

La vitesse des coalescents

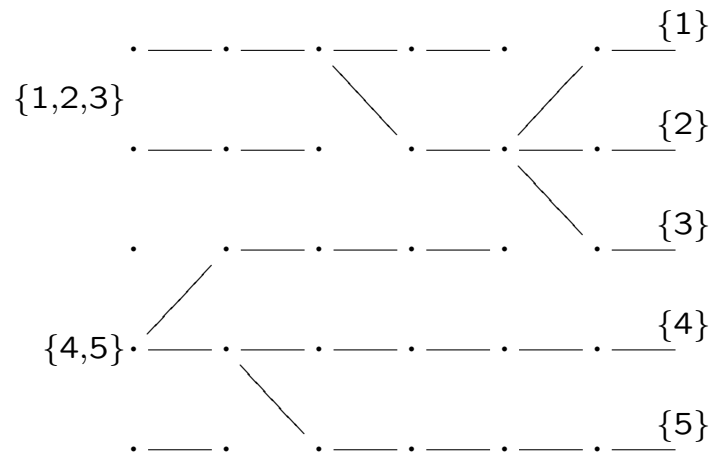
Julien Berestycki

Université Paris VI

Avec : Nathanaël Berestycki et Vlada Limic

Généalogies et modèle de cannings

- ▶ Taille de la population = N , générations discrètes.
- ▶ Pour une génération donnée g : $\nu_i^{(g)}$ = # enfants de indiv. i dans gén. $g + 1$. $\sum \nu_i^{(g)} = N$ et $\nu^{(g)}$ est échangeable.
- ▶ vecteurs $\nu^{(g)}$ iid



$R_r^{(N)}$ = partition de $\{1, \dots, N\}$ $i \sim j$ ssi i et j ont même ancêtre dans gén. $-r$. Ici $R_6^{(5)} = \{1, 2, 3\}, \{4, 5\}$

► Exemple : Wright-Fisher : Chaque individu choisit son parent dans la génération précédente uniformément. $P(\nu_1 = k_1, \dots, \nu_N = k_N) = \binom{N}{k_1, \dots, k_N} N^{-N}$.

► Dans ce cas $c_N := P(\text{deux indiv. ont même parent}) = 1/N$ et quand $N \rightarrow \infty$, on a $(R_{[t/c_N]}^{(N)})_{t \geq 0} \rightarrow (R_t)_{t \geq 0}$ (marg. fini-dim.) où $R = \text{coal. de Kingman}$.

Theorem (Kingman, Möhle-Sagitov) Reste vrai ssi $c_N \rightarrow 0$ et $d_N/c_N \rightarrow 0$ qd $N \rightarrow \infty$ où $d_N = \frac{\binom{N}{1}}{\binom{N}{3}} E[(\nu_1)_3]$.

Le coalescent de Kingman (1982)

On commence avec le n -coalescent

- ▶ Processus de Markov à valeur dans \mathcal{P}_n partitions de $[n] = \{1, \dots, n\}$.
- ▶ $\Pi_n(0) = \{1\}, \{2\} \dots, \{n\}$.
- ▶ Chaque paire de blocs fusionne à taux 1, pas d'autre transition.

► **Consistence.** on peut vérifier que $\Pi_{n+1} \big|_{[n]} =_d \Pi_n$.

► Ainsi on a un processus $\Pi(t) \in \mathcal{P} =$ partitions de $\{1, 2 \dots\}$ t.q.

$$\Pi(t) \big|_{[n]} = \Pi_n(t)$$

► C'est le **coalescent de Kingman**.

Descendre de l'infini

► Initialement :

$$N(0) = \infty, \text{ mais } \forall t > 0, N(t) < \infty$$

► En fait :

$$N_t \sim 2/t \text{ p.s. quand } t \rightarrow 0.$$

Heuristique : solution de $\frac{du}{dt} = -u^2/2$; $u(0) = \infty$ est $u(t) = 2/t$.

Une classe de modèle de Cannings naturelle

Chaque individu i produit un nombre d'enfants Z_i , les Z_i iid et $P(Z > k) \sim k^{-\alpha}, \alpha > 1$.

