

# Analyse du $\Lambda$ -coalescent : renouer avec ses racines.

SALA Raphaël, MUGISHA Axcél, GARCIA Hugo, COLLIN Thibault

Novembre 2025

## 1 Introduction

### 1.1 Fondements du $\Lambda$ -coalescent

La théorie de la coalescence modélise le phénomène par lequel des individus d'une population partagent un ancêtre commun. Nous souhaitons étudier rétrospectivement leur évolution.

Historiquement, le modèle de Wright-Fisher étudie une population de taille finie  $N$  où les individus d'une générations coalescent de manière uniforme entre eux dans la génération précédente [Fis30]. Ensuite, le modèle de Kingman [Kin82] est le modèle limite de Wright-Fisher où l'on s'intéresse à  $n < N$  lignées et en considérant  $N \rightarrow +\infty$ . Ce cadre asymptotique permet de simplifier grandement l'étude du phénomène de coalescence. Le modèle peut à présent être décrit comme un processus de Markov.

En 1999, Pitman et Sagitov généralisent le modèle de Kingman en autorisant la coalescence simultanée de plusieurs lignées. Des individus peuvent engendrer une proportion non négligeable de la population. Afin de définir un modèle, nous supposons raisonnablement que les lignées coalescent aléatoirement et indépendamment de leur histoire passée, c'est-à-dire en supposant l'absence de mémoire (propriété de Markov), que toutes les lignées ont les mêmes chances de coalescer entre elles que l'on appelle l'échangeabilité et enfin que nous ayons l'absence de collisions multiples signifiant qu'à tout instant donné, il ne peut y avoir qu'un seul événement de fusion en un même ancêtre.

**Théorème 1** (Pitman-Sagitov [Pit99, Sag99]). *Il existe un processus de Markov,  $(N_t)_{t \geq 0}$ , appelé  $\Lambda$ -coalescent, échangeable à collisions multiples simples si et seulement s'il existe une mesure finie  $\Lambda$  sur  $[0, 1]$  telle que, lorsqu'on a  $b$  lignées, pour tout  $2 \leq k \leq b$  le taux auquel chaque  $k$ -uplet fixé de lignées fusionne vaut,*

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

Nous ne définissons par formellement les conditions ici et donnons encore moins une preuve car cela est au-delà du cadre de ce rapport. Ce résultat montre que la dynamique de  $(N_t)_{t \geq 0}$ , indiquant le nombre de lignées à l'instant  $t$ , est entièrement caractérisée par une mesure finie. Sans perte de généralité nous considérons pour la suite une mesure de probabilité,  $\Lambda$  sur  $[0, 1]$ . Partant de  $b$  lignées, le taux d'une  $k$ -coalescence ( $2 \leq k \leq b$ ) est  $r_{b,k} := \binom{b}{k} \lambda_{b,k}$ . Le taux de sortie de l'état  $b$  est la somme des taux donc

$$\lambda_b = \sum_{k=2}^b r_{b,k} = \int_0^1 S_b(x) \Lambda(dx), \quad S_b(x) := \sum_{k=2}^b \binom{b}{k} x^{k-2} (1-x)^{b-k} = \frac{1 - (1-x)^b - bx(1-x)^{b-1}}{x^2} \quad (1)$$

D'après le lemme des réveils, à chaque événement de coalescence on passe de  $b$  à  $b - k + 1$  lignées avec probabilité,

$$\forall b \geq k \geq 2, \quad p_{b,k} := \frac{r_{b,k}}{\sum_{k=2}^b r_{b,k}} = \frac{\binom{b}{k} \lambda_{b,k}}{\lambda_b}$$

Ainsi, le squelette du processus est une chaîne de Markov décroissante sur  $\llbracket 1, n \rrbracket$ , commençant en  $n$  et absorbée presque sûrement en 1.

## 1.2 Exemple (Kingman)

Intéressons-nous au modèle de Kingman en guise d'introduction. On pose  $\Lambda = \delta_0$ . Pour  $2 \leq k \leq b$ ,

$$\lambda_{b,k} = \int_0^1 x^{k-2}(1-x)^{b-k} \delta_0(x) dx = [x^{k-2}(1-x)^{b-k}]_{x=0} = \begin{cases} (1-0)^{b-2} = 1 & \text{si } k = 2 \\ 0^{k-2}(1-0)^{b-k} = 0 & \text{si } k > 2 \end{cases}$$

Les coalescences se font que par paires. Une caractéristique intéressante du  $\Lambda$ -coalescent est le TMRCA (Time to the Most Recent Common Ancestor), c'est-à-dire le plus petit temps tel que toutes les lignées ont fusionné en un ancêtre commun. Dans la suite du rapport, nous le notons

$$\tau_{\Lambda,n} = \inf\{t \geq 0, N_t = 1, N_0 = n\}$$

Lorsque le contexte est clair sur  $\Lambda$  ou  $n$ , ceux-ci seront omis afin de rendre la lecture plus agréable.

**Lemme 1.** Soit  $\Lambda$  une mesure de probabilité sur  $[0, 1]$ . Notons  $H : b \in \mathbb{N}^* \mapsto \mathbb{E}_\Lambda(\tau \mid N_0 = b)$ . Alors  $H(1) = 0$ ,  $H(2) = 1$  et pour  $b \geq 3$ ,

$$H(b) = \frac{1}{\lambda_b} + \sum_{k=2}^{b-1} p_{b,k} H(b-k+1)$$

*Démonstration.* Pour  $b = 1$ ,  $N_t = 1$  donc  $H(1) = 0$ . Pour  $b = 2$ , le seul saut possible est de 2 vers 1 lignée avec taux  $\lambda_2 = \binom{2}{2} \lambda_{2,2} = 1$ , d'où  $\tau \sim \text{Exp}(1)$  et  $H(2) = 1/1 = 1$ .

