les deux derniers nucléotides lus sont différents et le dernier est A
- `XC`: les deux derniers nucléotides lus sont différents et le dernier est C
- `XG`: les deux derniers nucléotides lus sont différents et le dernier est G
- `XT`: les deux derniers nucléotides lus sont différents et le dernier est T


AA
CC
XA
XC
XT
XG
TT
GG
0.25
0.1
0.4

Exercice 3: TATAAT

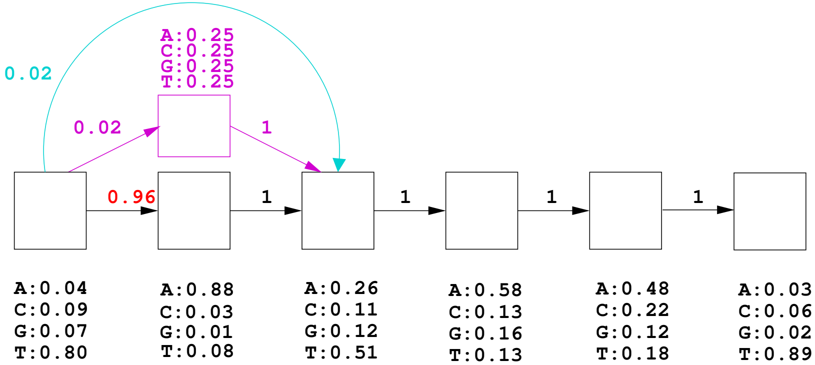
Question 1

	T	A	T	A	A	T
A	0.04	0.88	0.26	0.58	0.48	0.03
C	0.09	0.03	0.11	0.13	0.22	0.06
G	0.07	0.01	0.12	0.16	0.12	0.02
T	0.80	0.08	0.51	0.13	0.18	0.89



Question 2: erreurs de séquençage

la probabilité de chaque type d'erreur à une position donnée est 0,02. Qu'il s'agisse d'une insertion ou d'une délétion, il n'y a à chaque fois qu'un seul nucléotide ajouté ou supprimé, et dans le cas d'une insertion, il peut s'agir d'un A, d'un C, d'un G ou d'un T avec la même probabilité.



à reproduire pour chaque état + 1 état pour une délétion du premier nucléotide + 1 état pour une délétion du dernier nucléotide

Exercice 4: les magiciens

Magicien A: trois mouchoirs bleus et un mouchoir vert

Magicien B: un mouchoir bleu et deux mouchoirs verts

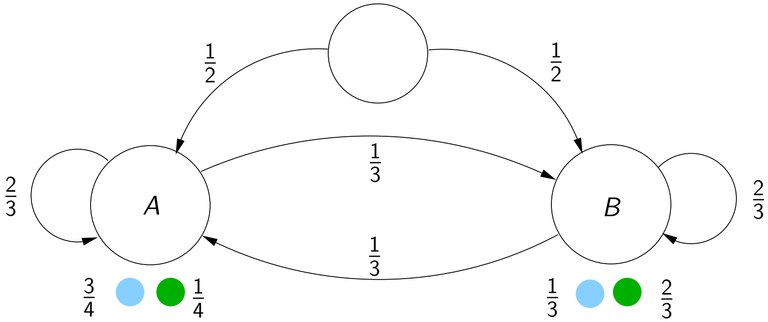
Pour savoir qui commence, les deux magiciens lancent le dé: si le nombre est pair, A commence et prend le dé, sinon B commence et prend le dé;

À chaque étape:

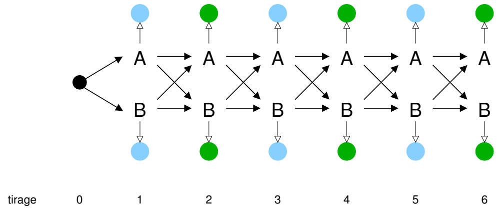
- le magicien qui a le dé tire un mouchoir au hasard dans son chapeau, et le passe à un assistant qui le montre au public, puis remet le mouchoir dans le chapeau;
- le magicien lance le dé: si le nombre est compris entre 1 et 4, il garde le dé. Sinon il le passe à l'autre magicien.

