**Examen du 5 mai 2017**

**Durée : 2 heures.**

**Exercice 1.** Soit $(\mathbb{t}, d)$ un arbre réel enraciné en $\rho \in \mathbb{t}$. On se propose ici de montrer quelques propriétés fondamentales des arbres réels qui avaient été admises en cours.

a) Rappeler les deux propriétés qui caractérisent les arbres réels et en particulier la définition de l'arc $[\![x, y]\!]$ et celle de l'isométrie $\phi_{x,y}$.

b) On définit la relation $x \prec y$ par

$$x \prec y \iff x \in [\![\rho, y]\!] \qquad x, y \in \mathbb{t}.$$

i) Montrer que pour tous $x, y \in \mathbb{t}$,

$$x \prec y \iff [\![\rho, x]\!] \subset [\![\rho, y]\!].$$

On pourra utiliser l'application $t \mapsto \phi_{\rho,y}(d(\rho, x)\, t)$.

ii) Montrer que $\prec$ est antisymétrique, c'est-à-dire que si $x \prec y$ et $y \prec x$ alors $x = y$.

iii) Montrer que $\prec$ est transitive.

c) Soient $x, y \in \mathbb{t}$ et $A := [\![\rho, x]\!] \cap [\![\rho, y]\!]$. Soit $\delta := \sup_{z \in A} d(\rho, z)$. On cherche à montrer qu'il existe un unique $z \in \mathbb{t}$ (habituellement noté $z = x \wedge y$) tel que $A = [\![\rho, z]\!]$, et que $d(\rho, z) = \delta$.

i) On définit $z := \phi_{\rho,x}(\delta)$. Montrer que $z = \phi_{\rho,y}(\delta)$.

ii) Montrer que $[\![\rho, z]\!] \subset A$.

iii) Démontrer l'inclusion opposée et conclure.

iv) Montrer que $\wedge$ est associatif.

*Solution de l'exercice* 1.

a) Un arbre réel $(\mathbb{t}, d)$ est un espace métrique complet tel satisfaisant les deux propriétés suivantes :
(A) Unicité des géodésiques. Pour tous $x, y \in \mathbb{t}$, il existe une unique isométrie $\phi_{x,y} : [0, d(x, y)] \to \mathbb{t}$ telle que $\phi_{x,y}(0) = 0$ et $\phi_{x,y}(d(x, y)) = y$.
L'arc $[\![x, y]\!]$ est la géodésique $\phi_{x,y}([0, d(x, y)])$.
(B) Absence de cycles. Pour toute application continue injective $\psi : [0, 1] \to \mathbb{t}$, on a $\psi([0, 1]) = [\![\psi(0), \psi(1)]\!]$.

b) Relation $\prec$.

i) Soit $x \in [\![\rho, y]\!]$ et $\psi : t \mapsto \phi_{\rho,y}(d(\rho, x)\, t)$. Alors $\psi$ est continue et injective avec $\psi(0) = \rho$ et $\psi(1) = x$. En effet, $x \in [\![\rho, y]\!]$ et $\phi_{\rho,y}$ est une isométrie donc $\phi^{-1}_{\rho,y}(x) = d(\rho, x)$. Par l'absence de cycle, $\psi([0, 1]) = [\![\rho, x]\!]$, or $\psi([0, 1]) = \phi_{\rho,y}([0, d(\rho, x)]) \subset [\![\rho, y]\!]$, d'où l'implication. L'implication réciproque est évidente.

ii) Si $x \prec y$ et $y \prec x$ alors d'après la question précédente, $[\![\rho, x]\!] \subset [\![\rho, y]\!]$ et $[\![\rho, y]\!] \subset [\![\rho, x]\!]$, donc $[\![\rho, x]\!] = [\![\rho, y]\!]$, ce qui implique $x = y$. En effet, cette égalité implique l'égalité $d(\rho, x) = d(\rho, y)$ puis par unicité des géodésiques, $\phi_{\rho,x} = \phi_{\rho,y}$. En conclusion, $x = \phi_{\rho,x}(d(\rho, x)) = \phi_{\rho,y}(d(\rho, y)) = y$.