Fixons  $b \geq 3$ . Définissons le temps de la première coalescence,

$$T_1 := \inf\{t \geq 0, N_t \neq b\}$$

$(N_t)_{t \geq 0}$  est un processus de Markov avec un taux de saut  $\lambda_b$ , donc  $T_1 \sim \text{Exp}(\lambda_b)$  et donc  $\mathbb{E}_\Lambda(T_1 \mid N_0 = b) = \frac{1}{\lambda_b}$ . De plus, si  $K$  est la taille de la fusion au temps  $T_1$ , alors  $K \sim \sum_{k=2}^b p_{b,k} \delta_k$  et  $N_{T_1} = b - K + 1$ .

Considérons la filtration naturelle  $(\mathcal{F}_t)_{t \geq 0}$  de  $(N_t)_{t \geq 0}$ . Par la propriété de Markov forte au temps  $T_1$  et l'absence de mémoire,

$$\mathbb{E}_\Lambda(\tau - T_1 \mid \mathcal{F}_{T_1}, N_0 = b) = \mathbb{E}_\Lambda(\tau \mid N_{T_1}) = H(N_{T_1})$$

Ainsi en conditionnant par  $\mathcal{F}_{T_1}$ ,

$$\begin{aligned} H(b) &= \mathbb{E}_\Lambda(\tau \mid N_0 = b) = \mathbb{E}_\Lambda(T_1 \mid N_0 = b) + \mathbb{E}_\Lambda(\tau - T_1 \mid N_0 = b) = \frac{1}{\lambda_b} + \mathbb{E}_\Lambda(H(N_{T_1}) \mid N_0 = b) \\ &= \frac{1}{\lambda_b} + \sum_{k=2}^b \mathbb{P}_b(N_{T_1} = b - k + 1) H(b - k + 1) = \frac{1}{\lambda_b} + \sum_{k=2}^b p_{b,k} H(b - k + 1) \end{aligned}$$

Or  $H(1) = 0$ , donc le terme  $k = b$  s'annule. D'où le résultat.  $\square$

Pour  $b$  lignées observées, on a  $\lambda_b = \sum_{k=2}^b \binom{b}{k} \lambda_{b,k} = \binom{b}{2} \lambda_{b,2} = \binom{b}{2}$  donc, d'après le Lemme 1, la taille moyenne d'un arbre pour le modèle de Kingman est donné par,

$$H(b) = \frac{1}{\lambda_b} + p_{b,2} H(b-1) = \frac{1}{\binom{b}{2}} + H(b-1)$$

Ainsi par récurrence,

$$H(b) = \sum_{k=2}^b \frac{1}{\binom{k}{2}} = \sum_{k=2}^b \frac{2}{k(k-1)} = \sum_{k=2}^b 2 \left( \frac{1}{k-1} - \frac{1}{k} \right) = 2 \left( 1 - \frac{1}{b} \right) \quad (2)$$

Là je propose donc (1 page grand max) (le but n'est pas de faire une étude de Kingman mais de présenter les différents objets du modèle de manière simple et visuelles)

- poser  $\Lambda = \delta_0$ ,
- l'intuition du modèle : la masse est vers 0 donc pas de grosse fusion. (Interprétation du modèle, du rôle des termes de l'intégrande)
- Les calculs en 5 lignes maximum ( $\lambda_{b,k}$ , TMRCA si utile, ...)
- Les notations :  $n$ ,  $N_t$ ,  $(C_t^i)_{0 < i < n}$ , TMRCA,  $T_k$

## 2 Analyse du TMRCA

### 2.1 Aux extrêmes de l'arbre.

Au vu du précédent exemple, on peut se demander l'influence de la mesure  $\Lambda$  sur le TMRCA. Intuitivement, ce temps moyen devrait diminuer lorsque la masse de  $\Lambda$  se rapproche de 1 puisqu'on autorise des coalescences multiples plus importantes. En première analyse on va étudier les deux cas extrêmes.

**Proposition 1.** *Soit  $n$  le nombre de lignées. Notons le TMRCA d'un  $\Lambda$ -coalescent,*

$$\tau := \inf\{t \geq 0, N_t = 1\}$$

*Alors, pour toute mesure de probabilité  $\Lambda$  sur  $[0, 1]$ , on a conditionnellement à  $\{N_0 = n\}$ ,*

$$1 = \mathbb{E}_{\delta_1}(\tau) \leq \mathbb{E}_{\Lambda}(\tau)$$

*Démonstration.* Prouvons l'égalité. Prenons  $\Lambda = \delta_1$ , nous avons  $\lambda_{n,k} = \delta_{n,k}$  (symbole de Kronecker), donc  $\lambda_n = \binom{n}{n} \lambda_{n,n} = 1$  donc  $\tau \sim \text{Exp}(1)$  et donc  $\mathbb{E}_{\delta_1}(\tau) = 1/1 = 1$ .

Soit  $\Lambda$  une mesure de probabilité sur  $[0, 1]$ . Notons  $H(b) := \mathbb{E}_{\Lambda}(\tau \mid N_0 = b)$ .

