Modèles stochastiques d'évolution et d'interaction

Habilitation à diriger les recherches

soutenue le 12.12.2012

par

Jean BÉRARD

Composition du jury :

Francis COMETS (Université Denis Diderot - Paris 7), Rapporteur
Sylvie MÉLÉARD (École Polytechnique)
Thomas MOUNTFORD (École Polytechnique Fédérale de Lausanne)
Didier PIAU (Université Joseph Fourier - Grenoble 1)
Stéphane ROBIN (INRA - AgroParisTech), Rapporteur
Christophe SABOT (Université Claude Bernard - Lyon 1)
Vladas SIDORAVICIUS (IMPA), Rapporteur

69-2012

Contents

Remerciements

Je tiens tout d'abord à remercier Francis Comets, Stéphane Robin, et Vladas Sidoravicius, de l'intérêt qu'ils ont manifesté pour mon travail en acceptant d'être les rapporteurs de ce mémoire. Mes remerciements vont également à Sylvie Méléard, Thomas Mountford, Didier Piau et Christophe Sabot, qui ont bien voulu faire partie du jury.

Je tiens aussi à remercier mes collaborateurs – d'hier et d'aujourd'hui –, pour les aventures mathématiques vécues en commun, et pour leur précieuse amitié.

Bien d'autres personnes ont également contribué, par leur exemple, leurs encouragements ou leurs conseils, à l'élaboration des travaux présentés dans ce mémoire : qu'elles trouvent ici l'expression de ma chaleureuse gratitude.

Je remercie également mes collègues de l'université Claude Bernard - Lyon 1, notamment ceux de l'Institut Camille Jordan, ainsi que les collègues de l'Unité de Mathématiques Pures et Appliquées de l'ENS Lyon, pour l'ambiance tout à la fois stimulante et amicale dans laquelle j'ai eu la chance d'évoluer au cours de ces années passées à Lyon.

Enfin, mes pensées vont à ma famille, dont l'affection constante est le plus grand soutien qui puisse être.

Introduction

Ce document donne un aperçu des travaux de recherche que j'ai menés depuis ma thèse. *Grosso modo*, ceux-ci s'articulent autour des trois thèmes suivants :

- les modèles stochastiques d'évolution de séquences d'ADN,
- les méthodes de renouvellement,
- les modèles de branchement-sélection.

L'étude de l'évolution biologique au niveau moléculaire s'appuie fortement sur la modélisation mathématique et l'utilisation de méthodes statistiques. Dans ce cadre, l'un des problèmes est de décrire le processus de survenue des mutations dans les séquences d'ADN. Le chapitre 1 rend compte de travaux, effectués pour partie en collaboration avec J.B. Gouéré, L. Guéguen, A. Huet, et D. Piau, traitant de *modèles de substitution de nucléotides avec dépendance au contexte* : il s'agit de modèles stochastiques décrivant les substitutions de nucléotides survenant le long d'une séquence d'ADN, en prenant en compte la dépendance des taux de substitution en un site donné vis-à-vis de la composition des sites voisins. Les résultats décrits portent à la fois sur l'étude théorique des propriétés de ces modèles, et sur les méthodes statistiques développées pour les appliquer à des données génomiques.

Le principe des *méthodes de renouvellement* est d'analyser le comportement d'un modèle stochastique en identifiant, au sein de celui-ci, une structure séquentielle constituée de variables aléatoires indépendantes et identiquement distribuées. Elles constituent aujourd'hui un outil fondamental pour l'étude des marches aléatoires en milieu aléatoire (voir par exemple **[144, 143]**) et de certaines marches aléatoires en auto-interaction (voir par exemple **[92]**), et permettent d'obtenir des résultats caractérisant précisément le comportement asymptotique des modèles étudiés (caractère ballistique de la marche aléatoire, loi des grands nombres, comportement de type limite centrale, grandes déviations, etc.). Le chapitre 2 décrit des travaux effectués en collaboration avec A. Ramírez, dans lesquels une approche par renouvellement est utilisée pour étudier des modèles stochastiques microscopiques de propagation de front de type $X + Y \to 2X$, dans le cas unidimensionnel, conduisant à des résultats de grandes déviations et de fluctuations pour différents modèles. Le chapitre 3 décrit un autre volet de cette collaboration, consacré aux modèles de marche aléatoire excitée, pour lesquels l'approche par renouvellement nous a permis d'obtenir un résultat caractérisant les fluctuations en dimension $d \geq 2$.

Les *modèles de branchement-sélection* constituent une description extrêmement simplifiée des processus de sélection naturelle à l'œuvre dans l'évolution des organismes vivants. L'étude de ces modèles a notamment

pour but de comprendre, dans un cadre simple, l'influence de facteurs tels que la taille de la population, ou l'amplitude et la fréquence des mutations, sur la vitesse d'évolution. La classe de modèles que nous étudions se trouve être liée à une théorie plus générale décrivant la propagation de fronts stochastiques, développée par les physiciens théoriciens E. Brunet et B. Derrida. Le chapitre 4 décrit principalement des travaux effectués en collaboration avec J.B. Gouéré, dans lequels nous avons obtenu des preuves mathématiques confirmant certaines des prédictions de Brunet et Derrida.

Même si les thèmes de recherche décrits ci-dessus apparaissent dans des chapitres séparés, un certain nombre de liens existent entre eux, qui méritent d'être signalés. Par exemple, tant les systèmes de particules de type $X + Y \to 2X$ discutés au chapitre 2, que les modèles de branchement-sélection abordés au chapitre 4, peuvent être vus comme des modèles stochastiques de propagation de front, du type de ceux décrits par l'équation de Fisher-Kolmogorov-Petrovsky-Piscounov, mais dans des régimes asymptotiques différents : limite d'un grand nombre de particules par site pour le chapitre 4, ou limite d'un «petit» nombre pour le chapitre 2. D'autre part, même si les points de vue adoptés sont très différents, l'objectif commun des modèles étudiés aux chapitres 1 et 4 est de décrire mathématiquement certains aspects des processus qui gouvernent l'évolution du vivant. On peut également noter que les modèles de branchement-sélection sont à la base des algorithmes particulaires de type Monte-Carlo séquentiel discutés au chapitre 1. Enfin, de manière peut-être moins explicite, des propriétés de renouvellement sont utilisées de manière cruciale dans l'étude des modèles de branchement-sélection du chapitre 4 et des modèles de substitution du chapitre 1.

Par ailleurs, ce manuscrit n'aborde pas les travaux **[25, 20, 24]**, qui s'inscrivaient dans un projet de recherche différent, ayant connu quelques infortunes[1].

Chaque chapitre contient non seulement l'énoncé des résultats obtenus sur le sujet abordé, mais également une description du contexte scientifique, des principales idées utilisées dans les preuves, ainsi que des extensions possibles et de certaines perspectives de recherche future. Les résultats mathématiques sont énoncés de façon précise, mais un style moins formel est généralement employé dans le reste de la présentation, dans l'espoir de communiquer les idées essentielles tout en évitant le caractère quelquefois aride des articles dans lesquels se trouvent les résultats correspondants. Pour permettre une identification facile, les résultats que j'ai obtenus sont signalés par une couleur différente de celle employée dans le reste du texte. Enfin, dans la mesure du possible, j'ai tenté de conserver une certaine cohérence des notations entre ce manuscrit et les articles auxquels il se réfère; cependant, je n'ai pas hésité à effectuer certains changements de notations lorsque qu'il m'a semblé que la clarté y gagnait.

[1] Les travaux **[113]** et **[11]**, menés indépendamment, ont considérablement limité l'apport de **[20]** et **[24]**.

CHAPTER 1

Context-dependent nucleotide substitution models

1. Introduction

1.1. Molecular evolution. Broadly speaking, the goal of the scientific discipline known as *molecular evolution* (see e.g. **[84]**) is to study biological evolution at the molecular level. On the one hand, one wants to understand the rates and patterns of changes in DNA (or RNA) sequences and their products (proteins or RNA molecules), over evolutionary time. On the other hand, one wants to use molecular data to reconstruct the evolutionary history of biological entities, such as sequences, organisms or species. Both approaches are intimately related, and both heavily rely on mathematical modeling to make sense of the flood of data produced by modern sequencing technology. We refer to **[68, 125, 81, 82]** for surveys of the mathematical, statistical and computational approaches used in molecular evolution.

Let us briefly recall some of the key aspects of molecular sequence evolution. First, at the level of an individual, *mutations*, i.e. errors in either DNA replication or repair, may occur, leading to a sequence that is not exactly identical to the one from which it was copied. Such mutations include *substitutions*, i.e. the replacement of one nucleotide by another, *recombinations*, i.e. the exchange of a piece of sequence with another, *deletions*, i.e. the suppression of one or more nucleotides from the sequence, *insertions*, i.e. the addition of one or more nucleotides to the sequence, and *inversions*, i.e. the reversal of a piece of the sequence. When a mutation affects a germ cell (as opposed to somatic cell), it may be transmitted and survive through the generations, possibly up to the present. Over evolutionary time scales, survival depends on many factors, among which the possible advantage or disadvantage the mutated sequence may confer to individuals bearing it, and the size of the population within which it spreads (through breeding).

Accurately modeling each of these aspects of sequence evolution, in which complex patterns of variability in both time and space can play a role, is an extremely difficult task, and most models use simplifying assumptions and focus on a few specific aspects. The models we consider in the sequel describe the evolutionary process leading to a sample of DNA sequences, using the following elements:

- a tree which describes the ancestral relationships between the sequences in the sample, up to a hypothetical common ancestral sequence, see Fig. 1, and also Fig. 2;
- a Markov model describing the time-evolution of DNA sequences along the edges of the tree, starting from the ancestral sequence down to the sequences in the sample;
- a probabilistic model for the ancestral sequence.

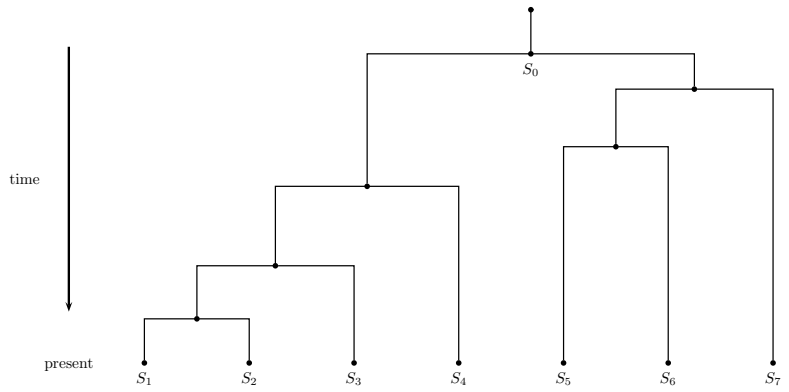


FIGURE 1. Hypothetical ancestral tree of a sample of DNA sequences $S_1, \ldots, S_7$. The ancestral sequence is denoted S_0. Edge lengths correspond to time durations.

Depending on the context, the branching events associated with internal nodes in the tree may simply represent one parent sequence giving rise to two offspring sequences, or to *speciation* events by which one species splits into two distinct species. In the latter case, the idea is that, over evolutionary time scales, one may neglect *polymorphism*, i.e. the existence of more than one version of the sequence within species.

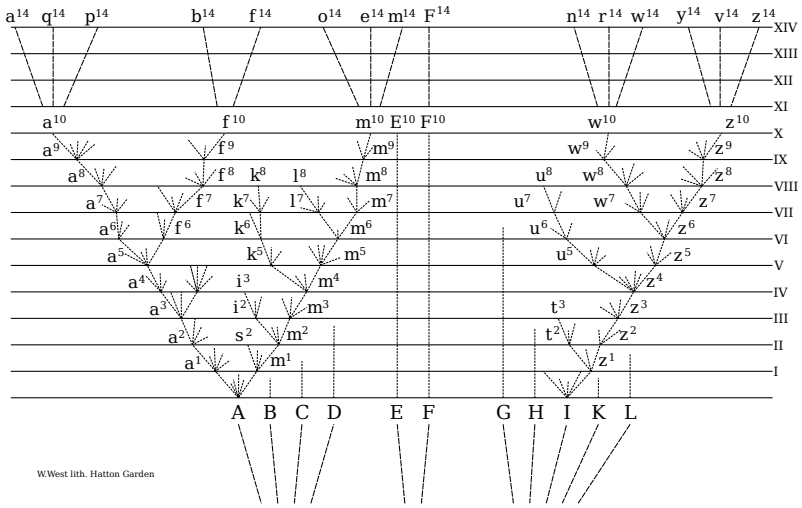
We also assume that, after a branching event, the two resulting sequences evolve independently, and that a single Markov model is appropriate to describe the sequence dynamics along the whole tree and across the whole sequence. Moreover, we neglect mutations other than substitutions, assuming the set of sequences in the sample to be *aligned*, so that nucleotide positions in the sequences that form the sample can directly be matched to those in the ancestral sequence.

1.2. Nucleotide substitution models. We now discuss more precisely the Markov models that are used to describe the evolution of DNA sequences. Since we restrict ourselves to nucleotide substitutions, we have to describe the rates at which single nucleotide changes occur along the sequence.

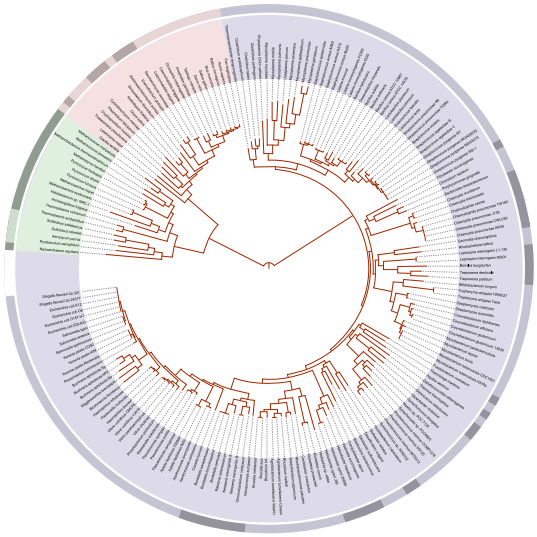
The most commonly used models describe the substitution process at a given site by a continuous-time Markov chain on the nucleotidic alphabet $\mathcal{A} := \{A, T, C, G\}$. Such a Markov chain is characterized by a 4×4 rate matrix, comprising 12 free parameters

$$(r_{u \to v},\ u, v \in \mathcal{A},\ u \neq v)\,,$$

describing the substitution rates between distinct elements of $\mathcal{A}$. Biological hypotheses (e.g. compatibility with DNA strand symmetry) are usually built



(a) The tree of life as it appears in Darwin's "On the Origin of Species" (1859).