iii) La transitivité de $\prec$ résulte de la transitivité de $\subset$.

c) Loi de composition $\wedge$. Soit $\delta := \sup_{z \in A} d(\rho, z)$ où $A = [\![\rho, x]\!] \cap [\![\rho, y]\!]$.

i) Soit $z := \phi_{\rho,x}(\delta)$. Par définition de $\delta$, il existe une suite $(z_n)$ de $A$ telle que $\delta_n := d(\rho, z_n)$ converge vers $\delta$. Comme $z_n \in [\![\rho, x]\!]$, par unicité de la géodésique, $z_n = \phi_{\rho,x}(\delta_n)$ et symétriquement, $z_n = \phi_{\rho,y}(\delta_n)$. Par continuité de ces deux isométries,

$$z = \phi_{\rho,x}(\delta) = \lim_{n\to\infty} \phi_{\rho,x}(\delta_n) = \lim_{n\to\infty} \phi_{\rho,y}(\delta_n) = \phi_{\rho,y}(\delta).$$

ii) Un arc est fermé comme image d'un fermé par une application continue, donc $A$ est fermé comme intersection de deux fermés. Or $z$ est limite de la suite $(z_n)$ à valeurs dans le fermé $A$, donc $z \in A$. Par la question b), on a donc $[\![\rho, z]\!] \subset A$.

iii) Soit $z' \in A$. Montrons que $z' \in [\![\rho, z]\!]$. Par définition de $\delta$, $\delta' := d(\rho, z') \leq \delta$. On sait également que $z' = \phi_{\rho,x}(\delta') = \phi_{\rho,y}(\delta')$. Soit $\psi : t \mapsto \phi_{\rho,x}(t\delta)$. Alors $\psi$ est continue, injective, avec $\psi(0) = \rho$ et $\psi(1) = z$. Donc par absence de cycle, $\psi([0, 1]) = [\![\rho, z]\!]$, or $z' \in \psi([0, 1])$ puisque $z' = \psi(\delta'/\delta)$, donc $z' \in [\![\rho, z]\!]$.
On a donc exhibé $z \in \mathbb{t}$ tel que $d(\rho, z) = \delta$ et par double inclusion, $[\![\rho, z]\!] = [\![\rho, x]\!] \cap [\![\rho, y]\!]$. L'unicité vient du fait que si $[\![\rho, z]\!] = [\![\rho, z']\!]$, alors $z = z'$ (voir question b).

iv) La loi de composition $\wedge$ est associative par associativité de $\cap$.

**Exercice 2.** Soit $(\mathbb{t}, d)$ un arbre ultramétrique à $n$ feuilles, sur lequel on superpose un nuage poissonnien de mutations ponctuelles, dont l'intensité est $\frac{\theta}{2} > 0$ par rapport à la mesure de Lebesgue sur $\mathbb{t}$. On fait l'hypothèse usuelle du **modèle à une infinité d'allèles**, c'est-à-dire que chaque feuille de l'arbre hérite le type, dit **allèle**, de la mutation la plus récente sur sa lignée. On définit $A_n(i)$ comme le **nombre de mutations dont le type est porté par exactement $i$ feuilles parmi les $n$**.
On échantillonne $k$ feuilles parmi les $n$ uniformément au hasard et l'on définit $p_{k,n}$ comme la **probabilité que les $k$ feuilles échantillonnées portent toutes le même allèle**.

a) Représenter graphiquement les objets considérés et donner une relation entre $n$ et les $A_n(i)$.

