Contrôle de Modèles et algorithmes stochastiques 1 et 2

Durée 1h30

Les documents, les calculatrices et les téléphones portables sont interdits.
Le barème sur 20 est approximatif.

Exercice 1 (Variante du modèle de Moran avec sélection) (12 points)
On considère une population de taille constante $N \geq 2$, pour laquelle la transmission des gènes d'une génération à une autre se fait suivant la version suivante du modèle de Moran.
À chaque génération, deux individus de la population sont choisis au hasard, sans remise. Le premier individu choisi n'a aucun descendant, le second individu choisi a deux descendants, et tous les autres individus de la population ont un seul descendant dans la génération suivante.

1. On considère un gène qui peut s'exprimer sous la forme de deux allèles différentes, A et a. Soit $r > 1$. On suppose que l'allèle A a un avantage sélectif sur l'allèle a, dans le sens suivant : à chaque génération, quand on choisit le premier individu, destiné à ne pas avoir de descendant, chaque individu portant l'allèle a (respectivement A) a une probabilité proportionnelle à r (respectivement 1) d'être choisi.
   Pour $n \geq 0$, soit X_n le nombre d'individus portant l'allèle A dans la population à la génération n. On note $P = (p_{ij})_{0 \leq i,j \leq N}$ la matrice de transition de la chaîne de Markov $(X_n)_{n \geq 0}$.
   (a) Donner p_{0j} et p_{Nj}, pour $j = 0, \ldots, N$.
   (b) Pour tout $1 \leq i \leq N-1$, donner $p_{i,i-1}, p_{ii}$ et $p_{i,i+1}$ et vérifier que $p_{i,i-1}+p_{ii}+p_{i,i+1} = 1$.
   (c) Vérifier si $(X_n)_{n\geq 0}$ est une martingale/sur-martingale/sous-martingale par rapport à la filtration naturelle.
   (d) Montrer que $(X_n)_n$ converge presque surement, quand $n \to \infty$, vers une variable aléatoire X_∞, en précisant les valeurs possibles pour X_∞.

   Pour $i = 0, \ldots, N$, on s'intéresse à $f_i := \mathbb{P}(X_\infty = 0 | X_0 = i)$, la probabilité de fixation de l'allèle a sachant que dans la population initiale i individus portent l'allèle A.
   (e) Donner f_0 et f_N. Pour $i = 1, ..., N-1$, exprimer f_i en fonction de f_{i-1} et de f_{i+1} et montrer que
   $$f_{i+1} - f_i = r^{-1}(f_i - f_{i-1}).$$
   (f) En déduire que pour tout $i = 0, ..., N$ on a
   $$f_i = \frac{r^{-i} - r^{-N}}{1 - r^{-N}}.$$
   (g) Donner $\mathbb{E}(X_\infty | X_0 = i)$. Déterminer ensuite $\mathbb{E}(X_\infty)$ si la loi de X_0 est Binomiale$(N, 1/2)$.

2. On s'intéresse maintenant à l'arbre généalogique d'un échantillon de $2 \leq n \leq N$ individus de la population présente. On appelle T'_n le nombre de générations qu'il faut rémonter en arrière jusqu'à ce que deux individus parmi les n trouvent leur ancêtre commun.
   (a) Donner la loi de T'_n.
   (b) Déterminer le nombre moyen de générations qu'il faut rémonter en arrière pour trouver le plus récent ancêtre commun de tous les n individus de l'échantillon.

Exercice 2 (Mouvement d'une particule dans $\mathbb{Z}$) (8 points)
On considère une particule qui se déplace dans $\mathbb{Z}$ suivant une marche aléatoire symétrique simple, et on note X_n sa position à l'instant n, pour tout $n \geq 1$. Plus précisément, on se donne une suite de variables aléatoires $(U_n)_{n\geq 1}$ i.i.d. de loi $\mathbb{P}(U_n = -1) = \mathbb{P}(U_n = 1) = 1/2$ et pour tout $n \geq 1$ on pose $X_n := \sum_{i=1}^n U_i$.