(b) Tree of life from the Interactive Tree of Life website, inferred from molecular data consisting of completely sequenced genomes.

FIGURE 2. Two versions of the tree of life.

into the model to somehow constrain the parameters and have it reflect some relevant biological features. Starting from the early one-parameter model of Jukes and Cantor **[95]**, in which all the substitution rates are assumed to be equal, a whole hierarchy of models has been developed over the years (we refer e.g. to **[71]** for a survey of these models and of the corresponding assumptions). What all these models have in common is that the substitution process at a given site is assumed to be independent from substitution processes occurring at other sites. We refer to these models as being *site-independent*.

One well-known phenomenon illustrating the inadequacy of the site-independence assumption is the hypermutability of CpG dinucleotides[1], which is believed to be the most important context-dependent effect in mammalian genomes. Because of cytosine methylation, the substitution rate of a CpG dinucleotide into TpG or CpA is about 10 times higher than the overall substitution rate of a C into a T or a G into an A. A natural way of introducing such dependencies in the evolution model is to complement the basal rates $(r_{u\to v})$ with context-dependent (or neighbour-dependent) substitution rates of the form

$$(r_{uv\to uw},\ u,v,w\in\mathcal{A},\ v\neq w)$$

to account for the possible influence of the left neighbour, and

$$(r_{uv\to wv},\ u,v,w\in\mathcal{A},\ w\neq v)$$

to account for the possible influence of the right neighbour. Informally, this means that, at a site where a nucleotide b is surrounded by an a at its left and a c at its right, the probability that b is substituted by another nucleotide d during an infinitesimal interval of time dt, i.e. the probability of the substitution

$$\begin{array}{ll}\cdots abc\cdots & \text{time } t\\ \quad\downarrow & \\ \cdots adc\cdots & \text{time } t+dt\end{array}$$

is given by

$$r_{b\to d}dt + r_{ab\to ad}dt + r_{bc\to dc}dt,$$

where the first term represents the basal rate of $b\to d$ substitutions, and the second and third terms represent the additional influences of the left and right neighbours respectively. One might consider more general combined influences of left and right neighbours, or influences from next-to-nearest or even further neighbours, but, since our main motivation is to take into account CpG hypermutability, we restrict ourselves to models of the above form.

To properly define the corresponding dynamics, consider a state space $\mathcal{S}$ of the form $\mathcal{A}^J$, where either $J=\mathbb{Z}$ or is a finite sub-interval[2] of $\mathbb{Z}$ of the form $J=\{a,\dots,b\}$. Given $\eta\in\mathcal{S}$, $x\in J$, and $u,v\in\mathcal{A}$ such that $u\neq v$, let $\mathfrak{R}^x_{u\to v}(\eta)$ denote the element of $\mathcal{S}$ defined by

$$\left(\mathfrak{R}^x_{u\to v}(\eta)\right)(x) := \begin{cases} v \text{ if } \eta(x)=u,\\ \eta(x) \text{ otherwise,}\end{cases}$$

$$\left(\mathfrak{R}^y_{u\to v}(\eta)\right)(y) := \eta(y),\ y\neq x.$$

Then, given $u,v,w\in\mathcal{A}$ such that $v\neq w$, one defines $\mathfrak{R}^x_{uv\to uw}(\eta)$ by

$$\left(\mathfrak{R}^x_{uv\to uw}(\eta)\right)(x) := \begin{cases} w \text{ if } (\eta(x-1),\eta(x))=(u,v),\\ \eta(x) \text{ otherwise,}\end{cases}$$

$$\left(\mathfrak{R}^y_{uv\to uw}(\eta)\right)(y) := \eta(y),\ y\neq x.$$

[1]The notation XpX' is used for pairs of consecutive nucleotides where a X is followed by a X' in the sequence, e.g. CpG, TpA, etc.

[2]In this case, one has to specify boundary conditions to properly deal with influences from the left-neighbour (resp. right-neighbour) at site a (resp. site b). We do not discuss boundary conditions in detail here.

Similarly, one defines $\mathfrak{R}^x_{uv\to wv}(\eta)$ by

$$\left(\mathfrak{R}^x_{uv\to wv}(\eta)\right)(x) := \begin{cases} w \text{ if } (\eta(x),\eta(x+1)) = (u,v), \\ \eta(x) \text{ otherwise,} \end{cases}$$

$$\left(\mathfrak{R}^y_{uv\to wv}(\eta)\right)(y) := \eta(y),\ y \neq x.$$

The nucleotide substitution dynamics is then defined as a Markov process $(\eta_t)_{t\geq 0}$ on $\mathcal{S}$, through the action of its infinitesimal generator on functions depending on a finite number of coordinates (see e.g. **[115]**):

$$G\phi(\eta) = G_\emptyset\phi(\eta) + G_\ell\phi(\eta) + G_r\phi(\eta),$$

where

$$\begin{aligned} G_\emptyset\phi(\eta) &:= \sum_{x\in J}\sum_{\substack{u,v\in\mathcal{A}\\ u\neq v}} r_{u\to v}\left(\phi(\mathfrak{R}^x_{u\to v}(\eta)) - \phi(\eta)\right), \\ G_\ell\phi(\eta) &:= \sum_{x\in J}\sum_{\substack{u,v,w\in\mathcal{A}\\ v\neq w}} r_{uv\to uw}\left(\phi(\mathfrak{R}^x_{uv\to uw}(\eta)) - \phi(\eta)\right), \\ G_r\phi(\eta) &:= \sum_{x\in J}\sum_{\substack{u,v,w\in\mathcal{A}\\ v\neq w}} r_{uv\to wv}\left(\phi(\mathfrak{R}^x_{uv\to wv}(\eta)) - \phi(\eta)\right) \end{aligned} \tag{1}$$

where $G_\emptyset$ corresponds to the basal substitution rates, and G_ℓ and G_r give the respective contributions of influences from left and right neighbours.

At this level of generality, it seems unlikely that many interesting properties of the model can be established. For instance, in the absence of monotonicity, the possible propagation of influences from site to site makes it unclear whether the corresponding model is ergodic under the non-degeneracy assumption that the basal rates are positive (see **[76, 85]** for a discussion of this issue in the more general context of one-dimensional interacting particle systems). Our own work revolves around a special class of nucleotide substitution models, called RN+YpR, which is made tractable by special structural properties it possesses.

1.3. RN+YpR nucleotide substitution models. The class of substitution models we consider, called RN+YpR, makes specific assumptions about the substitution rates allowed in the model. To state them, let us remember that adenine (A) and guanine (G) are *purines*, generically denoted R, while thymine (T) and cytosine (C) are *pyrimidines*, generically denoted Y. A substitution that changes a purine into another purine, or a pyrimidine into another pyrimidine, is called a *transition*. On the other hand, a substitution that changes a purine into a pyrimidine or vice-versa is called a *transversion*.

The first assumption (RN) is that the family of basal rates

$$(r_{u\to v},\ u,v\in\mathcal{A},\ u\neq v)$$

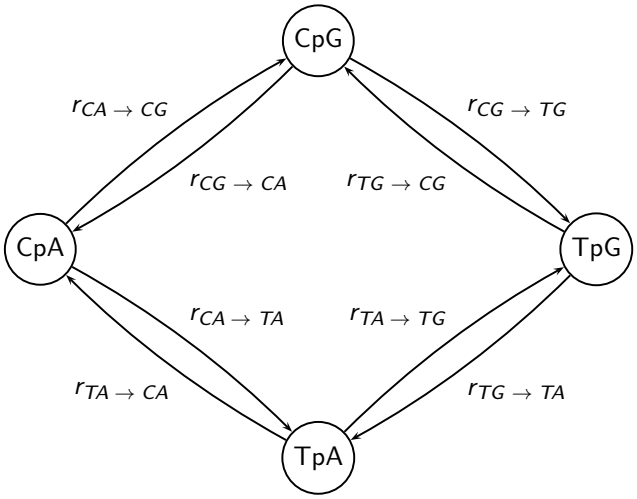


FIGURE 3. The 8 context-dependent substitutions allowed in a YpR model

satisfies the conditions defined by Rzhetsky and Nei **[139]** (whence the name "RN"), that is, the corresponding $\mathcal{A}\times\mathcal{A}$ matrix must have the following form

$$\begin{array}{c} \\ A \\ T \\ C \\ G \end{array} \begin{array}{c} \begin{array}{cccc} A & T & C & G \end{array} \\ \begin{pmatrix} - & v_T & v_C & w_G \\ v_A & - & w_C & v_G \\ v_A & w_T & - & v_G \\ w_A & v_T & v_C & - \end{pmatrix} \end{array}.$$

The above matrix is characterized by 8 free parameters (instead of 12 for the most general model), and reflects the assumption that the rate of a transversion resulting in a given nucleotide u depends only on u, and not on the nucleotide that has just been substituted.

The second assumption (YpR) is that the only non-zero context-dependent rates allowed in the model are those that turn a dinucleotide of the form (pyrimidine, purine), i.e. YpR, into another dinucleotide of the same form (whence the name "YpR" for the context-dependent part of the model). The 8 corresponding context-dependent substitutions are depicted in Fig. 3.

The combination of a family of basal rates satisfying assumption (RN) and of context-dependent rates satisfying assumption (YpR) defines the RN+YpR class of nucleotide substitution models. Note that the hypermutability of CpG dinucleotides is covered by these assumptions, since both CpG $\to$ CpA and CpG $\to$ TpG substitutions are of the form YpR $\to$ YpR.

2. Structural properties of RN+YpR models I

Let ρ denote the application which fuses the two purines together, and η the application which fuses the two pyrimidines together, that is

$$\rho(A) := R =: \rho(G),\ \rho(C) := C,\ \rho(T) := T,$$

and

$$\eta(A) := A,\ \eta(G) := G,\ \eta(C) := Y =: \eta(T).$$

Given a finite word $u_1 \dots u_m$ written in the nucleotidic alphabet, with $m \geq 2$, define the corresponding $\Phi-$encoding by

$$\Phi(u_1 \dots u_m) := \eta(u_1)u_2 \dots u_{m-1}\rho(u_m).$$

Given a finite interval $K := \{a, \dots, b\} \subset J$, define also π_K as the canonical projection[3] from $\mathcal{S} = \mathcal{A}^J$ to $\mathcal{A}^K$, i.e.

$$\pi_K((u_j)_{j\in J}) := (u_k)_{k\in K}.$$

THEOREM 1 (B., Gouéré, Piau [29]). *Given a family of pairwise disjoint intervals $K_1, \dots, K_m \subset J$, and given a fixed initial configuration $\eta_0 \in \mathcal{S}$, the random processes*

$$[\Phi(\pi_{K_n}(\eta_t))]_{t\geq 0},\ n = 1, \dots, m,$$

are independent Markov processes.

Note that, due to the translation-invariance of the infinitesimal generator G, the infinitesimal generator of the Markov process $\Phi(\pi_K(\eta_t))_{t\geq 0}$ depends only on $|K|$. We denote by $Q_{|K|}$ the corresponding infinitesimal generator on the set

$$\mathcal{A}_{|K|} := \{C, T, R\} \times \{A, T, C, G\}^{|K|-2} \times \{A, G, Y\}. \tag{2}$$

To explain Theorem 1, consider the evolution of a Φ-encoded polynucleotide $Z_t := \Phi(\eta_a(t) \dots \eta_b(t))$. We want to understand why the instantaneous rates associated with substitutions that modify the value of Z_t can all be expressed as functions of Z_t. Consider site a. By the YpR assumption, context-dependent substitutions at site a involving $\eta_{a-1}(t)$ can only occur if $\eta_a(t) \in \{A, G\}$, in which case the resulting nucleotide will also belong to $\{A, G\}$, so the value of $\rho(\eta_a(t))$, hence of Z_t, is not affected by such a substitution. As for basal substitutions, the (RN) assumption shows that, knowing only $\rho(\eta_a(t))$, one can deduce the rates of substitutions to C and T. Now consider site $a + 1$. Since the only possible context-dependent substitutions at site $a + 1$ involving $\eta_a(t)$ occur when $\eta_a(t) \in \{C, T\}$, we see that knowing $\rho(\eta_a(t))$ is enough to compute the corresponding instantaneous rates. A symmetric argument can be made at sites b and $b - 1$ for context-dependent substitutions involving right neighbours, and, finally, sites $a + 2 \leq x \leq b - 2$ do not raise any problem since the values of $\eta_{x-1}(t)$, $\eta_x(t)$ and $\eta_{x+1}(t)$ are part of Z_t.

To state the next result, we make the non-degeneracy assumption that all the basal rates are positive, i.e.

$$\forall u, v, \in \mathcal{A},\ u \neq v,\ r_{u\to v} > 0. \tag{3}$$

[3]In the sequel, we often identify the sequence $(u_k)_{k\in K}$ with the word obtained by concatenating the successive terms of the sequence, so that $\pi_K((u_j)_{j\in J})$ may also be viewed as a word written in the nucleotidic alphabet.

THEOREM 2 (B., Gouéré, Piau [**29**]). *Under assumption* (3)*, the process is ergodic, i.e. there exists a unique invariant distribution* μ *on* $\mathcal{S}$ *such that, for every initial configuration* η_0*, one has that, as* t *goes to infinity,*

$$\eta_t \xrightarrow{d} \mu. \tag{4}$$

Moreover, one has an exponential bound on the speed of convergence towards the distribution μ*: there exist explicit constants* $c_1, c_2 > 0$ *such that, for all* $K \subset J$*,*

$$d_{TV}\left(\pi_K(\eta_t), \pi_K(\mu)\right) \leq \exp(-c_1 t + c_2 \log|K|). \tag{5}$$

Finally, the image of μ *by* π_K *can be sampled perfectly by means of an efficient Propp-Wilson type algorithm.*

The proof of Theorem 2 is obtained through a suitable graphical construction of the dynamics based on marked Poisson processes (see [**115, 116**]). In view of Theorem 1, it is not too surprising that, within this construction, the evolution of a $\Phi-$encoded polynucleotide $\pi_K(\eta_t)$ is a function only of the Poisson processes attached to the sites in K. Thus, to control the evolution of the nucleotide at site x, one has to deal only with the graphical construction at sites $x-1, x, x+1$, for which one can define coupling times with exponential tail decay, leading to the proof of Theorem 2. Doing the graphical construction in the past leads to the Propp-Wilson type perfect simulation algorithm. This is described in more detail in Section 5 of the present chapter.