Montrons par récurrence forte l'inégalité, c'est-à-dire  $H(b) \geq 1$  pour  $b \geq 2$ . L'initialisation a été prouvée dans le lemme 1. Supposons l'inégalité vraie jusqu'à  $b-1$ . Remarquons que  $\lambda_{b,b} = \int_0^1 x^{b-2} \Lambda(dx) \leq \int_0^1 \Lambda(dx) = 1$ ,

$$H(b) = \frac{1}{\lambda_b} + \sum_{k=2}^{b-1} p_{b,k} H(b-k+1) \geq \frac{1}{\lambda_b} + \sum_{k=2}^{b-1} p_{b,k} = \frac{1}{\lambda_b} + 1 - p_{b,b} = 1 + \frac{1 - \lambda_{b,b}}{\lambda_b} \geq 1$$

D'où le résultat. □

Cette idée de déplacer la masse de  $\Lambda$  vers 1 pour diminuer la moyenne du TMRCA est intuitive. Pour le problème inverse de maximisation du TMRCA nous souhaiterions déplacer la masse de  $\Lambda$  vers 0. C'est-à-dire prouver que le modèle de Kingman soit celui maximisant le temps moyen du TMRCA. Toutefois, voilà une grande surprise : ce n'est pas le cas !

**Proposition 2.** *Il existe  $n > 1$  et une mesure de probabilité  $\Lambda$  sur  $[0, 1]$  telle que, conditionnellement à  $\{N_0 = n\}$ ,*

$$\mathbb{E}_{\Lambda}(\tau) > \mathbb{E}_{\delta_0}(\tau)$$

*Démonstration.* Soit  $n = 8$ , dans l'exemple de Kingman (voir sous-section 1.2), nous avons une formule explicite.

$$\mathbb{E}_{\delta_0}(\tau) = 2 \left(1 - \frac{1}{8}\right) = \frac{14}{8} = 1.75$$

Soit  $\Lambda = \delta_{1/4}$ , alors d'après (1),

$$\lambda_{n,k} = \left(\frac{1}{4}\right)^{k-2} \left(\frac{3}{4}\right)^{n-k} \quad \lambda_n = 16 \left(1 - \left(\frac{3}{4}\right)^n - \frac{n}{4} \left(\frac{3}{4}\right)^{n-1}\right)$$

Ainsi, en calculant nous obtenons,

$$\mathbb{E}_{\delta_{1/4}}(\tau) = \frac{1}{\lambda_n} + \sum_{k=2}^{n-1} \frac{\binom{n}{k} \lambda_{n,k}}{\lambda_n} \mathbb{E}_{\delta_{1/4}}(\tau \mid N_0 = n - k + 1) = \frac{19954284839411683}{11337879079537330} > 1.7599662 \dots > 1.75$$

□

Nous conjecturons que le théorème peut être étendu pour tout  $n > 6$ . À notre connaissance l'étude de ce phénomène n'est pas documenté pour  $n$  fini. Seul un article de [KLLS17] s'intéresse la croissance de  $\sup_{\Lambda} \mathbb{E}_{\Lambda}(\tau)$  lorsque  $n \rightarrow \infty$ .

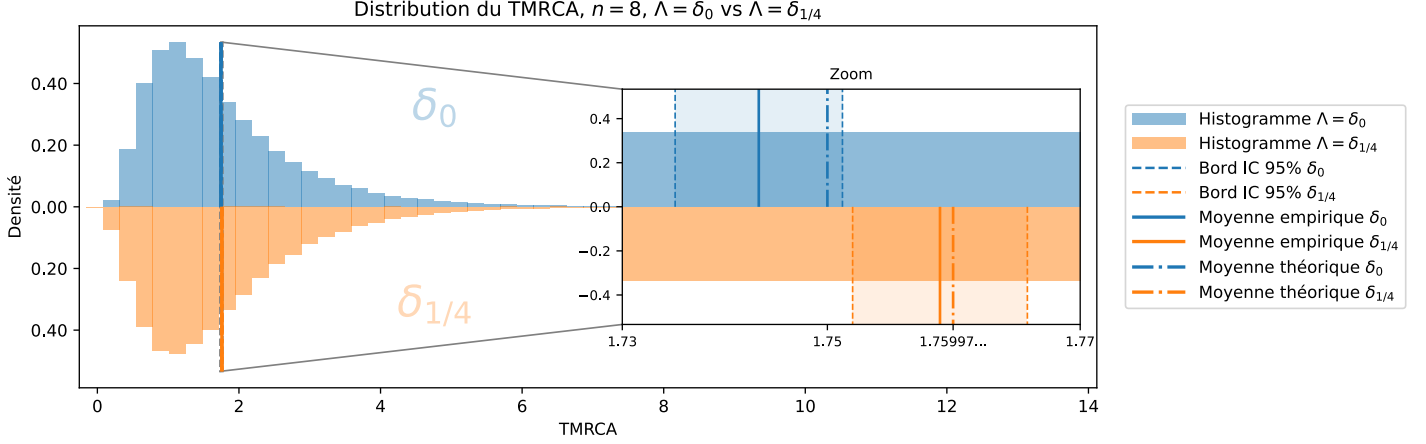


FIGURE 1 – Distribution empirique du TMRCA pour  $n = 8$  de  $\Lambda = \delta_0$  (Kingman, histogramme vers le haut en bleu) et  $\Lambda = \delta_{1/4}$  (histogramme vers le bas en orange). Les bandes verticales pointillées délimitent les intervalles de confiance (IC) à 95% pour la moyenne de  $\tau$  basés sur  $M = 1e5$  simulations indépendantes. Les lignes pleines indiquent les moyennes empiriques et les lignes pointillées ("-") les moyennes théoriques.