b) Montrer que

$$p_{k,n} = \mathbb{E}\left(e^{-\frac{\theta}{2} L_{k,n}}\right) = \sum_{i=k}^{n} \frac{g_k(i)}{g_k(n)} \mathbb{E}(A_n(i)),$$

où $L_{k,n}$ est une v.a. que l'on définira et $g_k(N) = N(N-1)\cdots(N-k+1)$ pour tout entier positif $N$.

c) On suppose dans cette question que l'arbre $(\mathbb{t}, d)$ est donné par un **coalescent de Kingman** issu de $n$ feuilles.

i) Expliquer pourquoi $p_{k,n} \equiv p_k$ ne dépend pas de $n$.

ii) Montrer l'égalité

$$p_k = \prod_{j=2}^{k} \frac{j-1}{\theta + j - 1}$$

iii) Soit $X_n(j)$ l'abondance (nombre de feuilles le portant) du $j$-ème allèle le plus abondant. On rappelle la convergence en loi lorsque $n \to \infty$ de la suite $\left(\frac{X_n(j)}{n}; 1 \leq j \leq n\right)$ vers une suite de v.a. $(X(j); j \geq 1)$.

iv) Démontrer que $\sum_j X(j) \leq 1$.

v) Montrer que

$$p_k = \sum_{j=1}^{n} \mathbb{E}\left(\frac{g_k(X_n(j))}{g_k(n)}\right)$$

vi) On admettra que $\sum_j X(j) = 1$. Démontrer que $p_k = \sum_{j\geq 1} \mathbb{E}\left(X(j)^k\right)$.
On pourra commencer par montrer que pour tout $\varepsilon > 0$, il existe $J \in \mathbb{N}$ tel que

$$\limsup_n \sum_{j\geq J} \mathbb{E}\left(\frac{g_k(X_n(j))}{g_k(n)}\right) \leq \varepsilon.$$

d) On suppose à présent que l'arbre $(\mathbb{t}, d)$ est donné par un **processus ponctuel de coalescence (CPP) de profondeur** $T$.

i) Montrer que

$$p_{n,n} = \mathbb{E}\left(e^{-\frac{\theta}{2}\left(\max_{j=1,\dots,n-1} H_j + \sum_{i=1}^{n-1} H_i\right)}\right),$$

où les $(H_i)$ sont des v.a. positives i.i.d. que l'on définira.

ii) En déduire l'égalité suivante, où l'on interprétera la variable d'intégration

$$p_{n,n} = (n-1)\int_0^T \mathbb{P}(H \in dt)\, e^{-\theta t} \left\{\mathbb{E}\left(e^{-\frac{\theta}{2}H}\mathbb{1}_{\{H<t\}}\right)\right\}^{n-2}.$$

iii) En déduire l'égalité suivante, où l'on interprétera la nouvelle variable d'intégration

$$p_{n,n} = e^{-\frac{\theta}{2}T}\left\{\mathbb{E}\left(e^{-\frac{\theta}{2}H}\right)\right\}^{n-1} + \int_0^T \frac{\theta}{2} dt\, e^{-\frac{\theta}{2}t} \left\{\mathbb{E}\left(e^{-\frac{\theta}{2}H}\mathbb{1}_{\{H<t\}}\right)\right\}^{n-2}.$$

iv) On n'a pas essayé de calculer $p_{k,n}$ dans le cas d'un CPP dont $k$ feuilles sont échantillonnées uniformément au hasard. Si on échantillonnait plutôt chaque feuille avec probabilité $q$ indépendamment, que pourrait-on faire ?