Another interesting consequence of Theorem 1 is the following.

THEOREM 3 (B., Gouéré, Piau [**29**]). *The finite marginals of* μ *solve explicit finite-size linear systems.*

Theorem 3 is just a consequence of the fact that the time-evolution of a $\Phi-$encoded polynucleotide can be described by a finite-state continuous-time Markov chain: the $k-$dimensional marginal of μ can be obtained by computing the stationary distribution of a $\Phi-$encoded polynucleotide of length $k+2$. As an example, we explicitly computed (with the help of a symbolic computation software) the $2-$dimensional marginals of μ in the simplest case where all basal rates $r_{u\to v}$ are equal to 1, and where the only non-zero context-dependent rates are $r_{CG\to CA} = r_{CG\to TG} = r$. Here is the list of all

the stationary frequencies[4] of dinucleotides starting with an A:

$$\begin{aligned} F(AA) &= \frac{1}{16}\left(1+\frac{r}{32+10r}\left(3+\frac{3r}{96+19r}\right)\right), \\ F(AC) &= \frac{1}{16}\left(1+\frac{r}{32+10r}\left(0-\frac{4r}{32+10r}\right)\right), \\ F(AG) &= \frac{1}{16}\left(1+\frac{r}{32+10r}\left(1-\frac{3r}{96+19r}\right)\right), \\ F(AT) &= \frac{1}{16}\left(1+\frac{r}{32+10r}\left(4+\frac{4r}{32+10r}\right)\right). \end{aligned}$$

Similar formulas are available for the 12 other dinucleotides. Note that, since for general k the corresponding linear systems are of size $(9\cdot 4^k)\times(9\cdot 4^k)$, it is unrealistic to use this approach for k larger than, say 6 or 7 (depending also on whether one wants a symbolic or numeric solution), although some tricks can be devised to alleviate the computational burden.

Theorem 1 imposes strong independence properties on μ. First, μ is $2-$dependent, meaning that, if $Z\sim\mu$, $(Z_x)_{x\in K_1}$ and $(Z_x)_{x\in K_2}$ are independent as soon as $d(K_1,K_2)>2$. Moreover, μ is a $3-$factor, meaning that we can find i.i.d. random variables $(\gamma_x)_{x\in\mathbb{Z}}$ (the marked Poisson processes attached to the sites of $\mathbb{Z}$ in the graphical construction) and a measurable function f such that the sequence $(f(\gamma_{x-1},\gamma_x,\gamma_{x+1}))_{x\in\mathbb{Z}}$ has distribution μ. Given such properties, one may be tempted to deduce that the dependence structure of μ must be Markov of order, say 1 or 2, which is not true. We do not provide a proof here, but rather give a caricatural example which illustrates why such a deduction is erroneous in general.

Let $(\varepsilon_x)_{x\in\mathbb{Z}}$ denote an i.i.d. sequence of symmetric Bernoulli random variables, and let $\zeta_x:=g(\varepsilon_x,\varepsilon_{x+1})$, where $g(0,1)=g(1,0)=0$, $g(0,0)=1$, $g(1,1)=2$. By construction, $(\zeta_x)_{x\in\mathbb{Z}}$ is a $2-$factor, but it is easy to check that, for all $n\geq 1$, conditional upon $(\zeta_{-2n},\dots,\zeta_{-1})=(1,0,\dots,0)$, the distribution of ζ_0 is $\frac{1}{2}\delta_0+\frac{1}{2}\delta_2$, while, conditional upon $(\zeta_{-2n},\dots,\zeta_{-1})=(0,0,\dots,0)$, the distribution of ζ_0 is $\frac{1}{2}\delta_0+\frac{1}{2}\delta_1$. This shows that $(\zeta_x)_{x\in\mathbb{Z}}$ cannot be a Markov chain, of any order.

Before we end this section, let us mention the works **[70, 15]**, in which the structural properties of RN+YpR models are used to study distance statistics between sequences evolving from a common ancestral sequence.

3. Statistical inference I

3.1. Inference with nucleotide substitution models. Let us define a little more precisely the statistical framework we are working with. The data consist of a (finite) sample of DNA sequences $(s_\alpha,\ \alpha\in\mathscr{S})$, where each s_α is an element of $\mathcal{S}=\mathcal{A}^J$. On the other hand, the model is specified by:

[4]Due to the invariance of the model with respect to space-translations, μ is also translation-invariant, and one can unambiguously define the stationary frequency $F(u_1\dots u_m)$ of a given word with respect to μ. It is also easily checked (e.g. by Theorem 1) that μ is ergodic with respect to space-translations, so that $F(u_1\dots u_m)$ also corresponds to the asymptotic empirical frequency of $u_1\dots u_m$ with respect to μ.

- a finite, rooted, bifurcating tree $\mathscr{T}$, whose leaves are identified with $\mathscr{S}$; to each edge (α, β) of $\mathscr{T}$ such that α is the parent of β (denoted $\beta \leftarrow \alpha$) is attached a positive number $t_{\alpha,\beta}$ counting the evolutionary time from α to β;
- families of basal and context-dependent rates $(r_{u\to v})$, $(r_{uv\to uw})$ and $(r_{uv\to wv})$
- an initial distribution ν on $\mathcal{S}$ describing the sequence at the root.

Of course, one is interested in inferring the model from the data, but several scenarios exist. For instance, in **[30]**, the topology of $\mathscr{T}$ is given, and the main focus is on the estimation of numerical parameters (rates and branch lengths). In other situations, one may be mainly interested in the topology of $\mathscr{T}$, with numerical parameters being subsidiary, or in goodness-of-fit comparisons between distinct classes of models. Note that we do not assume an a priori probabilistic model on $\mathscr{T}$.

Among the inference methods, maximum likelihood (including variations such as the expectation-maximization (EM) approach) plays a central role, along with bayesian strategies. In the sequel, we mostly discuss maximum likelihood, which is the method used in **[30]**.

Let us collectively denote the components of the model (tree, rates, ancestral sequence distribution) by θ, and use the notation $\mathbb{P}_\theta$ for the corresponding probability measure, which describes the evolutionary process leading from the ancestral sequence to the sequences in the sample. Define S_α as the random variable corresponding to the sequence attached to node α in the model. The likelihood of the sample of DNA sequences $(s_\alpha,\ \alpha \in \mathscr{S})$ with respect to the model θ is then defined by

$$L = L((s_\alpha)_{\alpha\in\mathscr{S}}|\theta) := \mathbb{P}_\theta(S_\alpha = s_\alpha,\ \alpha \in \mathscr{S}). \tag{6}$$

3.1.1. *Site-independent models.* For site-independent models, computing the likelihood of a sample of sequences with respect to the model can be done very efficiently, using a dynamic programming algorithm known as *Felsenstein's tree-pruning algorithm* (see e.g. **[81]**), which recursively computes likelihood values associated with the nodes of the tree, starting at the leaves and ending at the root. In accordance with the site-independence assumption, the distribution of the ancestral sequence is assumed to be of a product form, i.e. $\nu = \nu_0 \otimes \cdots \otimes \nu_0$, where ν_0 is a probability measure on $\mathcal{A}$.

Felsenstein's algorithm goes as follows. To each node α of the tree $\mathscr{T}$, and each site $x \in J$, is associated the map $L_{x,\alpha}$ from $\mathcal{A}$ to $\mathbb{R}$ defined by

$$L_{x,\alpha}(u) := \mathbb{P}_\theta(S_\beta = s_\beta,\ \beta \dashleftarrow \alpha,\ \beta \in \mathscr{S}|S_\alpha(x) = u),$$

where $\beta \dashleftarrow \alpha$ means that β is a descendant (but not necessarily a child) of α. One then has the key recursion identity:

$$L_{x,\alpha} := \prod_{\beta\leftarrow\alpha} e^{t_{\alpha,\beta}G_0} \times L_{x,\beta}, \tag{7}$$

where G_0 is the infinitesimal generator on $\mathcal{A}$ describing the Markov dynamics *of a single site*. The product $\times$ denotes the action of the corresponding Markov semi-group on real-valued functions defined on $\mathcal{A}$, which in practice means a matrix-vector product. The recursion (7) is initialized at the leaves

of $\mathscr{T}$ by

$$L_{x,\alpha}(\cdot) := \mathbf{1}_{s_\alpha(x)}(\cdot). \tag{8}$$

Finally, the likelihood for site x is obtained through

$$L_x := \sum_{u \in \mathcal{A}} \nu_0(u) L_{x,root}(u), \tag{9}$$

and the overall likelihood through

$$L := \prod_{x \in J} L_x. \tag{10}$$

Given the ability to compute L with Felsenstein's algorithm, maximum likelihood inference of the model can be performed. Moreover, the corresponding statistical framework is the simplest and best understood one (see e.g. **[148]**), since the sequence of observations forms an i.i.d. sequence under the model.

3.1.2. *Context-dependent models.* In the case of context-dependent substitution models, it is no longer true that the likelihood can be written under the product form (10). One could still consider a tree recursion such as (7), exploiting the fact that sequence evolution is a Markov process on $\mathcal{A}^J$. This would mean working with the full generator G (see (1)) on the set $\mathcal{A}^J$ whose cardinality is $4^{|J|}$, instead of G_0. Unfortunately, numerical computation of the exponential of a $4^{|J|} \times 4^{|J|}$ matrix is not feasible unless $|J|$ is less than, say 6 or 7, so this direct approach cannot be used for real data, where $|J|$ typically varies from a few hundreds to a few millions.

A first alternative approach consists in approximating the dependence structure induced by the model, using various ways of neglecting dependencies between non-neighbouring sites – which are generally expected to be small – to make computations tractable. A surrogate for the original likelihood of the model can thus be obtained and exploited within a maximum likelihood or a bayesian framework, see e.g. **[67, 5, 141, 117, 50, 49]**. This approach in general leads to computationally efficient algorithms, but the reliability of the corresponding approximations is usually difficult to assess other than empirically.

Another approach consists in using inference techniques developed for models with latent (unobserved) variables. Here, the set of latent variables corresponds to the full substitution history leading from the ancestral sequence to the sequences in the sample.

One key observation is that, due to the nearest-neighbour dependence of substitution rates in the model, the collection of substitution histories at sites $x \in J$ possesses an explicit Markov random field structure. To state this more precisely, introduce the notation $S_{\alpha,\beta,t}$ to denote the sequence that has evolved for t units of time along the tree from α to β. The full substitution history $\mathscr{H}(x)$ at site x is then defined by

$$\mathscr{H}(x) := \left(S_{\alpha,\beta,t}(x);\ \beta \leftarrow \alpha,\ 0 \le t \le t_{\alpha,\beta}\right).$$

The Markov random field property of $\mathscr{H}$ corresponds to the fact that $\mathscr{H}(x)$ depends on the histories at sites distinct from x only through the histories at neighbouring sites, which we write slightly informally as

$$dis.\left(\mathscr{H}(x) | \mathscr{H}(y);\ y \in J \setminus \{x\}\right) = dis.\left(\mathscr{H}(x) | \mathscr{H}(y);\ |y - x| = 1, 2\right). \tag{11}$$

Moreover, the conditional distribution of $\mathscr{H}(x)$ given the histories at neighboring sites can itself be described explicitly as a time-inhomogeneous Markov process running along the edges of the tree $\mathscr{T}$.

Thanks to this structure, it is possible to use Markov Chain Monte Carlo techniques such as Gibbs sampling, to sample from the conditional distribution of $(\mathscr{H}(x),\ x \in J)$ given the data, i.e. given $S_\alpha = s_\alpha,\ \alpha \in \mathscr{S}$. In this setting, one can then apply inference methods developed for latent variable models, such as Monte Carlo EM, or bayesian strategies, see e.g. **[137]**. This approach is used in several works (in discrete or continuous-time settings), see e.g. **[127, 91, 88, 66, 10]**.

One advantage of such methods is that the context-dependent dynamics is faithfully reflected instead of replaced by an approximation whose accuracy is difficult to assess. On the other hand, they lead to computer-intensive algorithms, for which convergence is an issue.

3.2. Inference for RN+YpR models (B. and Guéguen [30]).

3.2.1. *Theory.* In **[30]**, we developed a maximum-likelihood type approach[5] to inference, that exploits the specific properties of RN+YpR models.

Assuming (for the sake of simplicity) that $J = \{1, 2, \dots, 3q+2\}$ for some integer q, consider the division of the sequence $\eta_t \in \mathcal{A}^J$ into non-overlapping $\Phi-$encoded tri-nucleotides

$$\begin{array}{ccccc} \underbrace{\eta_t(1)\eta_t(2)\eta_t(3)} & \underbrace{\eta_t(4)\eta_t(5)\eta_t(6)} & \dots & \underbrace{\eta_t(3q-2)\eta_t(3q-1)\eta_t(3q)} \\ \Downarrow & \Downarrow & & \Downarrow \\ \Phi(\eta_t(1)\eta_t(2)\eta_t(3)) & \Phi(\eta_t(4)\eta_t(5)\eta_t(6)) & \cdots & \Phi(\eta_t(3q-2)\eta_t(3q-1)\eta_t(3q)) \end{array}$$

Thanks to Theorem 1, we see that these $\Phi-$encoded trinucleotides evolve as independent Markov processes on the alphabet $\mathcal{A}_3$, with infinitesimal generator Q_3, see (2). As a consequence, letting

$$Y^0_\alpha(k) := \Phi(S_\alpha(3k-2)S_\alpha(3k-1)S_\alpha(3k)),$$

$$y^0_\alpha(k) := \Phi(s_\alpha(3k-2)s_\alpha(3k-1)s_\alpha(3k)),$$

we can map our our sample $(s_\alpha,\ \alpha \in \mathscr{S})$ of DNA sequences of length $|J| = 3q$, onto a sample $(y^0_\alpha,\ \alpha \in \mathscr{S})$ of sequences of length q, written in the alphabet $\mathcal{A}_3$, whose past evolutionary history is described by a *site-independent* Markov model with infinitesimal generator Q_3 evolving along the tree $\mathscr{T}$. The corresponding likelihood

$$L^0_3(y^0_\alpha,\ \alpha \in \mathcal{S}|\theta) := \mathbb{P}_\theta(Y^0_\alpha = y^0_\alpha,\ \alpha \in \mathcal{S}),$$

can then be computed using Felsenstein's algorithm, using the alphabet $\mathcal{A}_3$ instead of $\mathcal{A}$ and generator Q_3 instead of G_0. Note that, to make the above approach work, we also have to assume that the distribution ν of the ancestral sequence is translation-invariant, and that, with respect to ν, $\Phi-$encoded polynucleotides with pairwise disjoint supports are independent. This is automatically the case if ν is e.g. assumed to be the stationary distribution of a RN+YpR model. In the sequel, we call ν_3 the distribution on $\mathcal{A}_3$ corresponding to a $\Phi-$encoded trinucleotide.