Dans la figure 1 nous utilisons des intervalles de confiances. Soit  $(T_i)_{1 \leq i \leq M}$  une collection de variables aléatoires i.i.d. de loi  $\tau_{\Lambda}$ . Posons la moyenne et l'estimateur de la variance

$$\bar{T}_M := \frac{1}{M} \sum_{i=1}^M T_i \quad s_M^2 := \frac{1}{M-1} \sum_{i=1}^M (T_i - \bar{T}_M)^2$$

Les intervalles de confiances sont construits génériquement. D'après le théorème central limite et le lemme de Slutsky,

$$\sqrt{M} \frac{\bar{T}_M - \mathbb{E}_{\Lambda}[\tau]}{\sqrt{s_M^2}} \xrightarrow[M \rightarrow \infty]{\mathcal{L}} \mathcal{N}(0, 1)$$

Ainsi, un IC asymptotique de niveau  $1 - \alpha$  est

$$[\bar{T}_M \pm q_{1-\alpha/2} \frac{\sqrt{s_M^2}}{\sqrt{M}}]$$

Dans la figure 1 le zoom à droite montre les intervalles de confiance à 95% disjoints renforçant l'observation  $\mathbb{E}(\tau_{\delta_0,8}) \leq \mathbb{E}(\tau_{\delta_{1/4},8})$ .

L'échelle de temps ici est en unités de  $N$  générations, avec  $N \gg n$  puisque nous considérons un modèle asymptotique.

## 2.2 Une forêt pas si grande

Un processus de Markov est entièrement déterminé par son générateur infinitésimal. Pour  $n$  lignées observées, celui d'un  $\Lambda$ -coalescent est la matrice triangulaire inférieure  $Q \in \mathcal{M}_n(\mathbb{R})$  définie pour tout  $1 \leq b, i \leq n$ , par

$$Q_{b,i} = \begin{cases} r_{b,k} & \text{si } b \geq 2 \text{ et } i = b - k + 1 \text{ pour } 2 \leq k \leq b \\ -\lambda_b & \text{si } b \geq 2 \text{ et } i = b \\ 0 & \text{sinon} \end{cases}$$

Le premier élément de sa diagonale,  $Q_{1,1}$ , est nul car l'état 1 est absorbant donc  $Q$  n'est pas inversible. En se restreignant à la sous-matrice  $R = (Q_{i,j})_{2 \leq i,j \leq n}$  la matrice devient inversible et nous pouvons exprimer

la densité de  $\tau$ . Posons  $p_R(t) = (p_k(t))_{2 \leq k \leq n}$  où  $p_k : t \geq 0 \mapsto \mathbb{P}(N_t = k \mid N_0 = n)$ . D'après la relation de Chapman-Kolmogorov,  $p_R$  vérifie pour tout  $t \geq 0$

$$\begin{cases} p'_R(t) = p_R(t)R \\ p_R(0) = (0, \dots, 0, 1) \end{cases} \iff p_R(t) = (0, \dots, 0, 1)e^{tR}$$

Définissons la fonction de survie,  $S : t \mapsto \mathbb{P}(\tau_n > t) = \mathbb{P}(N_t \neq 1 \mid N_0 = n) = \sum_{k=2}^n p_k(t) = p_R(t) \cdot \mathbf{1}$ . Donc la densité de  $\tau_n$  est donnée par,

$$f_\tau : t \mapsto d_t(1 - S(t)) = -S'(t) = -p'_R(t) \cdot \mathbf{1} = -p_R(t)R \cdot \mathbf{1} = -(0, \dots, 0, 1)e^{tR}R \cdot \mathbf{1} \quad (3)$$

On remarque également que ce processus est défini par  $(\lambda_{b,k})_{I_n}$  avec  $I_n := \{(b, k), 2 \leq k \leq b \leq n\}$ . Définissons pour  $r \in \llbracket 0, n-2 \rrbracket$

$$m_r : \Lambda \mapsto \int_0^1 x^r \Lambda(dx)$$

En développant l'intégrande des taux de fusions, pour tout  $(b, k) \in I_n$ , il existe  $A_n \in \mathcal{M}_{I_n, n-1}(\mathbb{R})$  tel que,

$$\lambda_{b,k} = \sum_{r=0}^{n-2} A_{(b,k),r} m_r(\Lambda)$$

Pour  $n$  fixé,  $Q$  est entièrement déterminée par  $(m_r(\Lambda))_{0 \leq r \leq n-2}$ , l'espace des mesures de probabilité sur  $[0, 1]$  se réduit à une projection de dimension finie,  $\mathbb{R}^{n-1}$ , donc un espace bien plus petit. Autrement dit, une infinité de mesures différentes deviennent indiscernables pour un processus considéré.