Question 1: modèle de Markov à états cachés

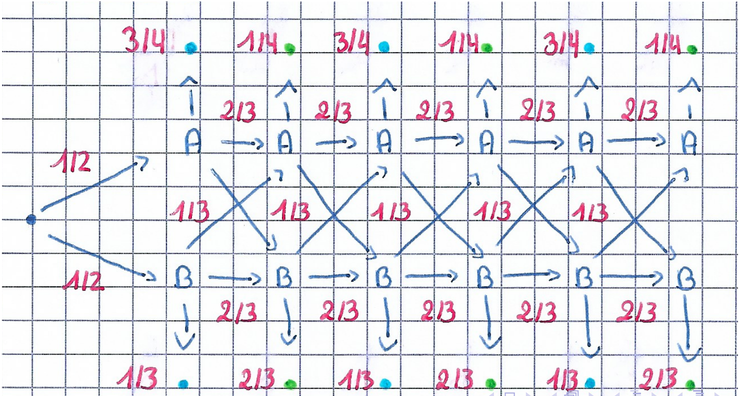
- états: les deux magiciens A et B + état de départ
- transition: lancers du dé
- émissions: couleur des mouchoirs, bleus ou verts



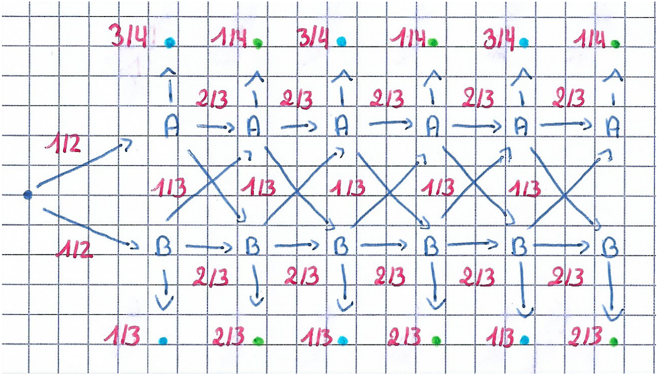
Question 2: observations `Bleu Vert Bleu Vert Bleu Vert`



+ probabilités de transition et d'émission



Question 3: chemin de plus grande probabilité



	1	2	3	4	5	6
A	$3/8$	$1/16A$	$1/32A$	$1/192A$	$1/324B$	$1/486A$
B	$1/6$	$1/12A$	$1/54B$	$1/81B$	$2/729B$	$8/6561B$

état précédent

Meilleur chemin : $ABBBAA$

Question 4: Formules de récurrence pour $P(i, A)$, la probabilité du meilleur chemin jusqu'au $i$ème tirage inclus se terminant sur l'état A et pour $P(i, B)$, la probabilité du meilleur chemin jusqu'au $i$ème tirage inclus se terminant sur l'état B.

$$P(1, A) = \frac{3}{8}$$
$$P(1, B) = \frac{1}{6}$$

$$P(i, A) = \frac{1}{4} \max\{2/3P(i-1, A), 1/3P(i-1, B)\} \text{ si } i \in \{2, 4, 6\}$$
$$P(i, B) = \frac{2}{3} \max\{1/3P(i-1, A), 2/3P(i-1, B)\} \text{ si } i \in \{2, 4, 6\}$$

$$P(i, A) = \frac{3}{4} \max\{2/3P(i-1, A), 1/3P(i-1, B)\} \text{ si } i \in \{3, 5\}$$
$$P(i, B) = \frac{1}{3} \max\{1/3P(i-1, A), 2/3P(i-1, B)\} \text{ si } i \in \{3, 5\}$$

A quoi cela vous fait-il penser ? Il s'agit de l'algorithme de Viterbi

Question 5: probabilité cumulée de tous les chemins compatibles avec l'observation `Bleu Vert Bleu Vert Bleu Vert`.