*Solution de l'exercice* 2.

a) Les allèles induisent une partition de l'ensemble des feuilles, donc $n = \sum_{i=1}^{n} iA_n(i)$.

b) Dans la première égalité, $L_{k,n}$ est simplement la mesure (longueur) de l'arbre généré par les $k$ feuilles échantillonnées (dont la racine est leur plus récent ancêtre commun). En effet $p_{k,n}$ est la probabilité que cet arbre ne porte aucune mutation, qui n'est autre que $\mathbb{E}\left(e^{-\frac{\theta}{2}L_{k,n}}\right)$.
Pour la deuxième égalité, il faut simplement remarquer que $p_{k,n}$ est aussi la probabilité d'échantillonner les $k$ feuilles dans un même bloc allélique. La probabilité d'échantillonner ces $k$ feuilles dans un bloc donné de taille $i$ vaut $\frac{g_k(i)}{g_k(n)}$, donc conditionnellement à l'arbre et à ses mutations, la probabilité que les $k$ feuilles portent le même allèle vaut $\sum_{i=k}^{n} \frac{g_k(i)}{g_k(n)} A_n(i)$. Le résultat vient en intégrant sur la loi de l'arbre avec ses mutations.

c) L'arbre $(\mathbb{t}, d)$ est maintenant donné par un coalescent de Kingman issu de $n$ feuilles.

i) Par cohérence du coalescent de Kingman par échantillonnage, l'arbre engendré par les $k$ feuilles échantillonnées uniformément parmi les $n$ a même loi qu'un coalescent de Kingman issu de $k$ feuilles. Par les propriétés classiques des mesures de Poisson, il en est de même pour l'arbre avec mutations poissonniennes. Ainsi $p_{k,n} \equiv p_k$ est simplement la probabilité qu'un coalescent de Kingman issu de $k$ ne porte aucune mutation.

ii) D'après ce qui précède, $p_k = \mathbb{E}\left(e^{-\frac{\theta}{2}L_k}\right)$, où $L_k$ est la longueur totale de l'arbre de Kingman issu de $k$ arrêté au dernier évènement de coalescence. Or

$$L_k = \sum_{j=2}^{k} jB_j,$$

où les $(B_j)$ sont indépendantes et $B_j$ suit la loi exponentielle de paramètre $j(j-1)/2$. On obtient donc

$$p_k = \mathbb{E}\left(e^{-\frac{\theta}{2}L_k}\right) = \mathbb{E}\left(e^{-\frac{\theta}{2}\sum_{j=2}^{k} jB_j}\right) = \prod_{j=2}^{k} \mathbb{E}\left(e^{-\frac{\theta}{2}jB_j}\right) = \prod_{j=2}^{k} \frac{j-1}{\theta+j-1}$$

Soit $X_n(j)$ l'abondance (nombre de feuilles le portant) du $j$-ème allèle le plus abondant.

iii) Le lemme de Fatou assure que $\sum_j X(j) = \sum_j \liminf \frac{X_n(j)}{n} \leq \liminf \sum_j \frac{X_n(j)}{n} = 1$, où l'on a implicitement plongé les $X_n(j)$ dans un même espace de probabilité (comme il est possible, d'après le théorème de plongement de Skorohod).

iv) Le raisonnement est le même que pour la question b).

v) Par le théorème de Fubini, $\sum_j \mathbb{E}(X_j) = 1$ et $\sum_j \mathbb{E}\left(\frac{X_n(j)}{n}\right) = 1$. On peut donc intervertir sommation et passage à la limite

$$\lim_{n\to\infty} \sum_{j\geq 1} \mathbb{E}\left(\frac{X_n(j)}{n}\right) = 1 = \sum_{j\geq 1} \lim_{n\to\infty} \mathbb{E}\left(\frac{X_n(j)}{n}\right),$$

où l'on a utilisé que

$$\lim_{n\to\infty} \mathbb{E}\left(\frac{X_n(j)}{n}\right) = \mathbb{E}(X(j)),$$

par convergence dominée. Comme on peut intervertir sommation et passage à la limite pour les sommes finies, on peut le faire pour les restes de la somme totale. Ainsi, choisissant $J \in \mathbb{N}$ tel que $\sum_{j\geq J} \mathbb{E}(X(j)) < \varepsilon$, on a

$$\lim_{n\to\infty} \sum_{j\geq J} \mathbb{E}\left(\frac{X_n(j)}{n}\right) = \sum_{j\geq J} \lim_{n\to\infty} \mathbb{E}\left(\frac{X_n(j)}{n}\right) = \sum_{j\geq J} \mathbb{E}(X(j)) < \varepsilon.$$