[5] More precisely, our approach belongs to the class of *composite likelihood methods* (see **[150, 151]** for general references on the subject).

On the practical side, the mapping we have just described makes it possible to recycle existing maximum-likelihood inference algorithms, developed for site-independent models, with an alphabet size now $|\mathcal{A}_3| = 3 \times 4 \times 3 = 36$ instead of just $|\mathcal{A}| = 4$, whence an additional but manageable computational cost. On the theoretical side, this approach automatically inherits the properties of maximum likelihood inference based on sequences of independent identically distributed observations, such as asymptotic consistency and efficiency (see e.g. **[148]**).

However, mapping the original sequence alignment data onto the family of $\Phi-$encoded trinucleotides $y^0_\alpha(k)$ leads to a substantial loss of information, since no distinction is made between As and Gs at sites $i = 3k-2$, or between Cs and Ts at sites $i = 3k$.

To make more efficient use of the information contained in the data, one should take into account not just one, but all three possible phases associated with the division of the sequence into non-overlapping trinucleotides. More precisely, one can generalize the definition of y^0_α and Y^0_α by letting, for $p = 1, 2$,

$$Y^p_\alpha(k) := \Phi(S_\alpha(3k-2+p)S_\alpha(3k-1+p)S_\alpha(3k+p)),$$
$$y^p_\alpha(k) := \Phi(s_\alpha(3k-2+p)s_\alpha(3k-1+p)s_\alpha(3k+p)).$$

Each $p = 0, 1, 2$ gives rise to a different likelihood, denoted L^p, defined by

$$L^p_3(y^p_\alpha,\ \alpha \in \mathcal{S}|\theta) := \mathbb{P}_\theta(Y^p_\alpha = y^p_\alpha,\ \alpha \in \mathcal{S}).$$

The approach used in **[30]** *consists in performing maximum-likelihood estimation with respect to the averaged log-likelihood* ℓ_3 defined by

$$\ell_3 := \tfrac{1}{3}\left(\log L^0_3 + \log L^1_3 + \log L^2_3\right). \tag{12}$$

On a purely practical side, maximum-likelihood inference with respect to ℓ_3 can still be carried out by recycling[6] algorithms devised for site-independent models, even though overlapping $\Phi-$encoded trinucleotides generally *do not* evolve independently under an RN+YpR model.

On the other hand, the use of ℓ_3 has to be given a theoretical justification, since it no longer corresponds to the likelihood associated with a sequence of i.i.d. random variables, due to the dependence between distinct phases[7]. In what follows, we sketch a justification based on asymptotic theory, assuming a sample of infinitely long DNA sequences (indexed by $\mathbb{Z}$), of which we study a larger and larger part as $J \nearrow \mathbb{Z}$.

Let us use a J subscript to indicate dependence on J in the following, and assume that the "true" evolutionary model describing the sequences corresponds to the value $\theta = \theta_0$. For $p = 0, 1, 2$, let $\hat{\ell}^p_{3,J}$ denote the random variable obtained by computing $\log L^p_{3,J}$ using the random variables $Y^p_{\alpha,J}$ as the input data:

$$\hat{\ell}^p_{3,J}(\theta) := \log L^p_{3,J}(Y^p_{\alpha,J}, \alpha \in \mathcal{S}|\theta).$$

[6]More specifically, maximum likelihood inference with respect to ℓ_3 can be achieved by applying a maximum likelihood algorithm that treats the whole family of (overlapping) $\Phi-$encoded trinucleotides as independently evolving sites governed by the infinitesimal generator Q_3.

[7]Note that ℓ_3, being based on mappings of the original sample, which erase some of the information contained in the data, has no reason to be equal to the true likelihood L.

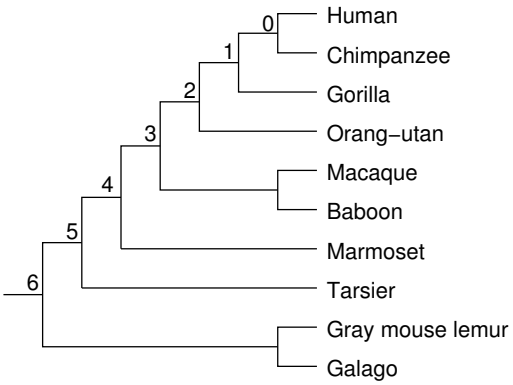


FIGURE 4. Topology of the phylogenetic tree associated with the 10 species of primates used in the alignment.

Using i.i.d.ness of the evolution of disjoint $\Phi-$encoded tri-nucleotides, the law of large numbers entails that, for $p = 0, 1, 2$, one has a.s. that

$$\lim_{J \nearrow \mathbb{Z}} |J|^{-1} \hat{\ell}^p_{3,J}(\theta) = \ell^*_3(\theta), \tag{13}$$

where $\ell^*_3(\theta)$ is the expected log-likelihood of the model describing a *single* $\Phi-$encoded trinucleotide evolving along $\mathscr{T}$ according to the infinitesimal generator Q_3 and starting with the ancestral sequence distribution ν_3 specified by θ_0, i.e.

$$\ell^*_3(\theta) := \mathbb{E}_{\theta_0}(\log L^0_{3,J_0}(Y^0_\alpha(1), \alpha \in \mathcal{S}|\theta)),$$

where $J_0 = \{1, 2, 3\}$.

The limit (13) is the key to the consistency of the maximum likelihood method, since, provided that the model is identifiable[8], θ_0 is the unique value of θ at which $\ell^*_3(\cdot)$ attains its maximum.

Now, exploiting mixing properties in a similar way as **[70]**, one can prove that, as $J \nearrow \mathbb{Z}$, the distribution of the vector

$$|J|^{-1/2} \left(\hat{\ell}^p_{3,J}(\theta) - \ell^*_3(\theta) \right)_{p=0,1,2}$$

is that of a tri-dimensional centered normal vector $Z = (Z^0, Z^1, Z^2)$ whose covariance matrix is of the form

$$\mathrm{cov}(Z) = \begin{pmatrix} v & c & c \\ c & v & c \\ c & c & v \end{pmatrix}.$$

As a consequence, the definition of ℓ_3 through (12) provides (in the sense of asymptotically unbiased with minimum variance) the optimal way of combining the three log-likelihood values ℓ^p_3, $p = 0, 1, 2$ to produce an estimate of ℓ^*_3.

3.2.2. *Applications.* We applied our method on a data set consisting of a portion of the human genome comprising 1,877,425 bases, aligned with genomes from nine other species of primates, see Fig. 4.

[8]We do not discuss identifiability issues here, which turn out to be non-trivial, and refer to **[30]** for references. Just note that only ν_3 can be identified here, not ν.

As a first application, we performed maximum likelihood computations with ℓ_3 for various substitution models with (+CpG) and without (+0) a context-dependent substitution rate $r = r_{CG\to CA} = r_{CG\to TG}$ modeling CpG hypermutability. The model describing basal (non-context-dependent) rates goes from the simplest (Jukes-Cantor, denoted JC69) up to the most general (Rzhetsky-Ney, denoted RN95) allowed in the RN+YpR class. The results are shown in Table 1. One observes, among other things, that including a context-dependent substitution rate always improves the fit in a dramatic way. Moreover, the very basic K80+CpG model outperforms all models with no context-dependent rates. Note that brutal comparison of likelihood values between distinct models goes against good statistical practice. Still, as is often the case when performing comparisons of nucleotide substitution models on large datasets, the likelihood differences are so large that using any of the usual criteria (classical likelihood ratio test for nested models, AIC or BIC), would not alter the result of the comparison.

Model	np	$\ell_3 - \ell_3^{\mathrm{JC}}$	ρ
JC69+CpG	1	132534.1	25.078
K80+0	1	118040.5	
K80+CpG	2	226295.3	12.719
T92+0	2	164839.6	
T92+CpG	3	234796.9	9.916
HKY85+0	4	164874.3	
HKY85+CpG	5	234829.6	9.916
TN93+0	5	164901.1	
TN93+CpG	6	234861.6	9.917
RN95s+0	3	169535.9	
RN95s+CpG	4	237311.3	9.242
RN95+0	7	169596.7	
RN95+CpG	8	237375.7	9.242

TABLE 1. Maximum ℓ_3 values for various models. Shown are the differences $\ell_3 - \ell_3^{\mathrm{JC}}$, where ℓ_3^{JC} is the value obtained for the basic Jukes-Cantor (JC69+0) model. Also shown are the estimated values of the normalized CpG hypermutability rate $\rho = r_{CG\to CA}/r_{C\to T} = r_{CG\to TG}/r_{C\to T}$, with various substitution models (np: number of free parameters).

We then performed an estimation of the CpG hypermutability rate along the human sequence in our dataset. Since CpG hypermutability is a consequence of methylation, a sufficiently low local value of this rate should indicate the presence of what biologists call a *hypomethylated island*. A more usual criterion is based on the so-called CpGo/e ratio

$$\mathrm{CpGo/e} = \frac{\text{frequency of CpG}}{\text{frequency of C} \ \times \text{frequency of G}}.$$

Note that the CpGo/e ratio is based solely on the composition of the human sequence, while our approach exploits the whole phylogenetic information

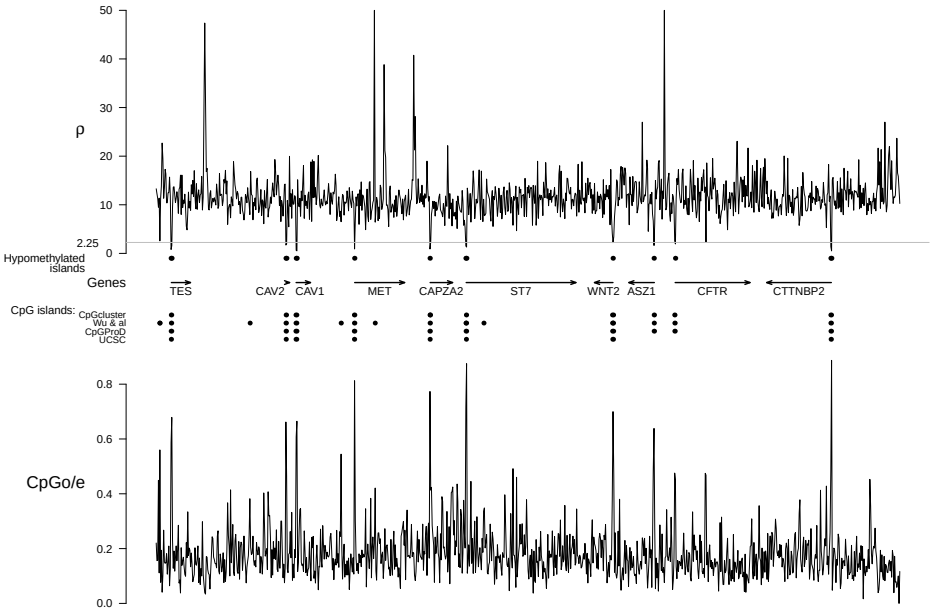


FIGURE 5. Top: CpG hyper-mutability rate $\rho = r_{CG\to CA}/r_{C\to T} = r_{CG\to TG}/r_{C\to T}$ along the human sequence, inferred over sliding windows of length 2000 by our method. Bottom: CpGo/e ratio on the corresponding windows.

contained in the alignment, and simulation studies (see **[30]**) suggest that this might lead to more accurate results. Fig. 5 shows the value of the hypermutability rate along the human sequence, inferred over sliding windows of length 2000, together with the corresponding value of the CpGo/e ratio.

4. Structural properties and statistical inference II

4.1. Likelihood computation for context-dependent nucleotide substitution models. In general, computing the exact value of the likelihood of a sample of DNA sequences for a context-dependent substitution model is an intractable problem. Inference methods either use a tractable substitute for the likelihood (among which the one we used in **[30]** in the special case of RN+YpR models), or EM/Bayes strategies in combination with Markov chain Monte Carlo methods to sample from the distribution of the (unobserved) substitution histories of sites conditional on the data. It turns out that the latter approach also has the potential to produce numerical estimates of the likelihood (or Bayes factor, in a bayesian framework), using a refined importance sampling scheme, as we now explain.

Given a family of sequences $s = (s_\alpha,\ \alpha \in \mathscr{S})$ and $h = (h(x),\ x \in J)$ describing the substitution history at sites $x \in J$, we use the notation $L(h, s|\theta)$ for the corresponding joint likelihood[9] with respect to the model θ.

[9]We do not enter into technical details here. More formally, $L(h, s|\ \theta)$ can be defined as the joint density of the random variables $\mathscr{H} = (\mathscr{H}(x),\ x \in J)$ and $S = (S_\alpha,\ \alpha \in \mathscr{S})$,

Now, for fixed s, consider the integral

$$I_{\theta_2,\theta_1} := \int \frac{L(h,s|\theta_2)}{L(h,s|\theta_1)} \cdot \frac{L(h,s|\theta_1)}{L(s|\ \theta_1)} dh. \tag{14}$$

On the one hand, I_{θ_2,θ_1} is the integral of

$$g_s^{\theta_2,\theta_1} \ : \ h \mapsto \frac{L(h,s|\theta_2)}{L(h,s|\theta_1)},$$

with respect to the conditional distribution of $\mathscr{H}$ given $S = s$. On the other hand,

$$I_{\theta_2,\theta_1} := \int \frac{L(h,s|\theta_2)}{L(s|\ \theta_1)} dh = \frac{L(s|\ \theta_2)}{L(s|\ \theta_1)}, \tag{15}$$

so that I_{θ_2,θ_1} is in fact the ratio between two values of the likelihood of the data s, computed for the two sets of model parameters θ_2 and θ_1.

Since $g_s^{\theta_2,\theta_1}$ admits an explicit expression, (14) allows one to numerically compute I_{θ_2,θ_1}, using substitution histories simulated from the conditional distribution of $\mathscr{H}$ given $S = s$. To compute the actual likelihood (and not just a ratio), one chooses for θ_2 the set of parameters of interest, and for θ_1 a set of parameters in which context-dependent substitution rates are zero, whence the possibility of evaluating $L(s|\theta_1)$ exactly by Felsenstein's algorithm.