Pour garder pop = N on tire N indiv. sans remise parmi tous ceux produits.

- ▶ Si $\alpha \geq 2$ moment d'ordre 2 et la généalogie = Kingman
- ▶ (Schweinsberg 2003) Si $1 < \alpha < 2$ la généalogie \rightarrow beta-coalescents.

Λ -coalescents (Pitman 99, Sagitov 99)

- ▶ Processus de Markov Π en temps continu à valeurs dans \mathcal{P} qui démarrent de : $\Pi(0) = \{1\}, \{2\}, \dots$
- ▶ Quand b blocs, chaque k -tuple fusionne au taux $\lambda_{b,k}$. (Kingman : $\lambda_{b,k} = \mathbf{1}_{\{k=2\}}$.) Tous les $\lambda_{b,k}$ ne sont pas possibles (car on a besoin de la consistance des $\Pi_{|[n]}$).

Theorem (Pitman 99) $\exists \Lambda$ une mesure finie $[0, 1]$:

$$\lambda_{b,k} = \int_0^1 p^{k-2} (1-p)^{b-k} \Lambda(dp)$$

- ▶ $\Lambda = \delta_0$ correspond au Kingman. $\Lambda = U(0, 1)$ au Bolthausen-Sznitman.

Descendre de l'infini

► Soit $(\Pi_t, t \geq 0)$ un Λ -coalescent.

Theorem (Pitman 99) Supposon $\Lambda(\{1\}) = 0$. Avec probabilité 1,

$$\#\text{blocks of } \Pi_t \begin{cases} = \infty & \forall t > 0 \\ < \infty & \forall t > 0 \end{cases}$$

► Quand Π descend de l'infini, on atteint $N(t) = 1$, et $N(t) \rightarrow \infty$ quand $t \rightarrow 0^+$.

Question

- ▶ Que dire du comportement en temps petit des Λ -coalescents ?

On commence avec le résultat suivant. Soit $\gamma_b = \sum_{k=2}^b (k-1)\lambda_{b,k}$.

Theorem (Schweinsberg 2000) :

$$CD_\infty \iff \sum_{b=1}^{\infty} \gamma_b^{-1} < \infty$$

- ▶ Mais à quelle vitesse ? i.e. Comment $N(t)$ se comporte-t-il quand t petit ?

► Cas des **Beta-coalescents**.

$\Lambda = \text{Beta}(2 - \alpha, \alpha)$ $1 < \alpha < 2$.

$$\Lambda(dx) = \frac{1}{\Gamma(2 - \alpha)\Gamma(\alpha)} x^{1-\alpha} (1 - x)^{\alpha-1} dx$$

► CD_∞

(Berestycki-B.-Schweinsberg 06)

Theorem Soit $N_t = \#$ blocs dans un Beta-coalescent.

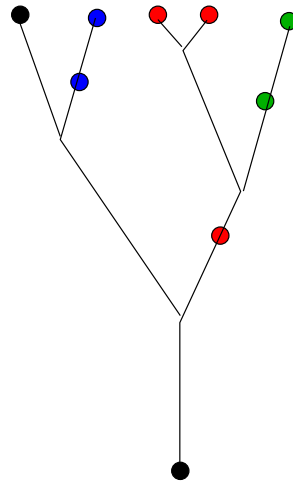
$$t^{\frac{1}{\alpha-1}} N_t \xrightarrow[t \rightarrow 0]{a.s.} (\alpha \Gamma(\alpha))^{\frac{-1}{\alpha-1}},$$

► Pour Λ général, pas de résultat jusqu'à présent.

Pourquoi est-ce intéressant ?

Des mutations surviennent à taux $\theta/2$ le long de chaque lignée.

- ▶ Les mutations arrivent le long d'une séquence ADN (...ATTGTCA...): chaque mutation affecte un nouveau site. Si la séquence code pour une couleur, chaque mutation crée une nouvelle couleur jamais vue.



