DE GRUYTER
OPEN

DEMONSTRATIO MATHEMATICA
Vol. XLVII No 2 2014

Krzysztof Bartoszek, Michał Krzemiński

CRITICAL CASE STOCHASTIC PHYLOGENETIC TREE MODEL VIA THE LAPLACE TRANSFORM

Abstract. Birth–and–death models are now a common mathematical tool to describe branching patterns observed in real–world phylogenetic trees. Liggett and Schinazi (2009) is one such example. The authors propose a simple birth–and–death model that is compatible with phylogenetic trees of both influenza and HIV, depending on the birth rate parameter. An interesting special case of this model is the critical case where the birth rate equals the death rate. This is a non–trivial situation and to study its asymptotic behaviour we employed the Laplace transform. With this, we correct the proof of Liggett and Schinazi (2009) in the critical case.

1. Introduction

Different viral types have phylogenetic trees exhibiting different branching properties, with influenza and HIV being two extreme examples. In the influenza tree, a single type dominates for a long time with other types dying out quickly until suddenly a new type completely takes over and the old type dies out. The HIV phylogeny is the complete opposite, with a large number of co–existing types.

In [6], a stochastic model is described and depending on the choice of parameters, it can exhibit both types of dynamics. We briefly describe the model after [6]. We only keep track of the number of different viral types at each time point t. Let $N(t)$ denote the number of distinct viral types at time t. In the nomenclature of phylogenetics, $N(t)$ counts the number of different species alive at time t. At each time point, the birth rate is $\lambda N(t)$ and the death rate is $N(t)$. If there is only one type alive then it cannot die. Clearly $N(t)$ is a Markov chain with discrete state space and continuous time. Each virus type is described by a fitness value that is randomly chosen at its birth. If a death event occurs, the type with smallest fitness dies. This

2010 *Mathematics Subject Classification*: Primary 60K35, Secondary 92D15.
Key words and phrases: phylogenetic tree, stochastic model, Tauberian theory.

DOI: 10.2478/dema-2014-0038

means that only the fitness ranks matter and so the exact distribution of a virus' fitness will not play a role.

The main result of [6] is the asymptotic behaviour of the dominating type, whether it is expected to remain the same for long stretches of time or change often. This is summarized in Theorem 1, [6]

Theorem 1.1. (Theorem 1, [6]) *Take* $\alpha \in (0,1)$. *If* $\lambda \leq 1$ *then*

$$\lim_{t\to\infty} \mathrm{P}(\textit{maximal types at times } \alpha t \textit{ and } t \textit{ are the same}) = \alpha,$$

while if $\lambda > 1$ *then this limit is* 0.

The proof of this theorem is based on considering successive visits to the state 1, in particular denote $\tau_1, \tau_2, \ldots$ to be the (random) times between visits of the chain to $N(t) = 1$ and $T_n := \tau_1 + \ldots + \tau_n$. In [6], the latter random variable is represented as

$$T_n = \sum_{i=1}^{n} X_i + \sum_{i=1}^{n} H_i,$$

where X_i are independent mean 1 exponential random variables and H_i are the (independent) hitting times of the state 1 conditional on starting in state 2. Descriptively H_i is the ith return from state 2 to the state of one virus type alive, since from 1 the chain has to jump to two types. The Markovian nature of the process ensures that the H_is are independent and identically distributed for distinct is. In the proof of Theorem 1, it is stated that the cumulative distribution function of H_i, $F(t)$, satisfies

$$\int_0^{F(t)} \frac{1}{1+s^2-2s}\, ds = t, \tag{1.1}$$

solved uniquely for

$$F(t) = \frac{t}{1+t}. \tag{1.2}$$

This gives the asymptotic behaviour (Lemma 3, [6])

$$\lim_{n\to\infty} \frac{T_n}{n\log(n)} = 1 \quad \text{in probability}, \tag{1.3}$$

from which the result of Theorem 1 is derived when $\lambda = 1$.

However Eq. (1.1) does not take into account that this model differs from a classical birth–death model where 0 is the absorbing state. We illustrate this in Fig. 1. We can easily re–numerate the state values, but the intensity values will differ between the two models. Correcting for this difference in intensity values one will still get the same asymptotics as in Eq. (1.3) and hence the same result as in Theorem 1 but with a more complicated proof.