This basic idea has to be refined, however, for the variance of $g_s^{\theta_2,\theta_1}$ with respect to the conditional distribution of $\mathscr{H}$ given $S = s$ may be infinite, in which case the Monte Carlo estimation of the ratio I_{θ_2,θ_1} is problematic. One refinement consists in choosing a sequence of parameters $\theta_1, \ldots, \theta_n$ in which θ_{i+1} is sufficiently "close" to θ_i so that the corresponding variance is finite. With $\theta_n = \theta$ being the parameter of interest, and θ_1 a parameter for which an exact computation of the likelihood is possible, one writes

$$L(s|\theta_n) = L(s|\theta_1) \times \frac{L(s|\ \theta_2)}{L(s|\ \theta_1)} \times \cdots \times \frac{L(s|\ \theta_n)}{L(s|\ \theta_{n-1})}, \tag{16}$$

each likelihood ratio being estimated by Monte Carlo simulation. Pushing this idea one step further, one may consider a smooth path in parameter space joining θ_1 to the target set of parameters θ, and replace the discrete telescopic product (16) by a continuous analogue, leading to the so-called *thermodynamic integration* approach (see **[83, 112]**).

4.2. Likelihood computations for RN+YpR models using sequential Monte Carlo methods (B. and Huet [31]). The approach developed in **[31]** is an alternative to thermodynamic integration for RN+YpR models, which consists in applying *sequential Monte Carlo methods*, see e.g. **[48, 4, 64, 57]**, and more precisely particle approximation algorithms developed for the filtering of hidden Markov models. We first describe the hidden Markov structure of RN+YpR models, then the corresponding particle approximation algorithm.

We have seen (see (11)) that, for general context-dependent substitution models with nearest-neighbour interaction, the family of full substitution histories $(\mathscr{H}(x), x \in J)$ has an explicit Markov random field structure (which

with respect to a suitable product reference measure (just the counting measure for S, and, for $\mathscr{H}$, the Lebesgue measure on $\mathbb{R}^k$ on the subspace where $\mathscr{H}$ has exactly k substitutions).

can be exploited to sample from the conditional distribution of these histories with respect to the sequences in the sample, using Markov chain Monte Carlo methods). The one-dimensional Markov random field structure implies that $(\mathscr{H}(x), x \in J)$ is in fact a Markov *chain* (of order 2). Thus, calling $\mathcal{O}(x)$ the family of nucleotides at site x in the DNA sequences that form the sample, i.e.

$$\mathcal{O}(x) := (S_\alpha(x), \alpha \in \mathscr{S}),$$

the pair $(\mathscr{H}(x), \mathcal{O}(x))_{x\in J}$ can be viewed as a hidden Markov chain. However, unlike the Markov random field structure, the Markov chain structure of $(\mathscr{H}(x), x \in J)$ cannot be described explicitly in general (to describe the transition from x to $x+1$, one has to integrate over all possible transitions from x to $x+1$ to $x+2$ etc.). However, in the case of RN+YpR models, an explicit Markov structure can be obtained for a slight modification of $(\mathscr{H}(x), \mathcal{O}(x))$ that uses the ρ and η encoding. Let

$$\mathscr{H}'(x) := (\rho(\mathscr{H}(x)), \eta(\mathscr{H}(x+1)),$$
$$\mathcal{O}'(x) := (\rho(\mathcal{O}(x)), \eta(\mathcal{O}(x+1))).$$

It can be proved that $(\mathscr{H}'(x), x \in J)$ has a Markov chain structure (of order 1), and that the corresponding transition kernel admits an explicit description in terms of a time-inhomogeneous Markov process running along the edges of the tree $\mathscr{T}$. Thus, a simulation procedure can be developed for this kernel, allowing the application of sequential Monte Carlo algorithms to the hidden Markov chain

$$(\mathscr{H}'(x), \mathcal{O}'(x))_{x\in J}.$$

We now describe the kind of sequential Monte Carlo algorithm we use to compute the likelihood. Define the event A_x by $A_x := \{\mathcal{O}'(x) = o'(x)\}$, where $o'(x) := ((\rho(s_\alpha(x)), \eta(s_\alpha(x+1))), \alpha \in \mathscr{S})$. Then define the conditioned transition probability kernel $\tilde{Q}$ by

$$\tilde{Q}(\xi, \cdot) := \mathbb{P}_\theta(\mathscr{H}'(x+1) \in \cdot | \mathscr{H}'(x) = \xi, A_{x+1}),$$

and

$$w_x(\xi) := \mathbb{P}_\theta(A_{x+1} | \mathscr{H}'(x) = \xi).$$

Among several possible variants, the algorithm we use, known in **[48]** as the sequential i.i.d. algorithm, consists in simulating for each $x \in J$, a population $(\xi_x^i,\ 1 \le i \le N)$, with the iteration leading from x to $x+1$ being of the following form:

- choose $I \in \{1, \dots, N\}$ according to the distribution
$$\frac{\sum_{i=1}^N w_x(\xi_x^i)\delta_{\xi_x^i}}{\sum_{i=1}^N w_x(\xi_x^i)},$$
- simulate ξ_{x+1}^i according to the probability distribution $\tilde{Q}(\xi^I, \cdot)$.

As a result, for large N, $(\xi_x^i,\ 1 \le i \le N)$ forms a particle approximation of the filtering distribution at x, i.e.

$$\frac{1}{N}\sum_{i=1}^N \delta_{\xi_x^i} \approx \mathbb{P}_\theta(\mathscr{H}'(x) \in \cdot | \mathcal{O}'(y) = o'(y),\ y \le x).$$

To compute an approximation of the likelihood, one writes,

$$L((s_\alpha, \alpha \in \mathscr{S})|\theta) = c_1 \times \frac{c_2}{c_1} \times \cdots \times \frac{c_m}{c_{m-1}}, \tag{17}$$

where, for $x \in J$,

$$c_x := \mathbb{P}_\theta(\mathcal{O}'(y) = o'(y),\ y \leq x),$$

and where we have assumed (without loss of generality) that $J = \{1, \ldots, m\}$. Using the particle approximation, one then has that, for large N,

$$\frac{c_{x+1}}{c_x} \approx \sum_{i=1}^{N} w_x(\xi_x^i),$$

and the overall likelihood can be estimated from (17).

Note that (17) shares some similarities with (16), and that, in both cases, we use Monte Carlo simulation to estimate individual factors in a product decomposition of the likelihood. However, here, we are using the sequential structure of the data, instead of relying on a sequence (or a path) of successive models. Moreover, the strong mixing properties of RN+YpR models make them especially nice candidates for sequential Monte Carlo algorithms.

Given the ability to obtain accurate numerical approximations of the likelihood, one can numerically investigate several interesting questions.

Validity of likelihood approximations. Several approaches to inference are based on neglecting some aspects of the dependence structure of the model to produce computationally tractable approximations of the likelihood. These can in turn be compared to the numerical estimates produced by the sequential Monte Carlo method. For instance, we have studied the validity of an approach where the sequence $\mathcal{O}'(x)$ is approximated by a Markov chain, showing that, on simulated data, the approximation can be both very accurate or very inaccurate, depending on the sample and on the region of the model's parameter space.

Loss of accuracy of ℓ_3-based inference. Since inference based on ℓ_3 does not use the true value of the likelihood, one expects inference based on ℓ_3 to be less accurate than inference based on the actual likelihood. However, in the examples we investigated, the difference between the estimates produced by both methods is negligible, compared to the fluctuations around the true value of the parameter that are due to the finite-size of the sample (the true value is known since we use simulated data).

Model comparisons. Using ℓ_3 does not allow us to perform likelihood-based comparisons with models outside the RN+YpR class. This is the case even for site-independent models that do not belong to the RN class, for which exact likelihood computations are nevertheless possible. Sequential Monte Carlo methods lead to numerical estimates that are accurate enough to detect likelihood differences between such models.

Accuracy of ancestral sequence reconstruction. Consider the problem of inferring the site-by-site conditional distribution of nucleotides in the ancestral sequence, conditional upon the data. Using $\Phi-$encoded polynucleotides, the distribution of the nucleotide at site x in the ancestral sequence can be inferred conditional upon truncated data, where only sites close to x (say at distance less than 2 or 3) are taken into account. On the other hand, the

sequential Monte Carlo algorithm we use automatically provides an estimation of this distribution, taking into account all the data. It turns out that, depending on the sample and of the region of the model's parameter space, the influence of non-nearest neighbours may or may not be negligible.

5. Perturbations

Our work on RN+YpR models exploits the specific structural properties of these models, which themselves depend crucially on the special assumptions on the rates that characterize the RN+YpR class. From a modeling perspective, one does not expect these special assumptions to hold exactly, but rather to provide, at best, a reasonably accurate approximation to the real substitution dynamics. Thus, one is naturally led to study perturbations of the RN+YpR class, where these assumptions are not met exactly. This question is also natural from a more mathematical perspective, where one would like to know how the specific dependence structure of RN+YpR models is transformed under slight perturbations of the assumptions.

Motivated by these questions, in **[32]**, we proved general results on perturbations of particle systems, first developed for the case of RN+YpR models in the preliminary work **[33]**. These results deal with the existence of *coupling from the past* (CFTP) times for the dynamics, named after the seminal paper by Propp and Wilson **[131]**.

The key idea of CFTP, formulated originally in the context of finite state-space Markov chains, is to simulate coupled trajectories of the process from further and further into the past, until eventually the present state of the Markov chain is the same for all these trajectories, regardless of their starting point. One then obtains an exact realization of the stationary distribution of the Markov chain, and, under a certain monotonicity condition on the transitions of the chain, CFTP leads to a practical algorithm for sampling from the stationary distribution. Many extensions of this scheme have been developed since, notably to include processes on more general state-spaces, and situations where the monotonicity condition is not met (see the online bibliography maintained by Wilson **[153]**).

In the context of interacting particle systems in the sense[10] of **[115]**, a natural coupling of the trajectories of the system is provided by the so-called graphical construction of the dynamics, based on Poisson processes (see below). Within this framework, the analog of CFTP is that, for any finite set of sites, the state of the system *restricted to these sites* does not depend on the initial configuration if one starts the dynamics far enough in the past. If such a property holds, we say that the particle system possesses a *CFTP time*.

The existence of a CFTP time is not only useful for simulation purposes, but leads to important theoretical properties of the particle system. Indeed, the existence of a CFTP time automatically implies that the interacting particle system is ergodic, and estimates on the tail of the CFTP time lead to bounds on the speed of convergence to the stationary distribution. Similarly,

[10]That is: continuous-time Markov processes which describe the evolution of a system of states attached to the sites of the lattice $\mathbb{Z}^d$ and such that the evolution at any site is governed by local transition rates involving the states of the neighboring sites.

estimates on the range of the space-dependence of the CFTP time yield bounds on the decay of spatial correlations for the stationary distribution.

Our main result is a general perturbation theorem, which can be stated informally as follows. Start with an interacting particle system possessing a CFTP time whose definition involves the exploration of an exponentially integrable number of points in the graphical construction, and which satisfies the positive rates property. Consider a perturbation obtained by adding new transition rates to the original dynamics. Then, provided that the perturbation is based on small enough rates, our result states that the perturbed interacting particle system possesses a CFTP time as well (with nice properties such as an exponentially decaying tail).

In the following, we give a more precise statement of this result, explain how it applies to RN+YpR models and other examples, and discuss its proof, which is based on the notion of *coupling time with ambiguities*.

5.1. A general perturbation theorem (B. and Piau [32]). The framework we consider is a natural extension of the one used to define context-dependent substitution models in Section 1. The four-letter nucleotidic alphabet $\mathcal{A}$ is replaced by a general finite alphabet $\mathscr{A}$, and the set of sites is $\mathbb{Z}^d$ for an arbitrary $d \geq 1$ instead of just $\mathbb{Z}$. Moreover, instead of the three transformations $\mathfrak{R}_{u\to v}$, $\mathfrak{R}_{uv\to uw}$ and $\mathfrak{R}_{uv\to wv}$, we consider general *transformation rules* of the form $\mathfrak{R} = (f, A, r)$, where A is a finite subset of $\mathbb{Z}^d$, $f : \mathscr{A}^A \to \mathscr{A}$ is a map, and $r \geq 0$ is a non-negative real number. Given a configuration of the system $\eta = (\eta(z))_{z\in\mathbb{Z}^d}$ in $\mathscr{A}^{\mathbb{Z}^d}$ and a site x in $\mathbb{Z}^d$, we denote by $\mathfrak{R}^x\eta$ the configuration such that

$$(\mathfrak{R}^x\eta)(x) = f((\eta(x+y))_{y\in A}),$$
$$(\mathfrak{R}^x\eta)(z) = \eta(z),\ z \neq x.$$

(Our convention when A is empty is that $\mathscr{A}^A$ is a singleton on which f takes a single well-defined value.)

The interacting particle system dynamics is defined by a finite list of such transition rules

$$\{\mathfrak{R}_i\,;\, i \in \mathcal{I}\}, \qquad \mathfrak{R}_i = (f_i, A_i, r_i),$$

through the infinitesimal generator G defined on functions that depend only on a finite number of coordinates by

$$G\phi(\eta) = \sum_{x\in\mathbb{Z}^d} \sum_{i\in\mathcal{I}} r_i \cdot (\phi(\mathfrak{R}_i^x\eta) - \phi(\eta))\,. \tag{18}$$

In the sequel, we assume that the dynamics is in fact built through the so-called graphical construction associated with the list of rules $(\mathfrak{R}_i)_{i\in\mathcal{I}}$ (see Liggett [**114**] for examples of this construction). Specifically, we consider a Poisson point process $\mathcal{P}$ on $\mathbb{Z}^d \times \mathcal{I} \times \mathbb{R}$, whose intensity is the product of counting (on $\mathbb{Z}^d$ and $\mathcal{I}$) and Lebesgue (on $\mathbb{R}$) measures, and whose realization prescribes the dynamics of the particle system as follows. First, for every x, $(\eta_t(x))_t$ is a jump process whose state may change only at times t for which there exists an (almost surely unique) index i such that (x, i, t) belongs to $\mathcal{P}$. Second, for every such time t,

$$\eta_t = \mathfrak{R}_i^x(\eta_{t-}). \tag{19}$$

Through the graphical construction, $\mathcal{P}$ induces a stochastic flow $\mathscr{F} = (\mathscr{F}_{t_1}^{t_2})_{t_1 \le t_2}$ on $\mathscr{A}^{\mathbb{Z}^d}$, defined by the fact that, for $t_1 \le t_2$ and ξ in $\mathscr{A}^{\mathbb{Z}^d}$, $\mathscr{F}_{t_1}^{t_2}(\xi)$ is the configuration of the system at time t_2 obtained by starting in configuration ξ at time t_1-, and using the transitions specified by $\mathcal{P}$ through (19).