Site de segregation: sites auxquels l'échantillon n'est pas homogène. (Exemple: $S_5 = 3$)

$N_n(k) = \#$ types portés par k indiv. (ici $N_5(1) = 3, N_5(2) = 1$).

Partition allélique $\Pi_n = \{\{1\}, \{2\}, \{3, 4\}, \{5\}\}$.

Si généalogie = Kingman

$$E[S_n] = \frac{\theta}{2} \sum_{b=2}^n b \binom{b}{2}^{-1} = \theta \sum_{b=2}^n \frac{1}{b-1} \sim \theta \log n.$$

Ewens sampling formula (ESF): si $\forall j, \pi$ a a_j blocs de taille j , alors

$$P(\Pi_n = \pi) = \frac{n!}{\theta(\theta+1)\dots(\theta+n-1)} \prod_{j=1}^n \frac{\theta^{a_j}}{j^{a_j} a_j!}.$$

► $E(N_n(1)) = n\theta/(n + \theta - 1)$

Si généalogie = Beta(2 - α , α)

Théorème (BBS). Fix $k \geq 0$.

Si la généalogie est décrite par un Beta-coalescent (α).

$S_n = \#$ sites de ségrégation

$$n^{\alpha-2} S_n \rightarrow_p \theta \frac{\alpha(\alpha-1)\Gamma(\alpha)}{2-\alpha}.$$

$$n^{\alpha-2} N_k(n) \rightarrow_p \theta \alpha(\alpha-1)^2 \frac{\Gamma(k+\alpha-2)}{k!}$$

- ▶ **Li and Hedgecock (1998)** *Genetic heterogeneity [...] among samples of larval Pacific oysters (Crassostrea gigas) supports the hypothesis of large variance in reproductive success* Can. J. Fish. Aquat. Sci.
- ▶ Analyse un échantillon de $n = 666$ larves d'huitre. Trouve $S_n = 67$ haplotypes différents, et parmi eux 72% ne se trouvent que dans 1 indiv.
- ▶ Prediction pour Kingman : $\approx \theta / \theta \log n \approx 15\%$.
- ▶ Eldon Wakeley *Coalescent processes when the distribution of offspring number among individuals is highly skewed* et Birkner Blath *Inference for Λ -coalescents*.

Principaux résultats

► Soit Λ une mesure finie.

$$\psi(q) := \int_0^1 (e^{-qx} - 1 + qx)x^{-2} \Lambda(dx)$$

► ψ est le **mécanisme de branchement** d'un CSBP (processus de branchement à espace d'état continu) $(Z_t, t \geq 0)$

Si

$$\int_0^\infty \frac{dq}{\psi(q)} < \infty \tag{1}$$

on pose

$$u(s) = \int_s^\infty \frac{dq}{\psi(q)}$$

et soit $v(t)$ son inverse cadlag.

Theorem (Berestycki - B. - Limic 2007) Soit Π un Λ -coalescent.

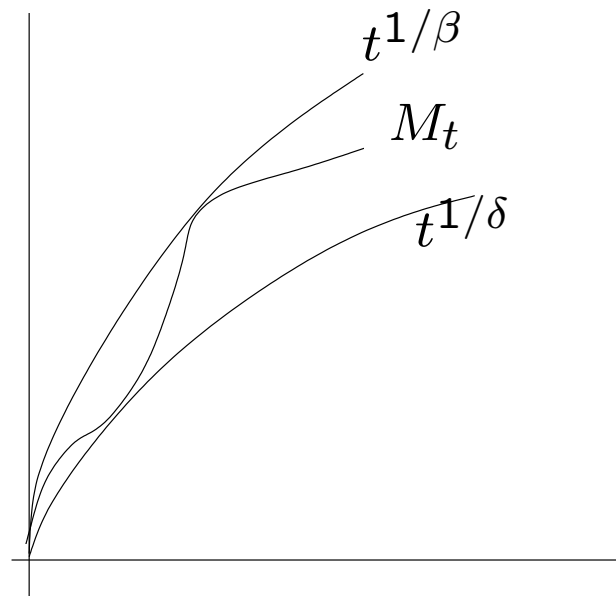
1. $\Pi \text{ CD}_\infty \Leftrightarrow Z \text{ s'\'eteint p.s.} \Leftrightarrow \int^\infty \frac{dq}{\psi(q)} < \infty.$

2. Quand c'est le cas, p.s.

$$\frac{N_t}{v(t)} \xrightarrow[t \rightarrow 0]{} 1$$

dès que **lower-index** $\delta > 1.$

► ψ est l'exposant de Laplace d'un processus de Lévy Y_t . Soit $M_t = \max_{s \leq t} Y_s$.