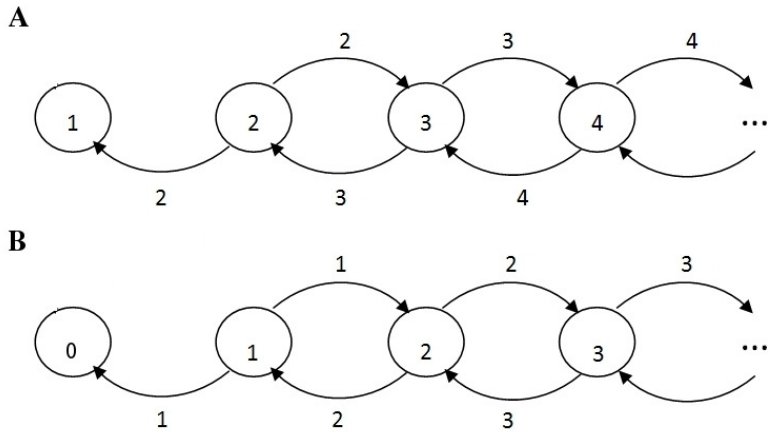

Fig. 1. A: Depiction of the Markov chain model described in [6]. B: Depiction of a classical Markov chain model for which Eq. (1.1) would be correct. The numbers inside the circles are the states (counting the number of virus type) and the numbers above and below the arrows are the birth and death rates, respectively, of the state from which they come out.

Below, we present a correct proof of Lemma 3, [6] in the case of $\lambda = 1$, based on Lemmas 2.1 and 2.2. From this, the statement of Theorem 1 of [6] follows.

2. Auxiliary lemmas

Lemma 2.1. *Let $P_1(t) \equiv P(H_1 \leq t) \equiv F(t)$. Then $P_1(t)$ solves the renewal equation*

$$P_1(t) = \frac{2t}{(1+t)^3} + \int_0^t \frac{2}{(1+(t-y))^3} P_1(y)\, dy. \tag{2.1}$$

Proof. In panel A of Fig. 1, we can see a representation of the studied Markov chain on the state space $S = 1, 2, 3, \ldots$. Due to the H_is being independent for different is, we can study the distribution of H_1 and treat 1 as an absorbing state. Let $P_n(t)$ denote the probability of being in state n at time t, when one starts in state 2 at time 0. The system of differential equations describing the probabilities is

$$\begin{cases} P_1'(t) = 2P_2(t), \\ P_2'(t) = -4P_2(t) + 3P_3(t), \\ P_n'(t) = -2nP_n(t) + (n+1)P_{n+1}(t) + (n-1)P_{n-1}(t), \quad n \geq 3, \end{cases} \tag{2.2}$$

with initial conditions

$$P_1(0) = 0, \qquad P_2(0) = 1, \qquad P_3(0) = \ldots = P_n(0) = \ldots = 0. \tag{2.3}$$

Let $P(s,t)$ denote the generating function of the sequence $\{P_n(t)\}_{n=1}^{\infty}$, i.e.

$$P(s,t) = \sum_{n=1}^{\infty} P_n(t)s^n.$$

Taking first derivatives, we get the partial differential equation

$$\frac{\partial P(s,t)}{\partial t} = -(s-1)^2 P_1(t) + (s-1)^2 \frac{\partial P(s,t)}{\partial s}, \tag{2.4}$$

with initial conditions

$$\left.\frac{\partial P(s,t)}{\partial s}\right|_{\substack{s=0\\t=t}} = P_1(t), \quad P(s,t)|_{\substack{s=s\\t=0}} = \sum_{n=1}^{\infty} P_n(0)s^n = s^2. \tag{2.5}$$

Following §2.1 [2] and using the substitution $z(x) = P(h(s,x), t+x)$ with

$$h(s,x) := 1 + \frac{1}{x + \frac{1}{s-1}}, \qquad s \neq 1,$$

we arrive at

$$\begin{aligned} P(s,t) - \left(\frac{s-(s-1)t}{1-t(s-1)}\right)^2 &= z(0) - z'(-t) = \int_{-t}^{0} z'(x)\,dx \\ &= \int_0^t \frac{-1}{\left(y - t + \frac{1}{s-1}\right)^2} P_1(y)\,dy. \end{aligned}$$

Evaluating the derivative of both sides with respect to s at 0, we find that the function $P_1(t)$ must satisfy the following integral equation (we can recognize it as a renewal equation)

$$\frac{\partial P}{\partial s}(0,t) = P_1(t) = \frac{2t}{(1+t)^3} + \int_0^t \frac{2}{(1+(t-y))^3} P_1(y)\,dy. \tag{2.6}$$ ∎

Lemma 2.2. *$F(t) \equiv P_1(t)$, the solution of the renewal equation* (2.1)*, has the following properties*

$$\int_0^h tF'(t)\,dt \sim \log(h) \quad \text{as } h \to \infty, \tag{2.7}$$

$$\int_0^h t^2 F'(t)\,dt \sim h \quad \text{as } h \to \infty. \tag{2.8}$$

Proof. The proof is based on Tauberian theory and we refer the reader to [3, 4] for details on this. Another approach would be to study the asymptotic behaviour of $F'(t)$ by renewal theory results (see e.g. [1, 5]). The Laplace transform of a density function $f(x)$, denoted $\widehat{f(x)}(s)$ is

$$\widehat{f(x)}(s) = \int_0^{\infty} e^{-xs} f(x)\,dx.$$

We will use the following theorem from [3], Theorem 2 §XIII.5.