Formules de récurrence

$$P(1,A)=\frac{3}{8}$$
$$P(1,B)=\frac{1}{6}$$

$$P(i,A)=\frac{1}{4}[2/3P(i-1,A)+1/3P(i-1,B)] \text{ si } i\in\{2,4,6\}$$
$$P(i,B)=\frac{2}{3}[1/3P(i-1,A)+2/3P(i-1,B)] \text{ si } i\in\{2,4,6\}$$

$$P(i,A)=\frac{3}{4}[2/3P(i-1,A)+1/3P(i-1,B)] \text{ si } i\in\{3,5\}$$
$$P(i,B)=\frac{1}{3}[1/3P(i-1,A)+2/3P(i-1,B)] \text{ si } i\in\{3,5\}$$

A quoi cela vous fait-il penser ? Il s'agit de l'algorithme Forward

Exercice 5: Les deux dés

La probabilité de l'événement *passer de deux dés à un dé* est égale à $1/10$. Quand le joueur joue avec un seul dé, la probabilité de repasser à deux dés est de $1/5$.

Deux états

- α: le joueur joue avec un seul dé
- β: le joueur joue avec deux dés

	un	deux	trois	quatre	cinq	six	sept	huit	neuf	dix	onze	douz
α	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	0	0	0	0	0	0
β	0	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

Question 1: comment a été calculée la table des émissions ?

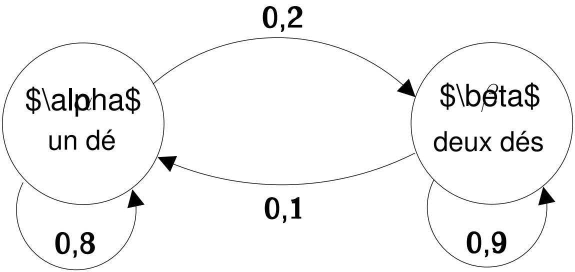
Pour α: le dé a 6 faces et toutes les observations ont la même probabilité

Pour β: on regarde toutes les combinaisons des deux dés

- il y a 36 jets de dés possibles (6×6)
- Pour chaque observation, on compte le nombre de jets dont la somme donne le résultat observé. Par exemple, pour l'observation `huit`, on a 5 jets possibles

$$
\begin{aligned}
2+6 &\rightarrow 8\\
3+5 &\rightarrow 8\\
4+4 &\rightarrow 8\\
5+3 &\rightarrow 8\\
6+2 &\rightarrow 8
\end{aligned}
$$

Question 2: HMM



Les probabilités d'émission sont données par la table précédente.

Question 3: observations `quatre onze six cinq`

Probabilité de chacune des trois suites d'états

- $s_1 = \alpha\alpha\alpha\alpha$
  La probabilité est 0, car l'état α ne peut pas émettre l'observation `onze`
- $s_2 = \beta\beta\beta\beta$

$$\frac{3}{36} * \frac{9}{10} * \frac{2}{36} * \frac{9}{10} * \frac{5}{36} * \frac{9}{10} * \frac{4}{36}$$

- $s_3 = \alpha\beta\beta\beta$

$$\frac{1}{6} * \frac{1}{10} * \frac{2}{36} * \frac{9}{10} * \frac{5}{36} * \frac{9}{10} * \frac{4}{36}$$

Question 4. au lieu d'annoncer le résultat de la somme obtenue, le joueur indique si le résultat est compris entre 1 et 4, ou compris entre 5 et 12.

Nouveau modèle de Markov caché :

- Les états et les probabilités de transition ne changent pas
- On modifie les probabilités d'émission, en sommant les probabilités précédentes.

	`un-quatre`	`cinq-douze`
α	$\frac{4}{6}$	$\frac{2}{6}$
β	$\frac{6}{36}$	$\frac{30}{36}$

Question 5. le joueur indique si le résultat obtenu est pair ou impair. Quelles sont les probabilités d'émission associées à cette règle ? Qu'en pensez-vous ?

- De nouveau, les états et les probabilités de transition ne changent pas
- On somme les probabilités d'émission de `deux`, `quatre`, `six`, `huit`, `dix`, `douze` d'une part, et de `un`, `trois`, `cinq`, `sept`, `neuf`, `onze` d'autre part.

	`pair`	`impair`
α	$\frac{1}{2}$	$\frac{1}{2}$
β	$\frac{1}{2}$	$\frac{1}{2}$

Les deux états ont les mêmes valeurs d'émissions: ils sont indistinguables. Cela signie que l'observateur ne peut pas décoder (retrouver la suite d'états) avec la seule information de la parité.

