Analyse du Λ-coalescent : renouer avec ses racines.

SALA Raphaël, MUGISHA Axcel, GARCIA Hugo, COLLIN Thibault

novembre 2025

1 Introduction

1.1 Fondements du Λ-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 $N \to +\infty$. 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 Λ-coalescent, échangeable à collisions multiples simples si et seulement s'il existe une mesure finie Λ 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 est entièrement caractérisée par une mesure finie, sans perte de généralité nous considérons pour la suite une mesure de probabilité, Λ 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

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

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 $[\![1,n]\!]$, 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 Λ-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 Λ ou n, ceux-ci seront omis afin de rendre la lecture plus agréable.

Lemme 1. *Soit Λ une mesure de probabilité sur $[0,1]$. Notons $H : b \in \mathbb{N}^* \longmapsto \mathbb{E}_\Lambda(\tau \mid N_0 = b)$. 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 λ_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 $\mathcal{F}_{T_1}$,

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

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

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,

$$H(b) = \sum_{k=2}^{b} \frac{1}{\binom{k}{2}} = \sum_{k=2}^{b} \frac{2}{k(k-1)} = \sum_{k=2}^{b} 2\Big(\frac{1}{k-1} - \frac{1}{k}\Big) = 2\Big(1 - \frac{1}{b}\Big) \tag{2}$$

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. (Inteprétation du modele, du role 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
Subplots (kingman)
— (Une réalisation) Arbre + TMRCA
— (Sur plusieurs réalisaiton) Distribution des fusions ((C_t^i))_{0<i<n } (donc uniquement en 2 normalement)
— (Sur plusieurs réalisation) distribution TMRCA + densité avec en légende "densité théorique" dans le caption mettre, ou en footnote, que la densité est explcite dans ce cas dont la formule n'est pas prouvée, ele peut être écritre matriciellement d'après (citer "une foret pas si grande")

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 Λ sur le TMRCA. Intuitivement, ce temps moyen devrait diminuer lorsque la masse de Λ 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 Λ-coalescent,*

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

Alors, pour toute mesure de probabilité Λ sur $[0,1]$, on a condtionnellement 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 Λ 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 Λ vers 1 pour diminuer la moyenne du TMRCA est intuitive. Pour le problème inverse de maximisation du TMRCA nous souhaiterions déplacer la masse de Λ vers 0. C'est-à-dire prouver que le modèle de Kingman soit celui maximisant le temps moyen du TMRCA. Toutefois, voila une grande surprise : ce n'est pas le cas !

Proposition 2. *Il existe $n > 1$ et une mesure de probabilité Λ 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),

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

Ainsi, en calculant nous obtenons,

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

□

Nous conjecturons que le théorème peut être étendu pour tout $n > 6$. A notre connaissance l'étude 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 $n \to \infty$.