Theorem 2.3. (Theorem 2 §XIII.5, [3]) *If L is slowly varying at infinity and $0 \leq \rho < \infty$, then each of the relations*

$$\widehat{f(t)}(s) = \int_0^\infty f(t)\exp(-st)\,dt \sim s^{-\rho}L\left(\frac{1}{s}\right), \qquad s \to 0, \tag{2.9}$$

$$F(t) = \int_0^t f(u)\,du \sim \frac{1}{\Gamma(\rho+1)}t^\rho L(t), \qquad t \to \infty \tag{2.10}$$

implies the other.

$F(t) \equiv P_1(t)$ defined as the solution to the renewal equation (2.1), after differentiating will satisfy

$$F'(t) = \frac{2-4t}{(1+t)^4} + \int_0^t \frac{2}{(1+t-y)^3}F'(y)\,dy. \tag{2.11}$$

We calculate the Laplace transforms of $F'(t)$, $tF'(t)$ and $t^2F'(t)$

$$\widehat{F'(t)}(s) = \widehat{\frac{2-4t}{(1+t)^4}}(s) + \widehat{\frac{2}{(1+t)^3}}(s)\cdot\widehat{F'(t)}(s), \tag{2.12}$$

$$\begin{aligned}\widehat{tF'(t)}(s) = {} & \widehat{\frac{(2-4t)t}{(1+t)^4}}(s) + \widehat{\frac{2(t)}{(1+t)^3}}(s)\cdot\widehat{F'(t)}(s) \\ & + \widehat{\frac{2}{(1+t)^3}}(s)\cdot\widehat{tF'(t)}(s),\end{aligned} \tag{2.13}$$

$$\begin{aligned}\widehat{t^2F'(t)}(s) = {} & \widehat{\frac{(2-4t)t^2}{(1+t)^4}}(s) + \widehat{\frac{2(t)^2}{(1+t)^3}}(s)\cdot\widehat{F'(t)}(s) \\ & + \widehat{\frac{4t}{(1+t)^3}}(s)\cdot\widehat{tF'(t)}(s) + \widehat{\frac{2}{(1+t)^3}}(s)\cdot\widehat{t^2F'(t)}(s).\end{aligned} \tag{2.14}$$

We are interested in the behaviour of the transforms as $s \to 0$, and for this, we will use the well known property (verifiable by the de L'Hôspital rule) of the exponential integral

$$\int_s^\infty \frac{\exp(-u)}{u}\,du \sim -\log(s),$$

to arrive at

$$\widehat{tF'(t)}(s) = \frac{\widehat{\frac{(2-4t)t}{(1+t)^4}}(s) + \widehat{\frac{2t}{(1+t)^3}}(s)\cdot\widehat{F'(t)}(s)}{1-\widehat{\frac{2}{(1+t)^3}}(s)} \sim \log\left(\frac{1}{s}\right), \quad s \to 0, \tag{2.15}$$

$$\widehat{t^2F'(t)}(s) = \frac{\widehat{\frac{(2-4t)t^2}{(1+t)^4}}(s) + \widehat{\frac{2t^2}{(1+t)^3}}(s)\cdot\widehat{F'(t)}(s) + \widehat{\frac{4t}{(1+t)^3}}(s)\cdot\widehat{tF'(t)}(s)}{1-\widehat{\frac{2}{(1+t)^3}}(s)} \sim \frac{1}{s}, \quad s\to 0, \tag{2.16}$$

$$\widehat{t(1-F(t))}(s) = \frac{\widehat{\frac{t}{(1+t)^2}}(s) - \widehat{\frac{2t^2}{(1+t)^3}}(s) + \widehat{\frac{2t}{(1+t)^3}}(s)\cdot\widehat{(1-F(t))}(s)}{1-\widehat{\frac{2}{(1+t)^3}}(s)} \sim \frac{1}{s}, \quad s\to 0. \tag{2.17}$$