The notion of CFTP time is defined in this context: we say that a negative and almost surely finite random variable T is a CFTP time (for site $x = 0$) if the following property holds:

$$\text{for all } \xi_1 \text{ and } \xi_2 \text{ in } \mathscr{A}^{\mathbb{Z}^d},\ [\mathscr{F}_T^{0-}(\xi_1)](0) = [\mathscr{F}_T^{0-}(\xi_2)](0) \text{ on } \{T > -\infty\}. \quad (20)$$

To formalize the notion of perturbation, we first consider a particle system whose dynamics is defined by a list of rules

$$\mathfrak{R}^u = \{\mathfrak{R}_i\,;\, i \in \mathcal{I}^u\}.$$

This corresponds to the original, unperturbed, particle system. The perturbed dynamics is defined by a list of rules of the form

$$\mathfrak{R}^u \cup \mathfrak{R}^p = \{\mathfrak{R}_i\,;\, i \in \mathcal{I}^u \cup \mathcal{I}^p\}, \qquad \mathcal{I}^u \cap \mathcal{I}^p = \varnothing.$$

Our result provides general conditions under which the existence of a CFTP time T^u for the unperturbed particle system leads to the existence of a CFTP time for the perturbed particle system, provided that the perturbation is small enough.

The first condition is that the unperturbed dynamics possesses the positive rates property. This means that, for every $v \in \mathscr{A}$, there exists a rule with index $\iota_v \in \mathcal{I}^u$ whose application unconditionally leads to the value v.

The second condition is that the definition of T^u involves the exploration of an exponentially integrable number of points in $\mathcal{P}$. To make this condition precise, we introduce the notion of an *exploration process* associated with T^u. This is a sequence $(\mathfrak{X}_n)_{n\ge 0}$ of subsets of $\mathcal{P}$, obtained by exploring the graphical construction further and further into the past. One starts with $\mathfrak{X}_0 := \emptyset$, and a current time value $t = 0$. From $\mathfrak{X}_n$, a set of active sites $\theta(\mathfrak{X}_n)$ is defined, and the graphical construction is searched for points that occur at active sites prior to the current time. The most recent such point is then added to $\mathfrak{X}_n$ to form $\mathfrak{X}_{n+1}$, and its time coordinate defines the new current time value. The process stops when the set of active sites is empty, in which case one sets $\mathfrak{X}_\infty := \mathfrak{X}_n$. We say that such an exploration process is associated with T^u if T^u is precisely the time coordinate of the last point added to the process, and if, on $\{T^u > -\infty\}$, the value of $[\mathscr{F}_{T^u}^{0-}(\xi)](0)$, which is the same for every ξ in $\mathscr{A}^{\mathbb{Z}^d}$, is measurable with respect to $\mathfrak{X}_\infty$.

Finally, the smallness of the perturbation is measured through two parameters ϵ and κ that admit explicit definitions in terms of $\mathfrak{R}^u$ and $\mathfrak{R}^u \cup \mathfrak{R}^p$,

$$\epsilon = \sup_{v\in\mathscr{A}} \left(\sum_{j\in\mathcal{I}^p;\ v\in f_j(A_j)} r_j \right) (r_{\iota_v})^{-1}, \qquad \kappa = \left(\sum_{i\in\mathcal{I}^p} |A_i| r_i \right) \left(\sum_{i\in\mathcal{I}} r_i \right)^{-1}. \quad (21)$$

We can now give a precise statement of the main theorem.

THEOREM 4 (B., Piau [32]). *Consider unperturbed dynamics $\mathfrak{R}^u$ with the positive rates property, and possessing a CFTP time T^u associated with*

an exploration process $(\mathfrak{X}_n)_n$ *such that* $|\mathfrak{X}_\infty|$ *has a finite exponential moment, i.e. for some* $\lambda > 0$*, one has* $\mathbb{E}(e^{\lambda|\mathfrak{X}_\infty|}) < +\infty$*. Then, for any perturbation with small enough parameters* ϵ *and* κ*, there exists a CFTP time* T^* *for the perturbed dynamics, which moreover possesses a finite exponential moment.*

Additional results on the range of the space-dependence of T^* with respect to the graphical construction are also obtained, see **[32]**. Although we do not discuss this question, non-asymptotic bounds (on the smallness of ϵ and κ, or on the exponential moment of T^*) can be derived from the proofs given in **[32]**. Also, let us point out again that the existence of a CFTP time with a finite exponential moment implies ergodicity of the particle system, with exponential bounds on the speed of convergence comparable to the one stated in Theorem 2.

Finally, note that, although **[32]** does not explicitly address issues related to the practical implementation of CFTP, the definition of T^* in terms of (T, H) makes it clear that T^* yields an actual CFTP algorithm for the perturbed particle system, provided that T^u and the associated exploration process are compatible with an actual algorithmic implementation.

5.1.1. *Applications.* We now discuss how Theorem 4 can be applied to RN+YpR models. In view of the positive rates assumption, we assume that all the basal substitution rates are positive, see (3), and re-express the dynamics of the model with the help of the following list of rules, each rule being of the form (f, A, r):

- Unconditional rules (denoted ι_v): for each v in $\mathcal{A}$, consider $A = \emptyset$, $r > 0$ and $f \equiv v$;
- Transversion rules: for each v in $\mathcal{A}$, consider $A := \{0\}$ and $f(w) = v$ if v and w are not both Y or both R, $f(w) = w$ otherwise;
- Transition rules: for each v in $\mathcal{A}$, consider $A = \{0\}$ and $f(w) = v$ if v and w are either both Y or both R, $f(w) = w$ otherwise;
- Left-dependent rules: for each u in Y, v in R, v' in R, consider $A = \{-1, 0\}$, $f(w_{-1}, w_0) = v'$ if $(w_{-1}, w_0) = (u, v)$, $f(w_{-1}, w_0) = w_0$ otherwise;
- Right-dependent rules: for each u in Y, v in R, u' in Y, consider $A = \{0, 1\}$, $f(w_0, w_1) = u'$ if $(w_0, w_1) = (u, v)$, $f(w_0, w_1) = w_0$ otherwise.

The key property is the following finite factor property of the dynamics, which holds for a similar reason as Theorem 1: the value of $\left[\mathscr{F}_t^{0-}(\xi)\right]$ is measurable with respect to the points in the graphical construction whose space coordinates lie in $\{-1, 0, 1\}$, and to $(\xi(x), x \in \{-1, 0, 1\})$. To explain how this property is used, introduce the notations

$$\mathcal{P}^x := \mathcal{P} \cap (\{x\} \times \mathcal{I} \times \mathbb{R}),\ \mathcal{P}^x_{s,s'} := \mathcal{P} \cap (\{x\} \times \mathcal{I} \times]s, s'[).$$

Say that a triple (t_{-1}, t_0, t_1) of negative times such that $t_0 > t_1$ and $t_0 > t_{-1}$ is a *locking triple* if

- for all $x \in \{-1, 0, 1\}$, one has that $(x, \iota_{v_x}, t_x) \in \mathcal{P}^x$ for some $v_x \in \mathcal{A}$;
- $\mathcal{P}^{-1}_{t_{-1},t_0} = \mathcal{P}^1_{t_1,t_0} = \varnothing$.

From the definition of the dynamics, one then has that, for all $x \in \{-1, 0, 1\}$, and all ξ,

$$[\mathscr{F}^{t_0}_{\min(t_{-1},t_1)}(\xi)](x) = v_x.$$

As a consequence of the finite factor property of the dynamics, this in turn implies that $[\mathscr{F}^{0-}_{\min(t_{-1},t_1)}(\xi)](0)$ takes the same value for all ξ. Thus, to define a CFTP time for the dynamics, one only has to run an exploration process on the graphical construction at sites $\{-1, 0, 1\}$ that looks for locking triples. From the independence properties of Poisson processes, it is easily seen that the time to wait before a locking triple appears possesses a finite exponential moment, and this is enough to ensure that $|\mathfrak{X}_\infty|$ has a finite exponential moment.

One of the limitations of the above definition is that, when context-dependent substitutions rates are large compared to the basal rates, locking triples are rare due to the second requirement in their definition. Accordingly, $\mathbb{E}(e^{\lambda|\Xi_\infty|})$ is finite only for small values of λ, and, as a result, the range of values of ϵ and κ for which Theorem 4 holds becomes small too. Using an alternative definition of locking triples, one gets a definition that is insensitive to large values of context-dependent rates, so that the magnitude of the perturbation allowed in Theorem 4 can be made independent of these rates. Alternative locking triples can e.g. be defined by

- for all $x \in \{-1, 0, 1\}$, one has that $(x, \iota_{v_x}, t_x) \in \mathcal{P}^x$ for some $v_x \in \mathcal{A}$ such that $v_{-1} \in R$ and $v_1 \in Y$;
- $\mathcal{P}^{-1}_{t_{-1},t_0}$ and $\mathcal{P}^{1}_{t_1,t_0}$ contain context-dependent rules only.

Note that Theorem 4 can be applied as soon as a finite factor property comparable to that of RN+YpR models hold, see **[32]**. However, RN+YpR models are by far our best motivated examples of interacting particle systems with such a property. Other examples of interacting particle systems to which Theorem 4 can be applied, and where the finite factor property does not hold, include one-dimensional *noisy voter models*, which are variants of the classical voter model (see **[114]**) in which unconditional rules are added. Two versions are considered in **[32]**, one in which the simplest version of the voter model is considered, and another version we call the voter model with asymmetric polling. We refer to **[32]** for a discussion of these examples.

5.2. Proofs.

5.2.1. *Coupling times with ambiguities.* The proof of Theorem 4 uses as an intermediate step the notion of *CFTP time with ambiguities*. This is a weakening of the notion of CFTP time, in which property (20) holds only when some so-called "ambiguities" associated with the rules attached to a specific random subset H of $\mathcal{P}$, are resolved. To give a precise definition, let us consider, for each point $\alpha = (x, i, t)$ in $\mathcal{P}$ and time $s < t$, the random variable $e(\alpha, \xi, s)$ defined as the value at site x produced by the application of the rule attached to α when starting in state ξ at time $s-$. More formally:

$$e(\alpha, s, \xi) = [\mathscr{F}^t_s(\xi)](x). \tag{22}$$

When there exist some states ξ_1 and ξ_2 such that $e(\alpha, s, \xi_1) \neq e(\alpha, s, \xi_2)$, we say that there is an ambiguity concerning the rule attached to α, starting at time $s-$.

Let us introduce the notation $\mathcal{P}_t = \mathcal{P} \cap (\mathbb{Z}^d \times \mathcal{I} \times [t, 0[)$. A CFTP time with ambiguities consists of a negative almost surely finite random variable T, together with a random subset H of $\mathcal{P}_T$, finite on the event $\{T > -\infty\}$, such that the following modification of (20) holds:

$$\begin{gathered}\text{for all } \xi_1 \text{ and } \xi_2 \text{ in } \mathscr{A}^{\mathbb{Z}^d},\ [\mathscr{F}_T^{0-}(\xi_1)](0) = [\mathscr{F}_T^{0-}(\xi_2)](0) \\ \text{provided that } e(\alpha, T, \xi_1) = e(\alpha, T, \xi_2) \text{ for all } \alpha \text{ in } H.\end{gathered} \tag{23}$$

Note that, when H is empty, (23) is identical to (20). In addition, we require that H has the stopping property in the sense that $H \cap \mathcal{P}_t$ is $\sigma(\mathcal{P}_t)$-measurable for all t.

Our first result is that, starting from a CFTP time with ambiguities (T, H), one can build an actual CFTP time T^*, provided that H contains few enough points on average. To give a precise statement, introduce the parameter

$$\mathfrak{g}(T, H) = \mathbb{E}\left(\sum_{(x,t,i) \in H} |A_i| \right).$$

THEOREM 5 (B., Piau [32]). *If there exists a CFTP time with ambiguities (T, H) such that $\mathfrak{g}(T, H) < 1$, then there exists a CFTP time T^*.*

Additionally, estimates on the tail of T^* and on the range of its space-dependence (in terms of bounds on exponential moments) can be derived from analogous properties for T and H, see **[32]**.

Here is an informal description of the construction of T^*. One recursively defines a sequence $(\mathrm{Amb}_n)_{n \geq 0}$ of random subsets of $\mathbb{Z}^d \times \mathbb{R}$ whose elements are called ambiguities, in the following way. Let $\mathrm{Amb}_0 = \{(0, 0)\}$ and fix $n \geqslant 0$. First apply the CFTP time with ambiguities (T, H) at each space-time point in Amb_n. This generates a set of elements of $\mathcal{P}$, with respect to which ambiguities have to be resolved. Then Amb_{n+1} is the set of space-time points upon which the resolution of these ambiguities directly depends or, in other words, for $\alpha = (x, i, t)$, the set $\{(x + y, t)\,;\ y \in A_i\}$. The overall set of points generated by this process is $\mathrm{Amb}_\infty = \bigcup_{n \geq 0} \mathrm{Amb}_n$ and T^* is the lowest value of T obtained when applying the CFTP time with ambiguities (T, H) to the space-time points in Amb_∞.

The idea which underlies the construction is that, when Amb_∞ is finite, we can resolve every ambiguity in a step-by-step manner, starting from the points in Amb_∞ that are furthest in the past and thus associated with an empty set of ambiguities, down to the origin where we can determine the value of $[\mathscr{F}_{T^*}^{0-}(\xi)](0)$. The almost sure finiteness of Amb_∞ is obtained by a first-moment argument which allows us to essentially bypass the analysis of dependencies which would otherwise be required to study the sequence of ambiguity sets $(\mathrm{Amb}_n)_{n \geq 0}$.

5.2.2. *Proof of Theorem 4.* The proof of Theorem 4 consists in constructing a CFTP time with ambiguities (T, H) for the perturbed dynamics, starting from the CFTP time T^u for the unperturbed dynamics. One then uses Theorem 5 to deduce the existence of a CFTP time for the perturbed dynamics.

The construction of (T, H) relies on what we call the exploration process with locking of perturbative ambiguities, attached to the perturbed dynamics. Informally, the construction can be described as follows. Run the exploration process associated with the *unperturbed* dynamics on $\mathcal{P}$. Each time a point $\alpha = (x, i, t)$ corresponding to a perturbative rule, in the sense that i belongs to $\mathcal{I}^p$, is encountered, split the exploration process into $|f(A_i)|$ exploration processes evolving in parallel, one for each v in $f(A_i)$, in which (x, i, t) is replaced by (x, ι_v, t).