**Proposition 3.** Soit  $n > 1$  et  $\Lambda_0^\alpha := \text{Beta}(2 - \alpha, \alpha)$  avec  $\alpha \in ]1, 2[$ , de densité

$$w_\alpha : x \in [0, 1] \mapsto \frac{1}{B(2 - \alpha, \alpha)} x^{1-\alpha} (1-x)^{\alpha-1}$$

Soit  $(J_n)_{n \geq 0}$  les polynômes de Jacobi. Pour tout  $n \geq 0$ ,  $J_n = \sum_{k=0}^n \binom{n+\alpha-1}{n-k+1} \binom{n-\alpha+1}{k} x^{n-k} (x-1)^k$  est de degré  $n$  et orthogonal à tous les polynômes de degré inférieur à  $n-1$  pour le produit scalaire [NEM94]

$$\langle f, g \rangle_\alpha = \int_0^1 f(x)g(x)w_\alpha(x)dx$$

On pose

$$M := \sup_{x \in [0,1]} |J_{n-1}(x)| \in ]0, +\infty[ \quad \text{et} \quad \varepsilon_n := \frac{1}{M}$$

Et on définit pour  $0 < \varepsilon < \varepsilon_n$ , la mesure de probabilité  $\Lambda_\varepsilon^\alpha \neq \Lambda_0^\alpha$  par sa densité

$$f_\varepsilon^\alpha : x \in [0, 1] \mapsto (1 + \varepsilon J_{n-1}(x))w_\alpha(x)$$

Alors, pour tout  $0 < \varepsilon < \varepsilon_n$ ,

$$\tau_{\Lambda_\varepsilon^\alpha, n} \stackrel{\mathcal{L}}{=} \tau_{\Lambda_0^\alpha, n}$$

*Démonstration.* Soit  $\varepsilon < 1/M$ , montrons que  $f_\varepsilon$  est bien une densité. Sur  $[0, 1]$ ,  $1 + \varepsilon J_{n-1} \geq 1 - \varepsilon M \geq 0$  donc par produit de termes positifs  $f_\varepsilon^\alpha \geq 0$ .

Puis, par orthogonalité de  $J_{n-1}$  avec la constante  $1 \in \mathbb{R}_{n-1}[X]$ ,

$$\int_0^1 J_{n-1}(x)w_\alpha(x)dx = 0$$

D'où,

$$\int_0^1 f_\varepsilon^\alpha(x)dx = \int_0^1 w_\alpha(x)dx + \varepsilon \int_0^1 J_{n-1}(x)w_\alpha(x)dx = 1 + \varepsilon \cdot 0 = 1$$

Ainsi  $f_\varepsilon^\alpha$  est bien une densité sur  $[0, 1]$ .

Montrons à présent que les générateurs infinitésimaux de  $\Lambda_0^\alpha$  et  $\Lambda_\varepsilon^\alpha$  coïncident. Pour tout  $r \in \llbracket 0, n-2 \rrbracket$ , on a  $X^r \in \mathbb{R}_r[X] \subset \mathbb{R}_{n-1}[X]$ , ainsi par orthogonalité de  $J_{n-1}$ ,

$$\int_0^1 x^r J_{n-1}(x) w_\alpha(x) dx = 0$$

Ainsi,

$$m_r(\Lambda_\varepsilon) = \int_0^1 x^r f_\varepsilon(x) dx = \int_0^1 x^r w_\alpha(x) dx + \varepsilon \int_0^1 x^r J_{n-1}(x) w_\alpha(x) dx = m_r(\Lambda_0)$$

Les taux de fusions sont donc égaux entre ces mesures, d'où le résultat.  $\square$

Ainsi nous venons de construire une infinité de mesures différentes qui induisent le même processus de coalescence. On s'attendait à obtenir une infinité d'arbres généalogiques différents mais ceux-ci sont identiques en loi. Pour la construction nous utilisons les polynômes de Jacobi. Toutefois au vu de la preuve, nous pouvons utiliser n'importe quelle familles de polynômes dits orthogonaux. **C'est vrai ça pourquoi Jacobi et pas tchbeychev ou legendre?? Pour avoir choisi ceux-là? toruve une raison pratique, j'ai pas internet pour chercher.**

**en admettant qu'on s'intresse à beta(alpha, 2-alpha), ca sera motivé dans la dernière section, pourquoi ici on ne peut pas considerer alpha entre 0 et 1?? Faut expliquer, donner une justification. Sinon c'ewst pas très gawe, on le mentionne pas e ton fait ocmme si rien n'était mais bizarre de pas prendre des plus petit alpha.**

