Contrôle continu sur les chaînes de Markov

Exercice 1.
La Drosophile est une mouche très utilisée en génétique. Sa facilité d'élevage et son cycle de vie court permettent d'obtenir sans trop de difficultés un grand nombre de générations. Les yeux d'une drosophile peuvent être de couleur rouge ou pourpre. La couleur des yeux d'une drosophile est un caractère héréditaire lié à la présence dans son patrimoine génétique de deux copies d'un gène (une sur chaque chromosome), chacune étant de deux types possibles. Notons C et c ces deux allèles : C code pour la couleur rouge tandis que c code pour la couleur pourpre. Il y a alors trois génotypes possibles pour chaque individu : CC, Cc et cc. L'allèle C est dominant et c est récessif. Ainsi, les génotypes CC et Cc correspondent au phénotype [C] tandis que seul le génotype cc correspond au phénotype [c].

1. On réalise l'expérience suivante. On choisit au hasard une drosophile et on note X_0 le génotype de cette drosophile. On la croise avec un individu hétérozygote (i.e. de génotype Cc). On choisit au hasard un individu de la descendance obtenue et on note X_1 le génotype de cette drosophile. Puis on réitère l'expérience : on croise la dernière drosophile sélectionnée avec un individu hétérozygote, on choisit au hasard un individu de la descendance obtenue et on note X_2 le génotype de cette drosophile, etc... On note X_n le génotype du n$^{\text{è}}$ individu obtenu.

a) Pour tous états i et j dans $E = \{CC, Cc, cc\}$, déterminer $p_{i,j} = \mathbb{P}[X_1 = j | X_0 = i]$.

b) La suite $(X_n)_{n\geq 0}$ est en fait une chaine de Markov homogène à espace d'états $E = \{CC, Cc, cc\}$ de matrice de transition $\mathcal{P} = (p_{i,j})_{i,j\in E}$. Tracer le graphe des états associé à $\mathcal{P}$ et déterminer la nature des différents états.

c) Montrer qu'il existe une unique loi stationnaire notée π puis déterminer son expression.

d) Pour tout i dans E, calculer $\mathbb{E}[T_i | X_0 = i]$ où $T_i = \inf\{n \geq 1, X_n = i\}$.

e) Déterminer $p^{(n)}_{CC,CC}$. En déduire que la chaine est apériodique. Que vaut $\lim\limits_{n\to\infty} \mathcal{P}^{(n)}$?

2. On réalise l'expérience suivante. On choisit au hasard une drosophile et on note X_0 le génotype de cette drosophile. On la croise avec un individu homozygote dominant (i.e. de génotype CC). On choisit au hasard un individu de la descendance obtenue et on note X_1 le génotype de cette drosophile. Puis on réitère l'expérience : on croise la dernière drosophile sélectionnée avec un individu homozygote dominant, on choisit au hasard un individu de la descendance obtenue et on note X_2 le génotype de cette drosophile, etc... On note X_n le génotype du n$^{\text{è}}$ individu obtenu.

a) Pour tous états i et j dans $E = \{CC, Cc, cc\}$, déterminer $p_{i,j} = \mathbb{P}[X_1 = j | X_0 = i]$.

b) La suite $(X_n)_{n\geq 0}$ la chaine de Markov homogène à espace d'états $E = \{CC, Cc, cc\}$ de matrice de transition $\mathcal{P} = (p_{i,j})_{i,j\in E}$. Tracer le graphe des états associé à $\mathcal{P}$ et déterminer la nature des différents états.

c) Montrer que, bien que le théorème donnant l'existance et l'unicité d'une loi stationnaire ne s'applique pas, il existe une unique loi stationnaire notée π puis déterminer son expression.

d) Pour tout i dans E, on note $T_i = \inf\{n \geq 1, X_n = i\}$. Calculer $\mathbb{E}[T_{CC} | X_0 = CC]$. Pourquoi a-t-on $\mathbb{E}[T_{cc} | X_0 = cc] = \infty$?

e) Calculer $\mathcal{P}^{(n)}$. En déduire $\lim\limits_{n\to\infty} \mathcal{P}^{(n)}$.

Exercice 2.
Soit $(X_n)_{n\geq 0}$ la chaine de Markov homogène à espace d'états $E = \{1,2,3,4\}$ de matrice de transition $\mathcal{P} = \begin{pmatrix} 1/4 & 3/4 & 0 & 0 \\ 0 & 0 & 1/2 & 1/2 \\ 1/3 & 2/3 & 0 & 0 \\ 0 & 0 & 1/4 & 3/4 \end{pmatrix}$.

1. Calculer $\mathbb{P}[X_1 = 2|X_0 = 3, X_2 = 3]$ et $\mathbb{P}[X_1 = X_2]$ en fonction des coefficients de la matrice $\mathcal{P}$ et de la loi initiale de la chaine notée $p^{(0)}$.
2. Tracer le graphe des états associé à $\mathcal{P}$ et déterminer la nature des différents états.
3. Montrer qu'il existe une unique loi stationnaire notée π puis déterminer son expression.
4. Pour tout i dans E, on note $T_i = \inf\{n \geq 1, X_n = i\}$. Calculer $\mathbb{E}[T_i|X_0 = i]$ pour tout i dans E.
5. Déterminer les limites presque-sûres de $\displaystyle\frac{1}{n}\sum_{k=0}^{n-1} X_k$ et de $\displaystyle\frac{1}{n}\sum_{k=0}^{n-1} X_k^2$.