Assuming that this process stops, one defines T as the lowest time-coordinate of a point encountered during the exploration, while H is defined as the set of such points that correspond to perturbative rules.

The fact that (T, H) is a CFTP time with ambiguities for the perturbed dynamics is a consequence of the fact that T^u is a CFTP time for the unperturbed dynamics. Informally, the idea is that fixing the ambiguities in H in a given way amounts to replacing the corresponding (x, i, t) with unconditional rules of the form (x, ι_v, t). The definition of the exploration process associated with T^u then guarantees that the value of $[\mathscr{F}_T^{0-}(\xi)](0)$ is the same for the perturbed dynamics whose ambiguities are fixed and for the unperturbed dynamics that uses the graphical construction in which the perturbative rules are replaced in the way described above.

The proof that the exploration process with locking of perturbative ambiguities almost surely stops is where the assumption that the perturbation is small enough and the exponential moment condition on $|\mathfrak{X}_\infty|$, are needed. These conditions allow a change-of-measure argument relating the perturbed and unperturbed graphical constructions to be made, which, in combination with a first-moment argument, leads to the desired result.

5.3. Discussion. For ergodic particle systems satisfying a monotonicity condition similar to the condition used by Propp and Wilson in [**131**], CFTP is always possible, as shown by van den Berg and Steif in [**145**]. For systems lacking this monotonicity condition, CFTP algorithms have been developed under so-called "high-noise" or "weak interaction" type assumptions, meaning that the strength of the interaction between neighboring sites has to be sufficiently small. In other words, the particle system under consideration must be a sufficiently small perturbation of a system in which distinct sites do not interact. An example is given by Haggström and Steif [**86**], where the authors use a bounding set approach to control the coalescence of trajectories (see also De Santis and Piccioni [**56**] for some refinements). Another example is given by Galves et al. [**77**] (see also Galves and al. [**78**] and Galves et al. [**79**]), who use a branching construction of which the approach we have developed can be seen as a generalization[11].

One interesting aspect of Theorem 4 is that it provides a general criterion under which small perturbations of an interacting particle system retain some of the CFTP properties of the original unperturbed system, allowing us to go beyond the weakly interacting case. Indeed, Theorem 4 can be

[11]A construction very similar to the one by Galves et al. [**77**] was already used in Ferrari et al. [**73**] to devise CFTP algorithms in a different framework. In fact, various other constructions of this kind appear in the literature, though not explicitly in the context of CFTP, see Ferrari [**72**] or the book edited by Dobrushin et al. [**61**] for examples.

applied to RN+YpR models that include arbitrarily strong interactions between neighbouring sites.

It would be interesting to find new examples of strongly interacting particle systems that satisfy the assumptions of Theorem 4, beyond those considered in **[32]**. Natural candidates are exponentially ergodic particle systems with a monotonicity property such as those considered in **[145]**, but naive upper bounds do not seem sufficient to prove that the finite exponential moment condition holds in general for these systems.

Finally, let us mention that some of our applications overlap with the recent paper by Mohylevskyy et al. **[121]**, where a special kind of perturbation of noisy voter models is considered and ergodicity is proved for sufficiently small perturbations.

6. Perspectives

To account for the variability of substitution rates with respect to both site and time, several extensions of the nucleotide substitution models described in Section 1 have been developed (see e.g. **[82]** Chap. 3). These include mixture models in which each site is assumed to have its own overall substitution rate, typically assuming an i.i.d. distribution of these rates across sites, chosen among a parametric family of distributions. These also include models in which substitution rates themselves evolve along the tree, e.g. according to a Markov process, leading to a Markov-modulated Markov process at the level of the sequence. A possible direction for extending the approach of **[30]** is to include such refinements.

One important point we have not discussed is the distinction between coding/non-coding parts of a DNA sequence. Due to the tri-nucleotide codon structure of the genetic code, the effect of a substitution on an organism's fitness can be dramatically different according to whether it occurs at the first, second or third position in a codon. To take this effect into account, substitution models which include a penalty for non-synonymous substitutions[12] are used. Unfortunately, adding such penalties to RN+YpR models seems to destroy some of their nice structural properties. We are currently working with L. Guéguen on the use of such models on coding sequences.

Another issue is that the models we have studied do not explicitly take into account the existence of insertions/deletions, but rather ignore the problem by using already aligned sequences. Adding an explicit stochastic model for insertions/deletions would clearly be an improvement over the present methodology.

Finally, one interesting perspective would be to develop computational methods that are able to deal with perturbations of RN+YpR models by taking advantage of their structural properties. This may enlarge the scope of the computational approach developed in **[30]** by allowing the inclusion of more realistic substitution models for which the restrictive assumptions of the RN+YpR class are not met exactly.

[12]A mutation on the DNA sequence is called synonymous if it does not alter the resulting amino-acid sequence.

CHAPTER 2

Interacting particle systems of the $X + Y \to 2X$ type

1. Introduction

1.1. The model(s). In this chapter, we consider interacting particle systems with two types of particles moving on the lattice $\mathbb{Z}^d$, denoted X and Y, with a local interaction rule modeling the irreversible autocatalytic reaction $X + Y \to 2X$. These particle systems have been introduced in the physical literature as microscopic stochastic models which, in the limit of a large average number of particles per lattice site, yield reaction-diffusion equations describing the propagation of a front, the prototypical example being the Fisher-Kolmogorov-Petrovsky-Piscounov equation. We refer to **[126]** for an extensive review of the subject from a theoretical physics perspective. In the models we consider, particles of both types move on $\mathbb{Z}^d$ by performing independent nearest-neighbour random walks in continuous time, with two possibly distinct jump rates $D_X \geq 0$ (for X particles) and $D_Y \geq 0$ (for Y particles). The reaction $X + Y \to 2Y$ is modeled by the following simple interaction rule: upon contact with an X particle, Y particles are instantaneously turned into X particles. Moreover, no particle (of either type) is injected into the system beyond those already present in the initial configuration, and no particle is ever removed from the system.

Typically, an initial configuration consists of a "gas" of Y particles spread over $\mathbb{Z}^d$, with constant or i.i.d. Poisson numbers of Y particles at each site, and a finite (nonzero) total number of X particles (this corresponds to the "small N" case in the terminology of **[126]**, meaning that the average number of particles per lattice site has a fixed value, rather than a large one with respect to which a limit is taken). One is then interested in characterising how the cluster formed by X particles gradually spreads into the space initially occupied by Y particles. Specifically, one would like to describe the large-time asymptotic shape of the set $\mathscr{B}_t$ formed by the sites that have been visited by an X particle prior to time t, i.e.

$$\mathscr{B}_t := \{x \in \mathbb{Z}^d;\ \exists s \in [0,t] \text{ such that } x \text{ bears an } X \text{ particle at time } s\}.$$

In the one-dimensional ($d = 1$) setting in which our own results have been obtained, we consider a slightly different kind of initial configuration, with constant or i.i.d. Poisson numbers of Y particles at sites located at the right of the origin, while all the particles at the left of the origin are assumed to be X particles. Figure 1 shows a simulation of the model in such a situation. One is then interested in characterising how the front separating X and Y particles moves towards $+\infty$. Specifically, one wants to describe the large-time asymptotic behaviour of the right-most position occupied by

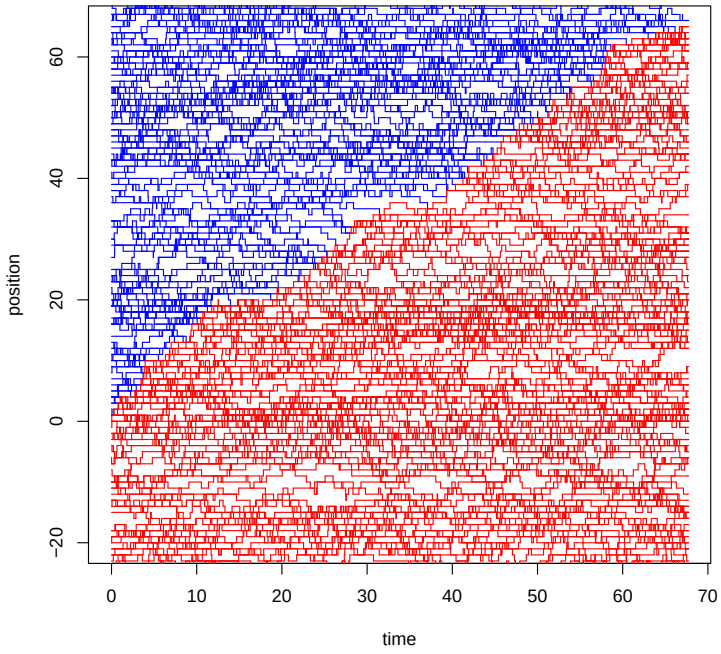


FIGURE 1. Time-evolution with $d = 1$, $D_X = D_Y = 1$, and i.i.d. Poisson numbers of mean 2. Red (resp. blue) trajectories correspond to particles of type X (resp. Y).

an X particle up to time t, i.e.

$$r_t := \sup\{x \in \mathbb{Z}, \quad \exists s \in [0,t] \text{ such that } x \text{ bears an } X \text{ particle at time } s\}.$$

To our knowledge, the first mathematical papers dealing with such particle systems, where both X and Y particles may have a non-zero jump rate, are **[100, 103]** by Kesten and Sidoravicius. The interpretation suggested in **[100, 103]** is that these particle systems can be viewed as microscopic stochastic models for the spread of a rumor or of an infection, where X particles represent informed/infected individuals, while Y particles represent ignorant/healthy individuals. Hence we use the generic term *KS infection model* to refer to them in the sequel. The special case where $D_X > 0$ and $D_Y = 0$ can also be interpreted as modeling the burning of a homogeneous solid, where X and Y particles correspond to heat packets and inert combustible molecules respectively, see **[134]** and the references to the physical literature therein. Therefore we use the term *stochastic combustion model* to refer to this case. Note that a fancier interpretation in terms of bouncing frogs can be found in **[1]**, leading to the somehow enigmatic name "frog model" that we chose not to use here.

One variation of the KS infection model we consider consists in making the infectious power of X particles remanent, in the sense that a Y particle turns into an X not only when it is in contact with a Y particle, but as

soon as it is located at a site that has previously been occupied by an X particle. We refer to this case as the *remanent KS infection model*. Other variations that we shall only briefly discuss include: the *activated random walk model*, where, in addition to the $X+Y \to 2X$ reaction, X particles turn to Y particles at a positive rate; the *modified DLA model*, where $D_X = 0$ and where the transition from Y to X happens when a Y particle attempts to jump to a vertex bearing an X particle, the jump then being cancelled.

1.2. Presentation of the results. We now proceed to the presentation of our results, with a short exposition of the relevant literature to put them into context. We refer to the survey paper by Kesten, Ramírez and Sidoravicius [**99**] for a more thorough discussion. The stochastic combustion case, i.e. the case $D_X > 0$ and $D_Y = 0$, is discussed first, then the general case where both $D_X > 0$ and $D_Y > 0$.

1.2.1. *Stochastic combustion model.* The first result we quote is an asymptotic shape theorem for $\mathscr{B}_t$, due to Ramírez and Sidoravicius. An analogous result for a discrete-time version of the model can be found in the paper [**1**] by Alves, Machado and Popov. We use the notation $[\![B]\!] := B \cap \mathbb{Z}^d$ for subsets $B \subset \mathbb{R}^d$.

THEOREM 6 (Ramírez, Sidoravicius [**134**]). *Consider the stochastic combustion model on $\mathbb{Z}^d$ starting with an initial configuration consisting of one X particle at the origin, and one Y particle at every other site. There exists a closed convex bounded subset $B^d \subset \mathbb{R}^d$, symmetric with respect to permutations of the coordinate axes, and with non-empty interior, such that, for all $\epsilon > 0$, almost surely for large enough t, one has*

$$[\![(1-\epsilon)tB^d]\!] \subset \mathscr{B}_t \subset [\![(1+\epsilon)tB^d]\!].$$

In view of Theorem 6 above, the next natural question to ask is that of the fluctuations of $t^{-1}\mathscr{B}_t$ around the limiting shape B^d. In dimension $d \geq 2$, there is no known answer, even at the level of a rough order of magnitude. On the other hand, in dimension $d = 1$, a central limit theorem with $t^{-1/2}$ scaling for the position of the front at time t has been obtained by Comets, Quastel and Ramírez [**53**], and we now describe this result in detail. The system is assumed to start in an initial configuration comprising a fixed number $\mathfrak{a} \geq 1$ of Y particles at each site $x \geq 1$, while all the particles initially at sites $x \leq 0$ are X particles (we assume that there is at least one such particle). For $x \leq 0$, we denote by $\eta(x)$ the number of particles at site x in the initial configuration, and make the assumption that, for a small enough θ (depending on $\mathfrak{a}$), the following condition is satisfied

$$\sum_{x \leq 0} \eta(x) e^{\theta x} < +\infty. \tag{24}$$

We first state the analog of Theorem 6 in the slightly different context we are now studying.

THEOREM 7 (Comets, Quastel, Ramírez [**53**]). *For the stochastic combustion model on $\mathbb{Z}$, there exists $0 < v < +\infty$ such that, for any initial configuration with $\mathfrak{a}$ particles of type Y at each site $x \geq 1$, satisfying the*

growth condition (24), *one almost surely has that*

$$\lim_{t\to+\infty} t^{-1} r_t = v.$$

The fluctuations are then characterised by the following result.

THEOREM 8 (Comets, Quastel, Ramírez [**53**]). *For the stochastic combustion model on* $\mathbb{Z}$, *there exists* $0 < \sigma^2 < +\infty$ *such that, for any initial configuration with* $\mathfrak{a}$ *particles of type* Y *at each site* $x \geq 1$, *satisfying the growth condition* (24), *one has that, as* ϵ *goes to zero,*

$$\epsilon^{1/2}\left(r_{\epsilon^{-1}t} - \epsilon^{-1} v t\right), \qquad t \geq 0,$$

converges in distribution on the Skorohod space to a Brownian motion with variance σ^2.

We now describe the results we have obtained on the probabilities of large deviations for the front. We assume that the initial configuration satisfies the following condition:

$$\text{for all } \theta > 0, \ \sum_{x\leq 0} \eta(x) e^{\theta x} < +\infty, \tag{25}$$

which is a strengthening of condition (24) above. With this assumption, we have the following large deviations principle.