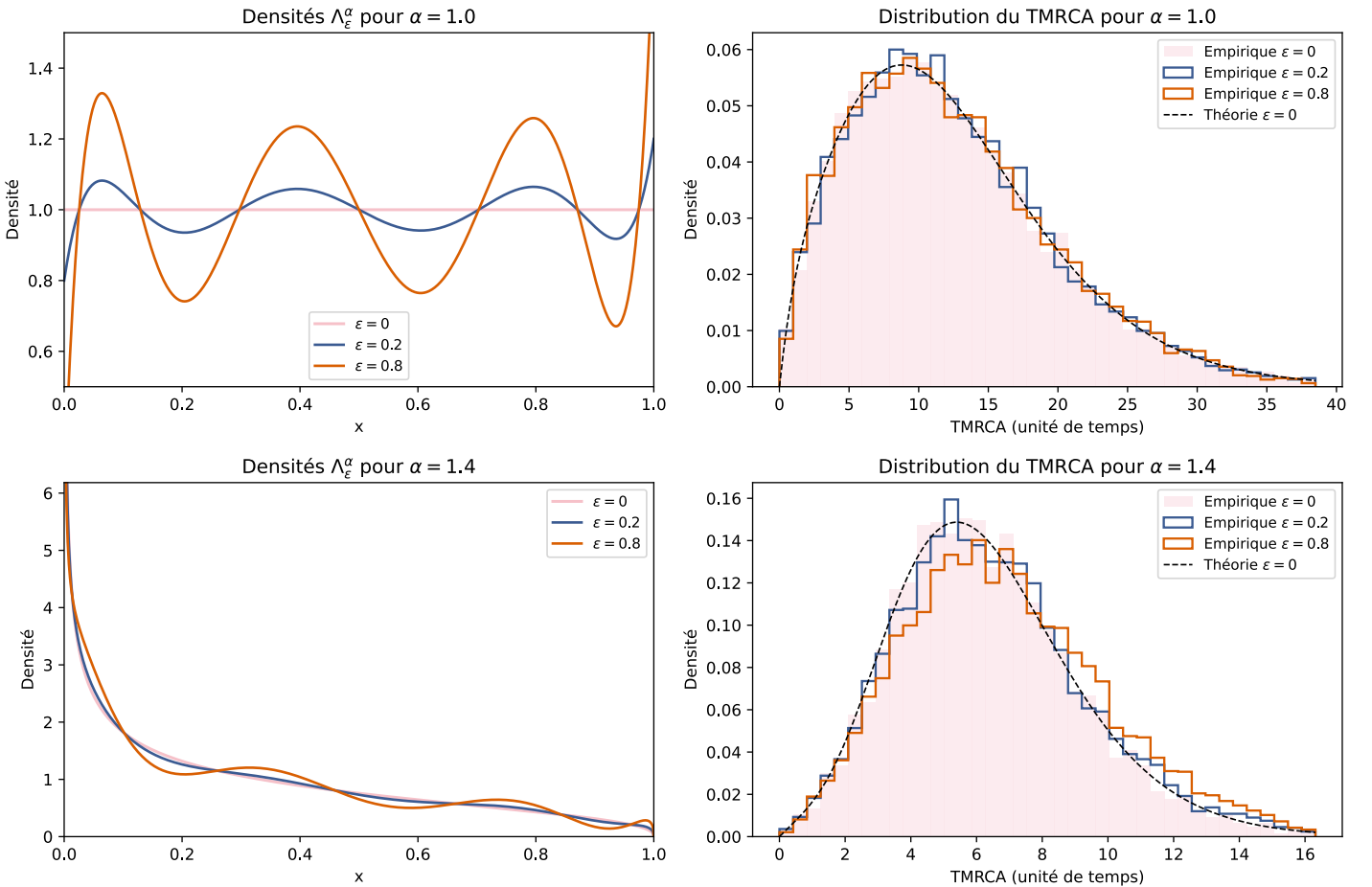


FIGURE 2 – (a) *Gauche* : comparaison des densités remarquablement différentes sur  $[0, 1]$ ,  $f_\varepsilon(x) = (1 + \varepsilon J_{n-1}(x)) w_\alpha(x)$ , pour  $n = 8$  : cas Beta( $2 - \alpha, \alpha$ ) (avec  $\alpha \in \{1, 1.4\}$ )  $\varepsilon = 0$  et perturbation de Jacobi  $\varepsilon \in \{0.2, 0.8\}$  (b) *Droite* : distribution empirique du TMRCA, issues de  $n$  lignées, à partir de 8000 simulations, avec superposition de la densité théorique donnée par (3).

**J'ai pas eu le temps (ni internet) pour reformuler cette partie comme je souhaiterai le faire (remarque WhatsApp) Dire que beta a été simulé par trapeze? montecarlo? faire quelques phrase dessus La figure**

met clairement en évidence le rôle crucial de la borne  $\varepsilon_n$ , qui garantit que la densité perturbée  $f_\varepsilon$  reste positive sur  $[1, 0]$  et définit bien une mesure de probabilité. Cette borne dépend directement du degré  $n$  du polynôme de jacobi [Sze39],

$$\varepsilon_n \sim \frac{C(\alpha)}{n^{|\alpha-1|}}$$

où  $C(\alpha)$  provient du comportement asymptotique des polynômes de Jacobi. Dans le cas,  $\alpha = 1$ , la mesure  $\Lambda_0^1$  est uniforme et  $\varepsilon_n$  ne dépend plus de  $n$ , d'où sur la figure les histogrammes empiriques collent partout. En revanche, pour  $\alpha = 1.4$ ,  $\varepsilon_n = 0.392062$ , et sur la figure on voit que si on prend  $\varepsilon$  plus grand que  $\varepsilon_n$  les histogrammes empiriques ne se superposent plus parfaitement.

### 2.3 Silence, ça pousse

Précédemment nous avons brièvement parlé de la mesure uniforme en prenant **si tu peux finir la phrase selon les notations prises**. Ce modèle est connu sous le nom de Bolthausen-Sznitman et décèle un résultat incontournable.

**Théorème 2** (Goldschmidt & Martin [GM05]). *Soit  $(N_t)_{t \geq 0}$  un Bolthausen-Sznitman coalescent. Alors,*

$$\tau_n - \log(\log(n)) \xrightarrow[n \rightarrow \infty]{\mathcal{L}} \mathcal{G}$$

où  $\mathcal{G}$  est la loi de Gumbel, de densité  $x \mapsto e^{-x-e^{-x}}$ .

La preuve est omise car elle dépasse le cadre de ce rapport. Toutefois, ce théorème renforce l'intuition qu'on a pu commencé à avoir à la Proposition 2 puisqu'on a que,

$$\lim_{n \rightarrow \infty} \mathbb{E}(\tau_n) = \lim_{n \rightarrow \infty} \log(\log(n)) + \mathbb{E}(\mathcal{G}) = \lim_{n \rightarrow \infty} \log(\log(n)) + \gamma = +\infty$$

où  $\gamma$  est la constante d'Euler-Mascheroni. En effet, comme l'illustre la figure 3, dès  $n \approx 50$  on observe  $\mathbb{E}(\tau_n) > 2$ , surpassant la borne du modèle de Kingman (2). Ainsi, la croissance de la hauteur des arbres généalogiques pour le modèle de Bolthausen-Sznitman est extrêmement lente mais permet d'obtenir des arbres aussi grand que l'on souhaite en moyenne.