Exercice 6: Le génie

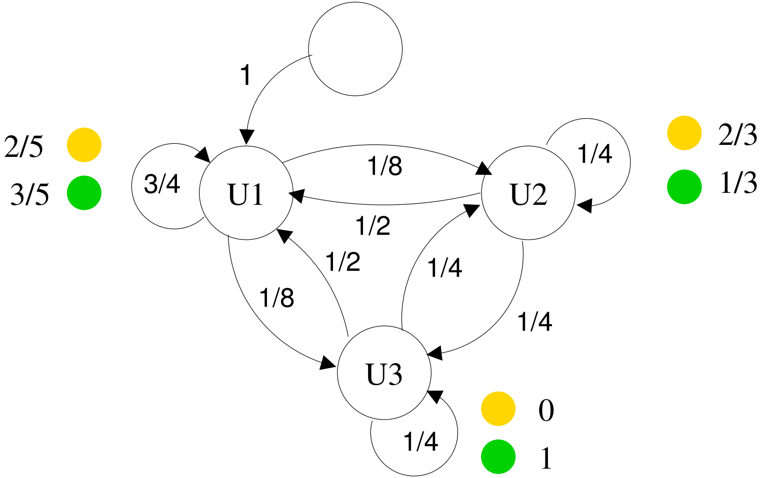
Un génie dispose de trois urnes, remplies respectivement de

- urne 1: 2 boules oranges et 3 boules vertes
- urne 2: 2 boules oranges et 1 boule verte
- urne 3: 3 boules vertes

A chaque étape, le génie tire une boule au hasard dans une des trois urnes, et montre la boule, sans dire de quelle urne elle provient. Pour le choix de l'urne, il a préférence pour l'urne 1, qu'il choisit systématiquement au départ puis 1 fois sur 2 quand il est sur l'urne 2 ou 3, et 3 fois sur 4 quand il est déjà sur l'urne 1. Les deux autres urnes, 2 et 3, sont choisies de manière équiprobables quand on est sur l'urne 1, 2 ou 3. L'objectif est de retrouver la suite des urnes pour une succession de boules donnée.

Question 1. HMM

On a trois états: un par urne.

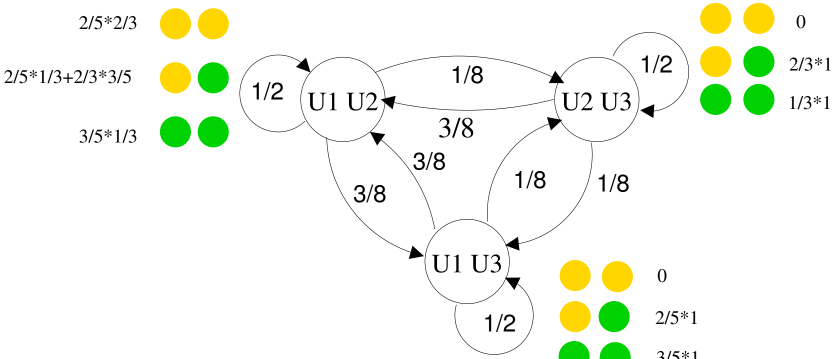


Question 2. Pour la suite d'observations `orange vert vert orange`, quelle est la suite d'états la plus probable parmi les quatre suites

- $s_1 = 1123$
- $s_2 = 2113$
- $s_3 = 1132$
- $s_4 = 1322$

Question 3. Le génie change de règle du jeu: à chaque tirage, il choisit deux urnes, et deux boules (une dans chaque urne). Le choix des deux urnes est donné par les règles suivantes:

- on reste sur le même couple d'urnes avec une probabilité de $1/2$;
- quand on change de couple d'urnes, on garde l'urne de plus petit numéro avec une probabilité de $3/4$.



2/5*2/3
2/5*1/3+2/3*3/5
3/5*1/3
U1 U2
U2 U3
U1 U3
1/2
1/8
3/8
3/8
3/8
1/8
1/8
1/2
1/2
0
2/3*1
1/3*1
0
2/5*1
3/5*1