As both the constant function (s^0) and $\log(1/s)$ are slowly varying functions for $s\to 0$, the Tauberian theorem allows us to conclude that

$$\int_0^h tF'(t)\,dt \sim \log(h), \tag{2.18}$$

$$\int_0^h t^2F'(t)\,dt \sim h. \tag{2.19}$$ ∎

3. Proof of Lemma 3, [6]

We will now use Lemmas 2.1 and 2.2 to prove Lemma 3 from [6]. Define as there

$$m_n := \int_0^{\rho_n} tF'(t)\,dt$$

and

$$s_n := \int_0^{\rho_n} t^2F'(t)\,dt,$$

with $\rho_n := n\sqrt{\log(n)}$. By Lemma 2.2, we know that

$$m_n \sim \log(\rho_n) \sim \log(n) \text{ and } s_n \sim \rho_n.$$

We now need to check how $n(1-F(\rho_n))$ behaves asymptotically. We do not know what $F(\rho_n)$ is, but using the Tauberian theorem and Eq. (2.17) from the proof of Lemma 2.2, we get that

$$\int_0^t u(1-F(u))\,du \sim t, \qquad t\to\infty. \tag{3.1}$$

Therefore, using integration by parts and Eq. (2.19)

$$1 \sim \frac{\int_0^t u(1-F(u))du}{t} = \frac{\frac{t^2}{2}(1-F(t))}{t} + \frac{\int_0^t \frac{u^2}{2}F'(u)\,du}{t}, \qquad t\to\infty,$$

$$1 \sim \frac{1}{2}(t(1-F(t))) + \frac{1}{2}, \qquad t\to\infty,$$

and so we arrive at

$$\lim_{t\to\infty} t(1-F(t)) \to 1. \tag{3.2}$$

The rest of the proof is a direct repeat of the one in [6] and so we get (as in [6]) that $T_n/(n\log(n))$ tends to 1 in probability implying Theorem 1, [6] for $\lambda = 1$.

If we applied the same chain of reasoning to the model of panel B in Fig. 1, starting off with the system of differential equations, the analogue of Eq. 2.4) would be a homogeneous partial differential equation

$$\frac{\partial P(s,t)}{\partial t} = (s-1)^2 \frac{\partial P(s,t)}{\partial s}, \tag{3.3}$$

with initial conditions

$$\left.\frac{\partial P(s,t)}{\partial s}\right|_{s=0,\, t=t} = P_1(t), \qquad P(s,t)|_{s=s,\, t=0} = s, \tag{3.4}$$

in agreement with $P_1(t) = t/(t+1)$. It would therefore be an interesting problem to see what conditions are necessary on the nonhomogeneous part of Eq. (2.4) to still get the same asymptotic behaviour of the Markov chain and what underlying model properties do these conditions imply.

Acknowledgments. We are grateful to Wojciech Bartoszek and Joachim Domsta for many helpful comments, insights and suggestions. K. B. was supported by the Centre for Theoretical Biology at the University of Gothenburg, Stiftelsen för Vetenskaplig Forskning och Utbildning i Matematik (Foundation for Scientific Research and Education in Mathematics), Knut and Alice Wallenbergs travel fund, Paul and Marie Berghaus fund, the Royal Swedish Academy of Sciences, Wilhelm and Martina Lundgrens research fund and Östersjösamarbete scholarship from Svenska Institutet (00507/2012).

References

[1] P. Embrechts, E. Omey, *Functions of power series*, Yokohama Math. J. 32 (1984), 77–88.
[2] L. C. Evans, *Partial Differential Equations*, AMS, 1998.
[3] W. Feller, *An Introduction to Probability Theory and Its Applications*, Vol. 2, Wiley, New York, 1971.
[4] J. Korevaar, *Tauberian Theory: a Century of Developments*, Springer, 2004.
[5] T. M. Liggett, *Total positivity and renewal theory*, in Probablity, Statistics and Mathematics, Papers in Honor of Samuel Karlin, Academic Press, 1989.
[6] T. M. Liggett, R. B. Schinazi, *A stochastic model for phylogenetic trees*, J. Appl. Probab. 46 (2009), 601–607.

K. Bartoszek
MATHEMATICAL SCIENCES
CHALMERS UNIVERSITY OF TECHNOLOGY AND THE UNIVERSITY OF GOTHENBURG
412 96 GÖTEBORG, SWEDEN
E-mail: krzbar@chalmers.se

M. Krzemiński
DEPARTMENT OF PROBABILITY THEORY AND BIOMATHEMATICS
FACULTY OF APPLIED MATHEMATICS AND TECHNICAL PHYSICS
GDAŃSK UNIVERSITY OF TECHNOLOGY
80-233 GDAŃSK, POLAND
and
INSTITUTE OF MATHEMATICS
POLISH ACADEMY OF SCIENCES
00-956 WARSZAWA, POLAND
E-mail: mkrzeminski@mif.pg.gda.pl

Received March 7, 2012; revised version April 17, 2013.
Communicated by J. Wesołowski.