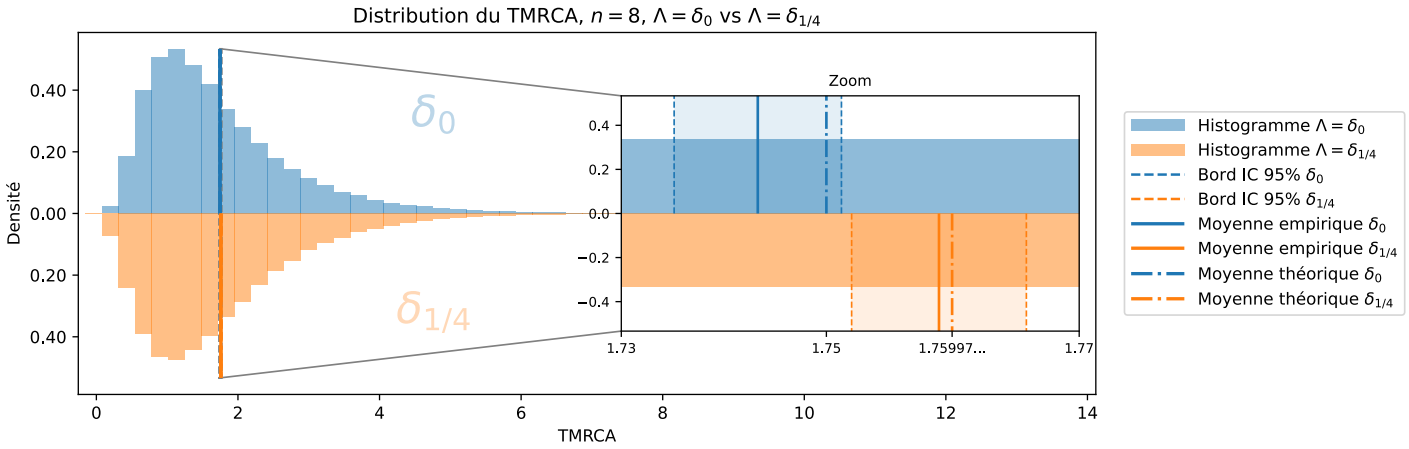

FIGURE 1 – Distribution empirique du TMRCA pour $n = 8$ de $\Lambda = \delta_0$ (Kingman, histogramme vers le haut en bleu) et $\Lambda = \delta_{1/4}$ (histogramme vers le bas en orange). Les bandes verticales pointillées délimitent les intervalles de confiance (IC) à 95% pour la moyenne de τ basés 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. Soient $(T_i)_{1\le i\le M}$ collections de variables aléatoires i.i.d. de loi τ_Λ. Posons la moyenne et l'estimateur de la variance

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

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

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

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

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

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

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 Λ-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 τ. 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 τ_n est donnée par,

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

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

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

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 $n > 1$. Notons P_k le polynôme de Legendre de degré k. Prenons Λ_0 la mesure uniforme sur $[0,1]$. Pour tout $0 < \varepsilon < 1$, définissons la mesure de probabilité $\Lambda_\varepsilon \neq \Lambda_0$ de densité sur $[0,1]$*

$$f_\varepsilon(x) = 1 + \varepsilon P_{n-1}(2x - 1)$$

Alors, pour tout $0 < \varepsilon < 1$,

$$\tau_{\Lambda_0} \overset{\mathcal{L}}{=} \tau_{\Lambda_\varepsilon}$$

Démonstration. Montrons que pour tout $0 < \varepsilon < 1$, f_ε est bien une densité de probabilité. Nous avons, par une analyse élémentaire que pour tout $k \geq 0$ $|P_k| \leq 1$ sur $[-1,1]$ [WW20] donc $f_\epsilon \geq 1 - 1 \cdot \varepsilon \geq 0$.

Rappelons que $(P_k)_{0 \leq k \leq n}$ est une base orthogonale de $\mathbb{R}_n[X]$ pour le produit scalaire $f, g \mapsto \int_{-1}^{1} f(x)g(x)dx$ donc $\int_0^1 f_\varepsilon(x)dx = 1 + \frac{\varepsilon}{2}\int_{-1}^{1} 1 \cdot P_n(x)dx = 1 + 0 = 1$. Montrons à présent que le générateur infinitésimal de Λ_0 et Λ_ε sont identiques. Soit $r \in [\![0, n-2]\!]$, puisque $X^r \in \mathbb{R}_{n-2}[X] \subset \mathbb{R}_{n-1}[X]$,

$$m_r(\Lambda_\varepsilon) = \int_0^1 x^r f_\varepsilon(x)dx = \int_0^1 x^r dx + \varepsilon \int_0^1 x^r P_{n-1}(2x-1)dx = \frac{1}{r+1} + \varepsilon \cdot 0 = m_r(\Lambda_0)$$

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

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.