$\beta \geq \delta$ est le **upper-index**. (Pruitt 1981, Blumenthal-Gettoor 1961). En général $0 \leq \delta \leq \beta \leq 2$

Corollaire 1. Si $\delta > 1$ alors

$$t^{-\frac{1}{\delta-\varepsilon-1}} \leq N_t \leq t^{-\frac{1}{\beta+\varepsilon-1}}$$

Corollaire 2. Kingman est le coalescent le plus rapide. Si $\Lambda([0, 1]) = 1$,

$$P(N_t \leq \frac{2}{t}(1 - \varepsilon)) \rightarrow 0$$

(pas de restriction sur Λ).

- ▶ Bertoin et Le Gall (2006) observent analytiquement que :

$$CD_{\infty} \iff \int^{+\infty} \frac{dq}{\psi(q)} < \infty$$

où

$$\psi(q) = \int_0^1 (e^{-qx} - 1 + qx)x^{-2} \Lambda(dx)$$

- ▶ ψ est le mécanisme de branchement d'un CSBP $(Z_t, t \geq 0)$, et un théorème de Grey 1974:

$$\iff Z \text{ s'éteint en temps fini.}$$

- ▶ Nous donnons une explication probabiliste.

D'où ça vient ?

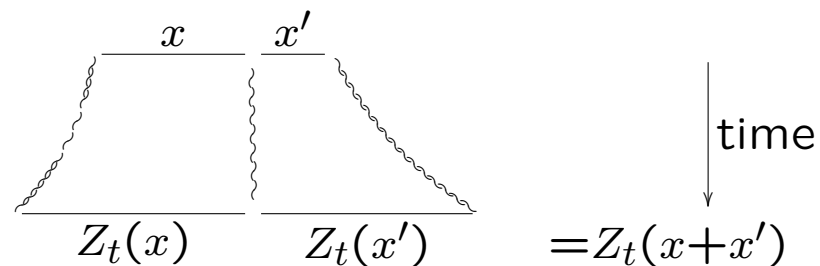
- ▶ Structure arborescente unit divers objets : coalescents, CSBP, arbres continus aléatoires et processus de Fleming-Viot généralisés.

Qu'est ce qu'un CSBP ? (continuous-state branching process)

- ▶ Décrit la taille d'une population continue (dans \mathbf{R}^+) : ce sont les analogues continus des processus de GW.
- ▶ Formellement : $(Z_t, t \geq 0)$ est un CSBP si Markov à valeur dans \mathbf{R}^+ , et **propriété de branchement**