On suppose que l'on dispose, pour tout $n \geq 1$, d'une observation bruitée Y_n de la position de la particule à l'instant n, avec $Y_n = X_n + \varepsilon_n$, où $(\varepsilon_n)_{n\geq 1}$ est une suite de v.a. i.i.d. à valeurs dans $\{-1, 0, 1\}$ de loi donnée par $\mathbb{P}(\varepsilon_n = 0) = p$ et $\mathbb{P}(\varepsilon_n = -1) = \mathbb{P}(\varepsilon_n = 1) = \frac{1}{2}(1-p)$, avec $p \in]0, 1[$ un paramètre. On suppose aussi que les suites $(U_n)_{n\geq 1}$ et $(\varepsilon_n)_{n\geq 1}$ sont indépendantes.

1. Montrer que $(X_n, Y_n)_{n\geq 1}$ est une chaîne de Markov cachée sur $\mathbb{Z} \times \mathbb{Z}$ et donner la loi initiale $\nu = \mathcal{L}(X_1)$, les probabilités de transition $P(x, x') = \mathbb{P}(X_{n+1} = x'|X_n = x)$, ainsi que les probabilités d'observation $Q(x, y) = \mathbb{P}(Y_n = y|X_n = x)$, pour $x, x', y \in \mathbb{Z}$ et $n \geq 1$.

   Pour tout $n \geq 1$, $x \in \mathbb{Z}$ et $y_{1:n} \in \mathbb{Z}^n$, on considère les probabilités de prédiction et filtration :

$$P_{n+1}(x; y_{1:n}) := \mathbb{P}(X_{n+1} = x|Y_{1:n} = y_{1:n}) \text{ et } F_n(x; y_{1:n}) := \mathbb{P}(X_n = x|Y_{1:n} = y_{1:n}).$$

2. Exprimer $P_{n+1}(x; y_{1:n})$ en fonction de $F_n(x-1; y_{1:n})$ et de $F_n(x+1; y_{1:n})$.
3. Soit $n \geq 2$. Montrer que pour $|x - y_n| \leq 1$ on a

$$F_n(x; y_{1:n}) = \frac{Q(x, y_n)P_n(x; y_{1:n-1})}{\sum_{z=y_n-1}^{y_n+1} Q(z, y_n)P_n(z; y_{1:n-1})}.$$

4. Exprimer $F_n(y_n; y_{1:n})$ en fonction de p et de $P_n(y_n; y_{1:n-1})$.
5. On suppose dans la suite que le paramètre p est inconnu et on s'intéresse à l'estimer à partir des observations $Y_{1:n}$, en utilisant l'algorithme E-M.
   Pour $p \in]0, 1[$ et $x_{1:n}, y_{1:n} \in \mathbb{Z}^n$, soit $L(p; x_{1:n}, y_{1,n}) := \log(\mathbb{P}_p(X_{1:n} = x_{1:n}, Y_{1:n} = y_{1:n}))$ la log-vraisemblance complète sous le paramètre p.
   Pour $p_0 \in]0, 1[$ donné, on définit la quantité

$$Q(p|p_0) := \mathbb{E}_{p_0}[L(p; X_{1:n}, y_{1:n}) \mid Y_{1:n} = y_{1:n}], \text{ pour tout } p \in]0, 1[.$$

   Montrer que $\hat{p}$ maximise la fonction $p \mapsto Q(p|p_0)$ si et seulement si $\hat{p}$ maximise la quantité

$$f(p) = \log(p) \sum_{k=1}^n \mathbb{P}_{p_0}(X_k = y_k|Y_{1:n} = y_{1:n}) + \log\left(\frac{1-p}{2}\right)\left(n - \sum_{k=1}^n \mathbb{P}_{p_0}(X_k = y_k|Y_{1:n} = y_{1:n})\right).$$

6. En déduire le paramètre $\hat{p}$ qui maximise $Q(p|p_0)$, en l'exprimant en fonction des probabilités de lissage $\mathbb{P}_{p_0}(X_k = y_k|Y_{1:n} = y_{1:n}),\ k = 1, \dots, n$.