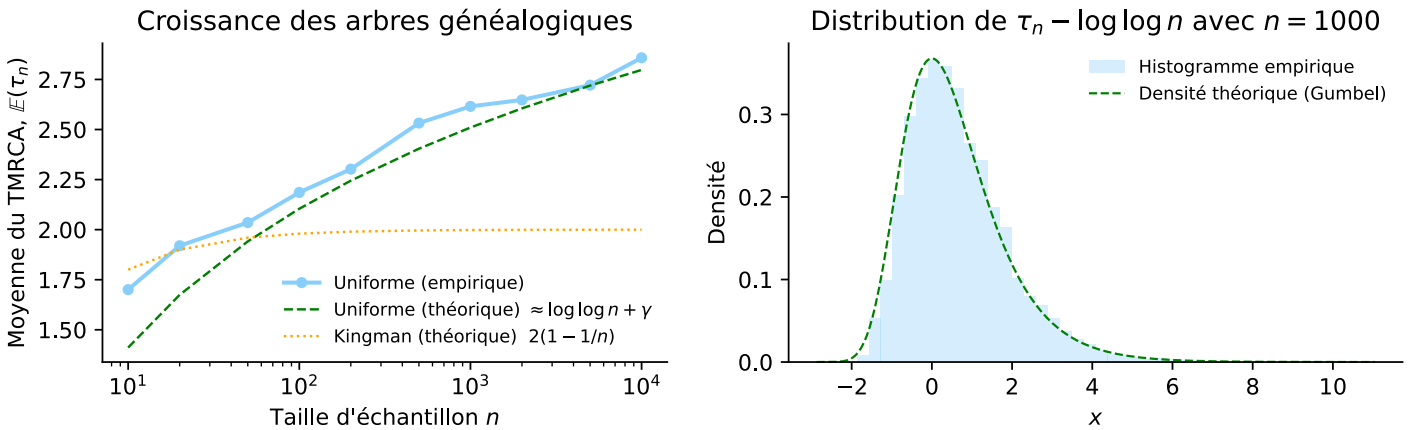


FIGURE 3 – (a) *Gauche* : sous le modèle de Bolthausen-Sznitman ( $\Lambda$  uniforme) nous déterminons les moyennes empiriques de  $\mathbb{E}(\tau_n)$  (500 répétitions par  $n \in \{10, 20, 50, 100, 200, 500, 1e3, 2e3, 5e3, 1e4\}$ ) comparées à l'approximation théorique  $\gamma + \log \log n$  et aussi à Kingman (2). (b) *Droite* : histogramme de  $\tau_n - \log \log n$  pour  $n = 1000$  (5000 simulations) avec superposition de la densité de Gumbel  $x \mapsto e^{-x-e^{-x}}$ .

### 3 Application : Le Paradoxe de la Sardine Japonaise

L'étude de Niwa et al. (2016) sur la sardine japonaise (*Sardinops melanostictus*) constitue une démonstration empirique de la nécessité du  $\Lambda$ -coalescent. Elle met en lumière l'échec du modèle de Kingman pour les espèces à stratégie de "Sweepstakes" ou "loterie reproductive" (haute fécondité, variance reproductive infinie), où quelques rares individus peuvent engendrer une fraction macroscopique de la génération suivante (fusions multiples).

L'idée de cette sous-partie est d'analyser une possible application du modèle de  $\Lambda$ -coalescence sur la mesure  $\text{Beta}(2 - \alpha, \alpha)$  au monde réel.

#### 3.1 Le Conflit

Chez les animaux à haute fécondité, l'analyse génétique révèle une contradiction flagrante si l'on reste dans le paradigme binaire classique (Wright-Fisher ou Kingman) :

On observe un excès massif de mutations uniques. Sous Kingman, cette observation est interprétée à tort comme une expansion démographique explosive et récente.

Si cette expansion était réelle, la distribution des distances par paires (mismatch distribution) devrait former une "vague" unimodale (loi de Poisson). Or, les données montrent une distribution en forme de "L" : un pic en zéro et une queue lourde.

#### 3.2 Résolution par le Beta-Coalescent et Simulation

Le modèle Beta-coalescent (avec  $\alpha \approx 1.3$ ) réconcilie ces deux signaux. Les fusions multiples fréquentes créent des groupes d'individus identiques (le pic à 0), tout en laissant survivre des lignées très anciennes (queue lourde), sans recourir à l'hypothèse d'expansion.

Nous avons simulé cette distinction (Fig. 4) en comparant un modèle de Kingman avec expansion (l'artefact) et un modèle Beta stationnaire (la réalité).