$$Z_t(x + x') \stackrel{d}{=} Z_t^{(1)}(x) + Z_t^{(2)}(x')$$

- ▶ **Interpretation :**



- La loi de Z est caractérisée par $\psi =$ mécanisme de branchement

$$E(e^{-\lambda Z_t(x)}) = \exp(-x u_t(\lambda))$$

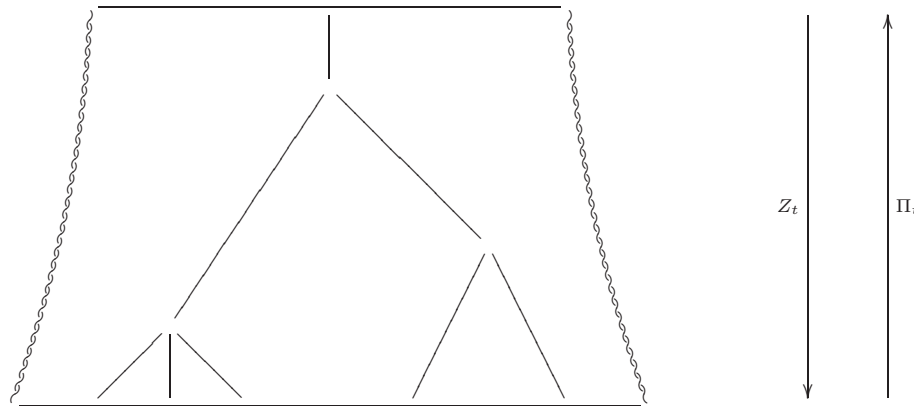
où

$$\frac{d}{dt} u_t(\lambda) = -\psi(u_t(\lambda))$$

et $u_0(\lambda) = \lambda$.

- ψ doit être l'exposant de Lévy d'un Lévy relié à Z par un changement de temps (Lamperti).
- Le branchement est α -stable si $\psi(u) = u^\alpha$ ($\alpha = 2$ est la diffusion de Feller $dZ_t = \sqrt{Z_t} dB_t$).

► Intuition:



► Pour un CSBP, # d'ancêtres de la population au temps $-t$ est $v(t)$. Donc partition ancestrale ($i \sim j \Leftrightarrow u_i, u_j$ ont même ancêtre) a $v(t)$ blocs.

► En temps petit, “genealogie” de $Z_t \approx \Lambda$ -coalescent.

► Généalogie d'un CSBP : Donnelly-Kurtz construction "look-down "

► Généalogie d'un Fleming-Viot généralisé est Λ -coalescent

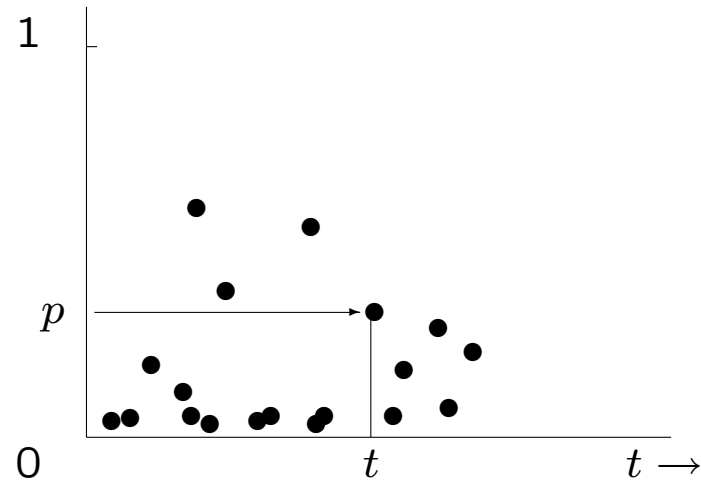
Clé : Construction par processus de points

Construction par PPP du Λ -coalescent

► On suppose $\Lambda(\{0\}) = 0$

$$\lambda_{b,k} = \int_0^1 p^{k-2} (1-p)^{b-k} \Lambda(dp) = \int_0^1 p^k (1-p)^{b-k} p^{-2} \Lambda(dp)$$

$(t_i, p_i) =$ atoms of a P.P.P. $dt \otimes p^{-2} \Lambda(dp)$



Pour chaque atome (t, p) on fait une *p-fusion* au temps t (tirage à pile ou face).

Si (t, p) est un atome : pour chaque bloc *tirage* avec $P(\text{face}) = p$. Tous les blocs "face" fusionnent.

Construction d'un CSBP par PP

Qu'est ce que la généalogie de Z ? On définit un processus à valeur mesure

$$M_t([0, x]) = Z_t(x)$$

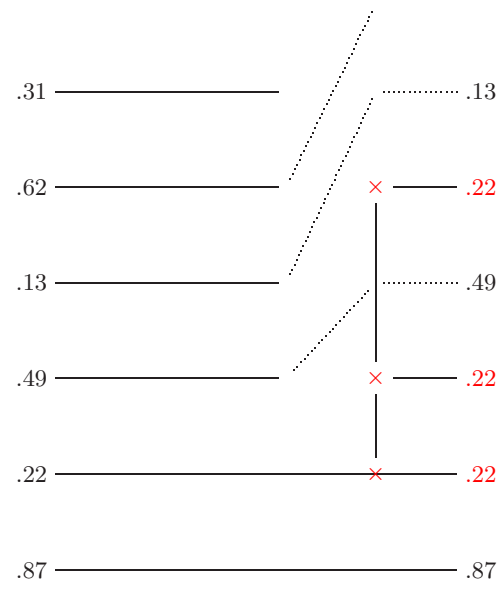
et $R_t(\cdot) = M_t(\cdot)/M_t([0, 1])$.