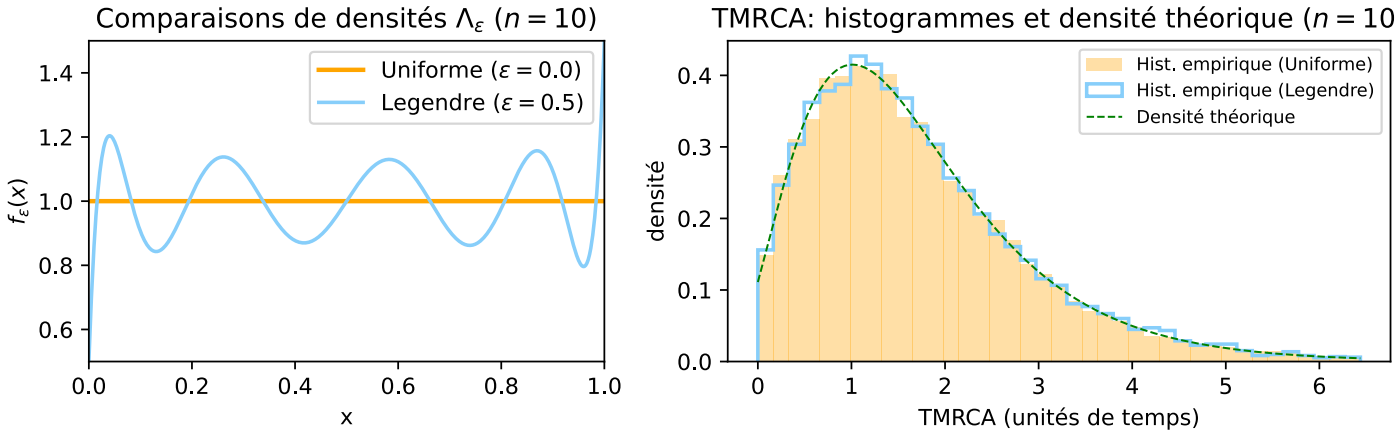

FIGURE 2 – (a) *Gauche* : comparaison des densités remarquablement différentes sur $[0, 1]$, $f_\varepsilon(x) = 1 + \varepsilon P_{n-1}(2x - 1)$, pour $n = 10$: cas uniforme $\varepsilon = 0$ et perturbation de Legendre $\varepsilon = 0.5$ (b) *Droite* : distribution empirique du TMRCA, issues de n lignées, à partir de 8000 simulations, avec superposition de la densité théorique donnée par (3). Comme attendu les densités se confondent.

2.3 Silence, ça pousse

Précedemment nous avons parlé de la mesure uniforme. 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,*

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

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,

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

où γ 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.

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 Löhr, and Jason Schweinsberg. Exchangeable coalescents, ultrametrics, and trees. 2017.

[Pit99] Jim Pitman. Coalescents with multiple collisions. *The Annals of Probability*, 27(4) :1870–1902, 1999.

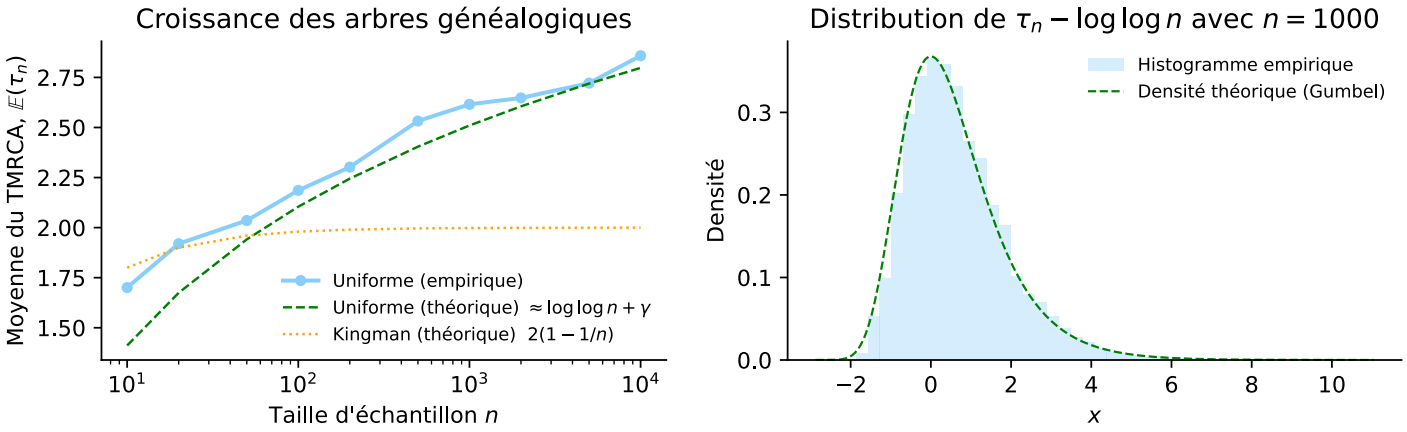

FIGURE 3 – (a) *Gauche* : sous le modèle de Bolthausen-Sznitman (Λ uniforme) nous déterminons les moyennes empiriques de $\mathbb{E}(\tau_n)$ (500 répétitions par $n \in \{10, 20, 50, 100, 200, 500, 1e3, 2e3, 5e3, 1e4\}$) comparées à l'approximation théorique $\gamma + \log\log n$ et aussi à Kingman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

[Sag99] Serik Sagitov. The general coalescent with asynchronous mergers of ancestral lines. *Journal of Applied Probability*, 36(4) :1116–1125, 1999.

[WW20] E. T. Whittaker and G. N. Watson. *A Course of Modern Analysis : An Introduction to the General Theory of Infinite Processes and of Analytic Functions ; with an Account of the Principal Transcendental Functions*. Cambridge University Press, Cambridge, 3rd edition, 1920. Chap. 15 : Legendre functions.