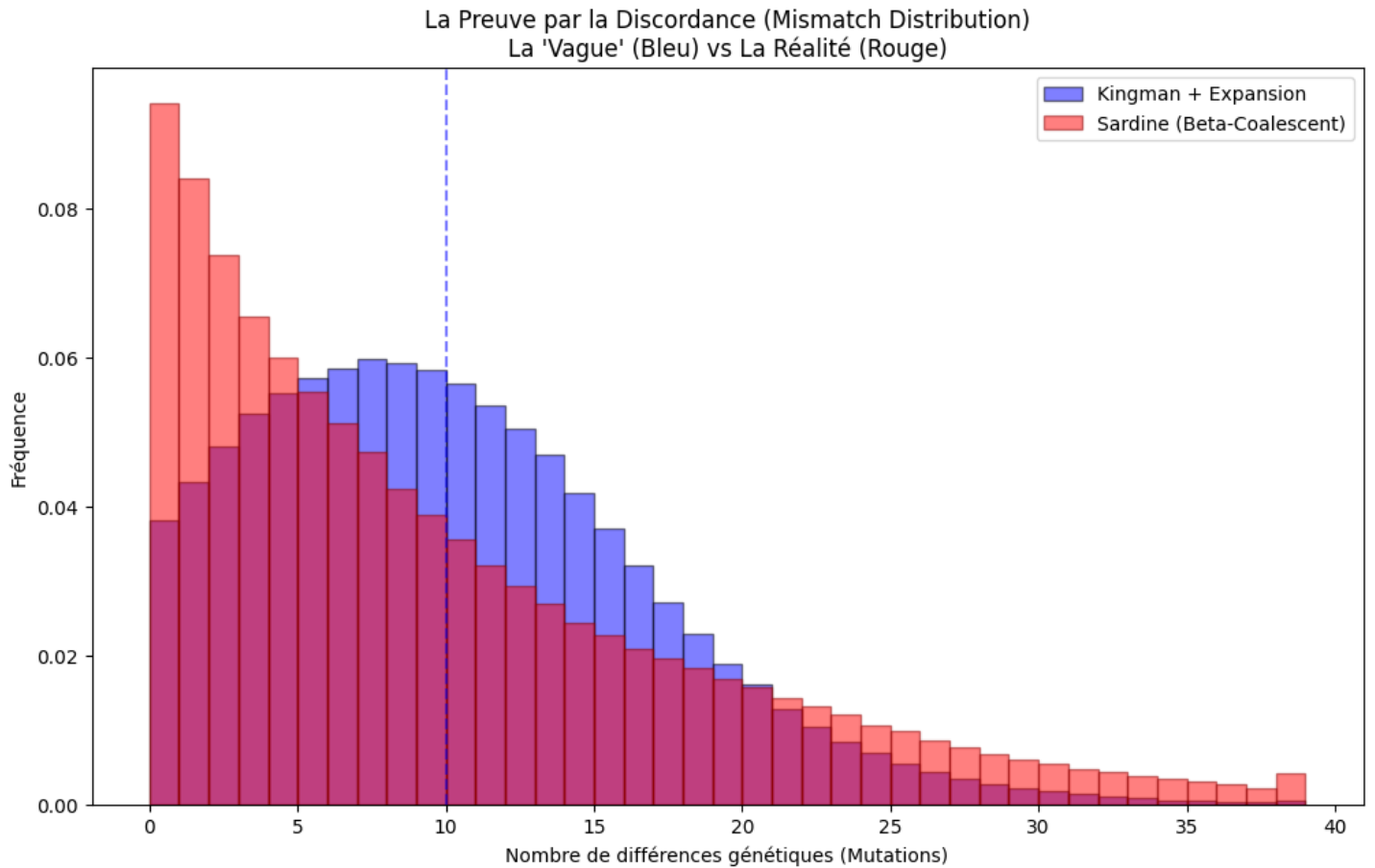


FIGURE 4 – Distribution des discordances ( $n = 150$ ). **En bleu (Kingman + Expansion)** : Une "vague" créée par le phénomène de croissance démographique. **En rouge (Sardine  $\alpha = 1.3$ )** : Un pic de clones (sweepstakes récents) et une queue de lignées persistantes (résultat attendu).

### 3.3 Conséquence : Millions de sardines, même génome

Ce résultat invalide la relation linéaire entre taille de recensement  $N$  et taille efficace  $N_e$ . Sous le régime Beta-coalescent, la dérive génétique est régie par une loi de puissance :

$$N_e \propto N^{\alpha-1}$$

Pour la sardine ( $\alpha \approx 1.3$ ), cela implique  $N_e \propto N^{0.3}$ . Cette relation explique pourquoi des populations marines gigantesques conservent une diversité génétique très faible et instable. Ce modèle permet non seulement d'expliquer un ancien paradoxe chez les animaux à haute fécondité mais sert en plus d'outil pour estimer la taille effective d'une population.

## Références

- [Fis30] Ronald A. Fisher. *The Genetical Theory of Natural Selection*. Clarendon Press, Oxford, 1930.
- [GM05] Christina Goldschmidt and Jeremy B. Martin. Random recursive trees and the bolthausen-sznitman coalescent. *Electronic Journal of Probability*, 10 :718–745, 2005.
- [Kin82] J. F. C. Kingman. The coalescent. *Stochastic Processes and their Applications*, 13(3) :235–248, 1982.
- [KLLS17] Tobias Kluge, Koenigs Leckey, Wolfgang Löhner, and Jason Schweinsberg. Exchangeable coalescents, ultrametrics, and trees. 2017.

- [NEM94] Paul Nevai, Tamás Erdélyi, and Alphonse P Magnus. Generalized jacobi weights, christoffel functions, and jacobi polynomials. *SIAM Journal on Mathematical Analysis*, 25(2) :602–614, 1994.
- [Pit99] Jim Pitman. Coalescents with multiple collisions. *The Annals of Probability*, 27(4) :1870–1902, 1999.
- [Sag99] Serik Sagitov. The general coalescent with asynchronous mergers of ancestral lines. *Journal of Applied Probability*, 36(4) :1116–1125, 1999.
- [Sze39] G. Szegő. *Orthogonal Polynomials*, volume 23 of *Colloquium Publications*. American Mathematical Society, 1939.