► Construction lookdown : processus $(\xi_1(t), \xi_2(t), \dots)$ tel que $\Xi_t(\cdot) := \lim_n \frac{1}{n} \sum_{i=1}^n \delta_{\xi_i(t)} \stackrel{d}{=} R_t(\cdot)$

Au départ $\xi_i(0)$ sont iid $u_{[0,1]}$.

On prend $U_{i,j}$ i.i.d. uniformes sur $[0, 1]$

Processus	$Z(t)$ CSBP ψ
PP	$(t_i, z_i) = (t_i, \Delta Z_{t_i}/Z_{t_i})$
$\forall i : \forall j : \eta_j^{(i)}$	$\mathbf{1}_{U_{i,j} \leq z_i}$
Au temps t_i	niveaux j t.q. $\eta_j^{(i)} = 1$ participent (autres sont poussés)



And now something (not so) completely different

Processus de Fleming Viot généralisé (Bertoin - Le Gall) $\Theta_t(\cdot)$ est un processus à valeur mesure, dual du Λ -coalescent.

On peut décrire son évolution comme suit : avec un certain taux $x^{-2}\Lambda(dx)$ on tire un point y selon Θ_{t-} et on remplace Θ_{t-} par $(1-x)\Theta_{t-} + x\delta_y$. L'indiv. y produit une fraction x de la génération suivante.

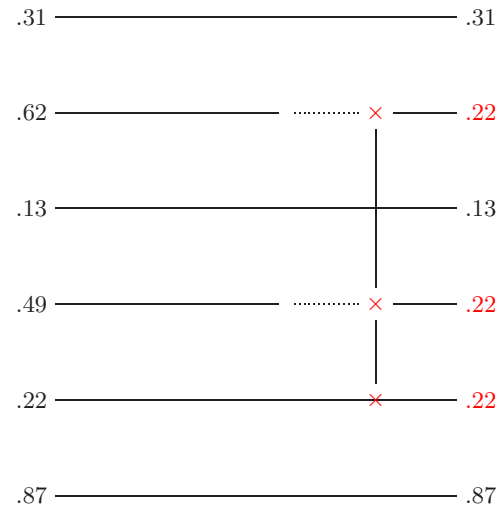
► On peut construire Θ comme mesure empirique d'un système de particules : $(\theta_1(t), \theta_2(t), \dots)$ t.q.

$$\lim_n \frac{1}{n} \sum_{i=1}^n \delta_{\theta_i(t)} \stackrel{d}{=} \Theta_t(\cdot)$$

Au départ $\theta_i(0)$ sont iid $u_{[0,1]}$.

Soit $U_{i,j}$ i.i.d. uniformes sur $[0, 1]$

Processus	$X(t)$ Lévy Ψ
PP	$(t_i, x_i) = (t_i, \Delta X_{t_i})$
$\forall i : \forall j : \eta_j^{(i)}$	$\mathbf{1}_{U_{i,j} \leq x_i}$
Au temps t_i	niveaux j t.q. $\eta_j^{(i)} = 1$ participent (autres sont tués)



- ▶ En temps petit $(t_i, \Delta Z_{t_i}/Z_{t_i}) \sim (t_i, \Delta X_{t_i})$ car Z est X changé de temps et $Z_t \rightarrow 1$.
- ▶ On peut donc comparer $\#$ de blocs d'un Λ -coalescent avec $\#$ de familles d'un CSBP.
- ▶ puis techniques de martingale et limites fluides pour la preuve.