Comme $\frac{g_k(X_n(j))}{g_k(n)} \leq \frac{X_n(j)}{n}$, on a

$$\limsup_n \sum_{j\geq J} \mathbb{E}\left(\frac{g_k(X_n(j))}{g_k(n)}\right) \leq \limsup_n \sum_{j\geq J} \mathbb{E}\left(\frac{X_n(j)}{n}\right) \leq \varepsilon.$$

Le reste de la démonstration consiste à écrire pour tout $n \geq J$,

$$\left| p_k - \sum_{j\geq 1} \mathbb{E}\left(X(j)^k\right) \right| \leq \sum_{j\geq J} \mathbb{E}\left(\frac{g_k(X_n(j))}{g_k(n)}\right) + \left| \sum_{j=1}^{J-1} \left\{ \mathbb{E}\left(\frac{g_k(X_n(j))}{g_k(n)}\right) - \mathbb{E}\left(X(j)^k\right) \right\} \right| + \sum_{j\geq J} \mathbb{E}\left(X(j)^k\right),$$

et à utiliser le théorème de convergence dominée pour montrer que le terme central tend vers 0.

d) L'arbre $(\mathbb{t}, d)$ est maintenant donné par un processus ponctuel de coalescence (CPP) de profondeur $T$.

i) Utiliser la question b) et se servir du fait que $L_{n,n} = \max_{j=1,\dots,n-1} H_j + \sum_{i=1}^{n-1} H_i$, où les $H_i$ sont les $n-1$ profondeurs des nœuds de l'arbre, qui sont i.i.d. par définition du CPP. En particulier, ici le CPP étant conditionné par son nombre de feuilles, les $H_i$ sont les profondeurs usuelles conditionnées à être plus petites que $T$.

ii) Soient $J := \arg\max H_i$ et $H^\star := \max H_i$. D'après la question précédente, en intégrant par rapport à la loi de $(J, H^\star)$, on obtient

$$\begin{aligned} p_{n,n} &= \sum_{j=1}^{n-1} \int_0^T \mathbb{E}\left( e^{-\frac{\theta}{2}\left(H^\star + \sum_{i=1}^{n-1} H_i\right)} \mathbb{1}_{\{H^\star \in dt, J=j\}} \right) \\ &= \sum_{j=1}^{n-1} \int_0^T \mathbb{E}\left( e^{-\theta H_j} e^{-\frac{\theta}{2}\sum_{i\neq j} H_i} \mathbb{1}_{\{H_j \in dt\}}\, \mathbb{1}_{\{H_i < t, i\neq j\}} \right) \\ &= \sum_{j=1}^{n-1} \int_0^T \mathbb{P}(H \in dt)\, e^{-\theta t} \left\{ \mathbb{E}\left( e^{-\frac{\theta}{2}H} \mathbb{1}_{\{H<t\}} \right) \right\}^{n-2} \\ &= (n-1) \int_0^T \mathbb{P}(H \in dt)\, e^{-\theta t} \left\{ \mathbb{E}\left( e^{-\frac{\theta}{2}H} \mathbb{1}_{\{H<t\}} \right) \right\}^{n-2}. \end{aligned}$$

iii) L'égalité se déduit d'une intégration par parties et la nouvelle variable d'intégration est l'âge de la mutation la plus récente. Le premier terme correspond aux cas où il n'y a aucune mutation sur tout l'arbre.

iv) Lorsqu'on échantillonne chaque feuille avec probabilité $q$ indépendamment, l'arbre engendré par les feuilles échantillonnées est un CPP dont la profondeur $H'$ est donnée par $H' = \max_{j=1,\dots,G} H_j$ où $G$ est une variable géométrique de probabilité de succès $q$.