Exercice 3.
Soit $(X_n)_{n\geq 0}$ la chaine de Markov homogène à espace d'états $E = \{1,2,3,4,5,6\}$ de matrice de transition $\mathcal{P}$ dont les termes hors-diagonaux sont donnés par : $\mathcal{P} = \begin{pmatrix} . & 1/4 & 1/3 & 0 & 0 & 0 \\ 1/4 & . & 0 & 1/4 & 1/3 & 0 \\ 1/2 & 0 & . & 0 & 0 & 0 \\ 0 & 0 & 0 & . & 1/2 & 1/3 \\ 0 & 0 & 0 & 1/2 & . & 1/2 \\ 0 & 0 & 0 & 1/3 & 1/4 & . \end{pmatrix}$.

1. Déterminer les termes diagonaux de $\mathcal{P}$.
2. Tracer le graphe des états associé à $\mathcal{P}$ et déterminer la nature des différents états.
3. Pourquoi cela a-t-il un sens de considérer la chaine restreinte à $\{4,5,6\}$? Montrer que la chaine (X_n) restreinte à $\{4,5,6\}$ admet une unique loi invariante notée π et déterminer son expression.
4. Pour i dans E, on pose $T_i = \inf\{n \geq 0 : X_n = i\}$. Déduire de ce qui précède $\mathbb{E}[T_i|X_0 = i]$ pour i dans $\{4,5,6\}$ ainsi que la limite presque-sûre de $\displaystyle\frac{1}{n}\sum_{k=0}^{n-1} X_k$ conditionnellement à l'évènement $\{X_0 \in \{4,5,6\}\}$.
5. Déterminer $\displaystyle\lim_{n\to\infty} \frac{1}{n}\sum_{k=0}^{n-1} I(X_k = i)$ pour tout i dans E (bien que le théorème donnant cette limite ne s'applique pas directement).

Exercice 4.
L'évolution des nucléotides de l'ADN au fil des générations peut se modéliser comme une chaîne de Markov homogène $(X_n)_{n\geq 0}$ à espace d'états $E = \{A, C, G, T\}$ de matrice de transition $\mathcal{P}$. Les nucléotides A et G sont des purines tandis que les nucléotides C et T sont des pyrimidines. Les purines et les pyrimidines ont des structures moléculaires différentes aussi on peut s'attendre à ce que la probabilité de substitution d'une purine par une purine diffère de la probabilité de substitution d'une pyrimidine par une pyrimidine, diffère aussi de la probabilité de substitution d'une purine par une pyrimidine et diffère de la probabilité de substitution d'une pyrimidine

par une purine. Aussi, on spécifie ainsi la matrice de transition $\mathcal{P}$ de la façon suivante :

$$\mathcal{P} = \begin{pmatrix} \alpha & a & \alpha & a \\ b & \beta & b & \beta \\ \alpha & a & \alpha & a \\ b & \beta & b & \beta \end{pmatrix}$$

où a, b, α et β appartiennent à $]0, 1[$.

1. Quelles relations doivent satisfaire a, b, α et β pour que la matrice $\mathcal{P}$ soit stochastique ? Dans toute la suite, on supposera ces relations satisfaites.

2. Tracer le graphe des états associé à $\mathcal{P}$ et déterminer la nature des différents états.

3. Montrer qu'il existe une unique loi stationnaire notée π.

4. Dans la suite, on fait l'hypothèse que la loi initiale $p^{(0)}$ de la chaîne (X_n) coïncide avec la probabilité stationnaire π. Montrer que la loi de X_n est π pour tout $n \in \mathbb{N}$.

5. Notons p un sous-ensemble des paramètres de la matrice de transition choisis pour être non liés, par exemple

$$p = (a, b)$$

Pour toute suite de réels $(i_0, ..., i_n)$ à valeurs dans E^{n+1}, on appelle fonction de log-vraisemblance la fonction L_p définie par

$$L_p(i_0, ..., i_n) = \log\Big(\mathbb{P}[X_0 = i_0, ..., X_n = i_n]\Big)$$

Exprimer $L_p(i_0, ..., i_n)$ en fonction de p (et de π).

6. Déterminer la valeur de p notée $\widehat{p}(i_0, ..., i_n)$ qui maximise $L_p(i_0, ..., i_n)$. On admettra que la matrice hessienne de L_p est définie négative en la solution annulant le vecteur gradient de L_p.

7. Déterminer la limite presque-sûre de $\widehat{p}(X_0, ..., X_n)$. On utilisera le fait que $(Y_n)_{n\in\mathbb{N}}$ où $Y_n = (X_n, X_{n+1})$ est une chaine de Markov à valeurs dans $E \times E$ irréductible récurrente positive d'unique probabilité invariante $\widetilde{\pi} = (\widetilde{\pi}_{i,j} = \pi_i p_{i,j})_{i,j\in E}$.

8. Considérons l'évolution suivante :

$$AGAAAAGAGGGAAATAGGCGGCGGCCCCTTCTTTCTTTCCC$$

Proposer une "bonne" approximation de la matrice de transition.