THEOREM 9 (B., Ramírez, [**36**]). *For the stochastic combustion model on* $\mathbb{Z}$, *there exists a rate function* $I : [0, +\infty[\to [0, +\infty[$ *such that, for any initial configuration with* $\mathfrak{a}$ *particles of type* Y *at each site* $x \geq 1$, *satisfying the growth condition* (25), *one has that*

$$\limsup_{t\to+\infty} \frac{1}{t} \log \mathbb{P}\left[\frac{r_t}{t} \in C\right] \leq -\inf_{b\in C} I(b), \qquad \text{for} \quad C \subset [0, +\infty[\quad \text{closed},$$

and

$$\liminf_{t\to+\infty} \frac{1}{t} \log \mathbb{P}\left[\frac{r_t}{t} \in G\right] \geq -\inf_{b\in G} I(b), \qquad \text{for} \quad G \subset [0, +\infty[\quad \text{open}.$$

Furthermore, I *is identically zero on* $[0, v]$, *while* I *is positive, convex and increasing on* $]v, +\infty[$.

More precise estimates for the probabilities of slowdown large deviations are available. To state them, let

$$U := \limsup_{x\to-\infty} \frac{1}{\log|x|} \log\left(\sum_{y=0}^{x} \eta(y)\right), \ u := \liminf_{x\to-\infty} \frac{1}{\log|x|} \log\left(\sum_{y=0}^{x} \eta(y)\right),$$

and

$$s := \min(1, U).$$

THEOREM 10 (B., Ramírez, [**36**]). *Consider the stochastic combustion model on* $\mathbb{Z}$, *and an initial configuration with* $\mathfrak{a}$ *particles of type* Y *at each site* $x \geq 1$, *satisfying the growth condition* (25). *Then the following results hold true.*

(a) *For all* $0 \leq c < b < v$*, as* t *goes to infinity,*

$$\mathbb{P}\left[c \leq \frac{r_t}{t} \leq b\right] \geq \exp\left(-t^{s/2+o(1)}\right). \tag{26}$$

(b) *In the special case where* $\eta(x) \geq \mathfrak{a}$ *for all* $x \leq 0$*, one has that, for every* $0 \leq b < v$*, as* t *goes to infinity,*

$$\mathbb{P}\left[\frac{r_t}{t} \leq b\right] \leq \exp\left(-t^{1/3+o(1)}\right). \tag{27}$$

(c) *When* $u < +\infty$*, as* t *goes to infinity,*

$$\exp\left(-t^{U/2+o(1)}\right) \leq \mathbb{P}\left[r_t = 0\right] \leq \exp\left(-t^{u/2+o(1)}\right). \tag{28}$$

1.2.2. *KS infection model with* $D_X > 0$ *and* $D_Y > 0$. We now turn to the more general case where both D_X and D_Y are non-zero. We first quote an asymptotic shape theorem similar to Theorem 6, which is due to Kesten and Sidoravicius and holds for the case where $D_X = D_Y$.

THEOREM 11 (Kesten, Sidoravicius **[103]**). *Consider the KS infection model on* $\mathbb{Z}^d$ *with* $D_X = D_Y$*, starting with an initial configuration consisting of i.i.d. Poisson numbers of* Y *particle at each site, and a finite (positive) number of* X *particles. There exists a closed convex bounded subset* $B^d \subset \mathbb{R}^d$*, symmetric with respect to permutations of the coordinate axes, and with non-empty interior, such that, for all* $\epsilon > 0$*, one has almost surely that, for large enough* t*,*

$$[\![(1-\epsilon)tB^d]\!] \subset \mathscr{B}_t \subset [\![(1+\epsilon)tB^d]\!].$$

As is already the case for the stochastic combustion model, in dimension $d \geq 2$, no result concerning the fluctuations around the limiting shape B^d is available. However, in dimension $d = 1$, we have obtained a central limit theorem with $t^{-1/2}$ scaling, that we now describe. The system is assumed to start in an initial configuration comprising i.i.d. Poisson numbers of particles at each site, with every particle at the right of the origin being of type Y, while every particle at the left of the origin is of type X. In this context, Theorem 11 above shows that there exists $0 < v < +\infty$ such that a.s.,

$$\lim_{t \to +\infty} t^{-1} r_t = v. \tag{29}$$

Fluctuations around the limiting speed v are then described by the following theorem.

THEOREM 12 (B., Ramírez **[35]**). *For the KS infection model on* $\mathbb{Z}$ *with* $D_X = D_Y$*, starting with i.i.d. Poisson numbers of* X *(resp.* Y*) particles at the left (resp. right) of the origin, there exists* $0 < \sigma^2 < +\infty$ *such that, as* ϵ *goes to zero,*

$$B_t^\epsilon := \epsilon^{1/2}\left(r_{\epsilon^{-1}t} - \epsilon^{-1} v t\right), \qquad t \geq 0,$$

converges in law on the Skorohod space to a Brownian motion with variance σ^2.

The KS infection model is not well-understood when $D_X \neq D_Y$, even in dimension $d = 1$. One still has the following upper bound showing that $\mathscr{B}_t$

spreads at most linearly in time, but it is not known whether a matching (i.e. growing linearly with time) lower bound holds[1].

THEOREM 13 (Kesten, Sidoravicius [**100**]). *Consider the KS infection model on $\mathbb{Z}$, starting with an initial configuration consisting of i.i.d. Poisson numbers of Y particle at each site, and a finite (positive) number of X particles. There exists a constant $C > 0$ such that almost surely, for all large enough t, one has that*

$$\mathscr{B}_t \subset [\![t[-C, C]^d]\!]$$

For the remanent KS model however, the following result shows that both a law of large numbers and a central limit theorem hold, under the assumption that $D_X \geq D_Y$.

THEOREM 14 (B., Ramírez [**35**]). *Consider the remanent KS infection model on $\mathbb{Z}$, with $D_X \geq D_Y$, starting with i.i.d. Poisson numbers of X (resp. Y) particles at the left (resp. right) of the origin. There exists $0 < v < +\infty$ such that one almost surely has that*

$$\lim_{t \to +\infty} t^{-1} r_t = v.$$

Also, there exists $0 < \sigma^2 < +\infty$ such that, as ϵ goes to zero,

$$B_t^\epsilon := \epsilon^{1/2} \left(r_{\epsilon^{-1} t} - \epsilon^{-1} v t \right), \qquad t \geq 0,$$

converges in law on the Skorohod space to a Brownian motion with variance σ^2.

1.3. Organization of the chapter. The rest of this chapter is organized as follows. Sections 2 and 3 describe the various approaches leading to the proofs of the results quoted above, in the stochastic combustion case (Section 2), and in the general case where $D_X > 0$ and $D_Y > 0$ (Section 3). Finally, Section 4 discusses open questions and perspectives, and contains additional comments and references to the literature.

2. Proofs: stochastic combustion model ($D_X > 0$, $D_Y = 0$)

Since, for the stochastic combustion model, changing the value of D_X amounts to rescaling time by a constant factor, we assume in the following discussion that $D_X = 2$, which corresponds to the choice made in [**53, 36**].

2.1. Sub-additivity. The approach used by Ramírez and Sidoravicius in [**134**] to prove Theorem 6 (and also by Alves, Machado and Popov in [**1**]) is based on a sub-additivity property of hitting times, which can be stated as follows. Assign to each site $x \in \mathbb{Z}^d$ a random walk path $W(x)$ in an i.i.d. manner. Then, for every $u, v \in \mathbb{Z}^d$, define $T(u, v)$ as the first hitting time of v by an X particle, i.e.

$$T(u, v) := \inf\{t \geq 0;\ v \text{ bears an } X \text{ particle}\}, \tag{30}$$

[1] One can still prove a lower bound of the form $[\![t(\log t)^{-p}[-c, c]^d]\!] \subset \mathscr{B}_t$ for some constants $p, c > 0$, see [**100**].

where the initial configuration consists of one X particle at site u, and one Y particle at every other site, and where, for every x, the particle initially located at x moves according to the random walk $W(x)$ as soon as it is turned into an X particle (which means at time zero for the particle initially at u). Then, for all $u, v, w \in \mathbb{Z}^d$, one has a.s. that

$$T(u, w) \leq T(u, v) + T(v, w). \tag{31}$$

Having established (31), the proof then mainly consists in checking the other assumptions of Kingman's sub-additive ergodic theorem (see e.g. **[114]**), which is by no means an easy task. In fact, most of the work in **[134]** and **[1]** is devoted to proving that the random variables $T(u, v)$ have a finite expectation.

2.2. Regeneration structure. On the other hand, the approach used by Comets, Quastel and Ramírez in **[53]** to prove Theorems 7 and 8 is based on a completely different idea, which consists in introducing a renewal structure for the stochastic combustion model, yielding a sequence of a.s. finite random times $0 =: \kappa_0 < \kappa_1 < \kappa_2 < \dots$ such that

(i) the r.v.s $(\kappa_{n+1} - \kappa_n, r_{\kappa_{n+1}} - r_{\kappa_n})_{n \geq 0}$ are independent,
(ii) the r.v.s $(\kappa_{n+1} - \kappa_n, r_{\kappa_{n+1}} - r_{\kappa_n})_{n \geq 1}$ are identically distributed,
(iii) $\mathbb{E}(\kappa_2 - \kappa_1)^2 < +\infty$ and $\mathbb{E}(r_{\kappa_2} - r_{\kappa_1})^2 < +\infty$.

Given such a renewal structure, Theorems 7 and 8 can be derived in a standard way, applying to r_{κ_n} the law of large numbers and the central limit theorem for sums of i.i.d. square-integrable random variables, then approximating r_t by $r_{\kappa_{s(t)}}$, where $s(t) := \sup\{n \geq 1;\ \kappa_n \leq t\}$. In fact, the core of the work lies in finding an appropriate definition for the renewal structure, and then proving the required tail-estimates.

Broadly speaking, the idea is to define κ_n in such a way that the history of the front after time κ_n does not depend (up to translation) on the trajectories of particles located below r_{κ_n} at time κ_n. This is achieved by considering times after which the front remains forever above a (space-time) straight line, while particles lying below the front at these times remain forever below the straight-line. Specifically, given a slope parameter α, let us say that a time t at which the front jumps is a *forward super-α time* if, for every $s \geq t$, one has $r_s \geq r_t + \lfloor \alpha(s - t) \rfloor$, and a *forward sub-$\alpha$ time* if, for any particle whose location at time t is $\leq r_t - 1$, the corresponding random walk trajectory, denoted $(W_s)_{s \geq 0}$, satisfies $W_s \leq r_s - 1 + \alpha(s - t)$ for all $s \geq t$. If t is both a forward super-α time and a forward sub-α time, let us say that t is a *forward α time*. These definitions are illustrated in Fig. 2, 3, 4. What makes the existence of forward α times plausible for small enough α is the fact that the front r_t moves ballistically, while any individual particle trajectory moves diffusively. It turns out that the sequence defined by[2] $\kappa_0 := 0$ and, for all

[2]This is not the way the random variables $(\kappa_n)_n$ are defined in **[53]**. Indeed, the definition given in **[53]** includes several additional conditions that are used later in the proof of the tail-estimates for the renewal structure, but are not necessary to establish the basic structural properties (i)-(ii) of independence and identity of distributions. Moreover, the tail-estimates proved in **[53]** imply the square integrability property (iii) for the random variables $(\kappa_n)_n$ as defined by (32). We chose to present this definition instead of the

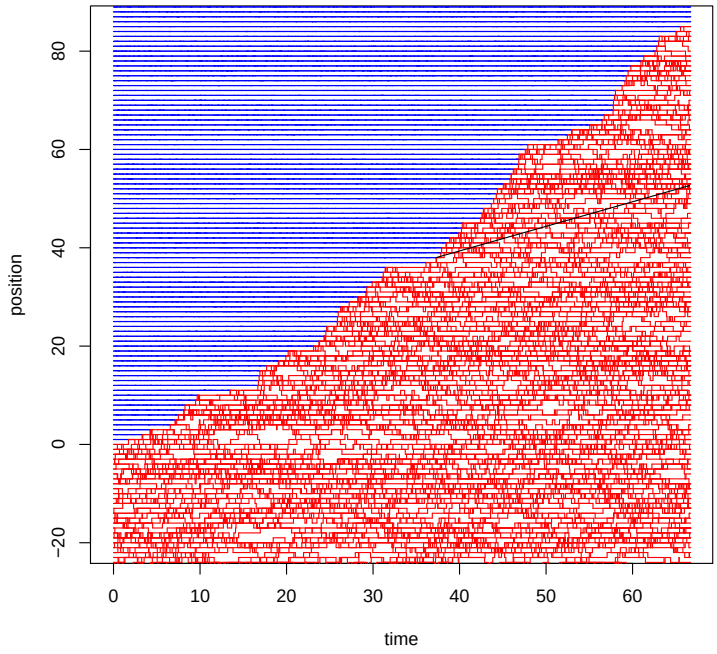


FIGURE 2. A realization of the stochastic combustion model with a forward α time t. Red (resp. blue) trajectories correspond to particles of type X (resp. Y). The line of slope α associated with the forward α time is drawn in black.

$n \geq 0$,

$$\kappa_{n+1} := \inf\{t > \kappa_n;\ t \text{ is a forward } \alpha \text{ time}\}, \tag{32}$$

provides a.s. finite random variables enjoying the required properties (i)-(ii)-(iii), for a suitable choice of the slope parameter α.

To explain why (i) and (ii) hold (leaving aside the fact that one must also prove that the κ_ns are a.s. finite), let us observe that the two conditions involved in the definition of a forward α time ensure that, for $n \geq 1$, the particles located strictly below r_{κ_n} at time κ_n cannot possibly touch the front after time κ_n, so we may as well remove them at time κ_n without altering the future evolution of the front.

Moreover, the value of $(\kappa_{n+1} - \kappa_n, r_{\kappa_{n+1}} - r_{\kappa_n})$ is identical to the value of (κ_1, r_{κ_1}) that would be obtained if we shifted the origin of space and time to (r_{κ_n}, κ_n) after removing these particles. Now the configuration of the remaining particles, i.e. those above r_{κ_n} at time κ_n, is completely determined, with exactly $\mathfrak{a}$ particles of type Y particles at each site $x \geq r_{\kappa_n} + 1$, and $\mathfrak{a} + 1$ particles of type X at site r_{κ_n} (the $\mathfrak{a}$ particles of type Y originally at site r_{κ_n}, plus the single particle of type X that reaches r_{κ_n} at time κ_n and instantaneously turns them into particles of type X). Moreover, the

original one because it is simpler to state, and also because it makes the comparison with the renewal structure used for the KS infection model, defined in Section 3 below, easier.

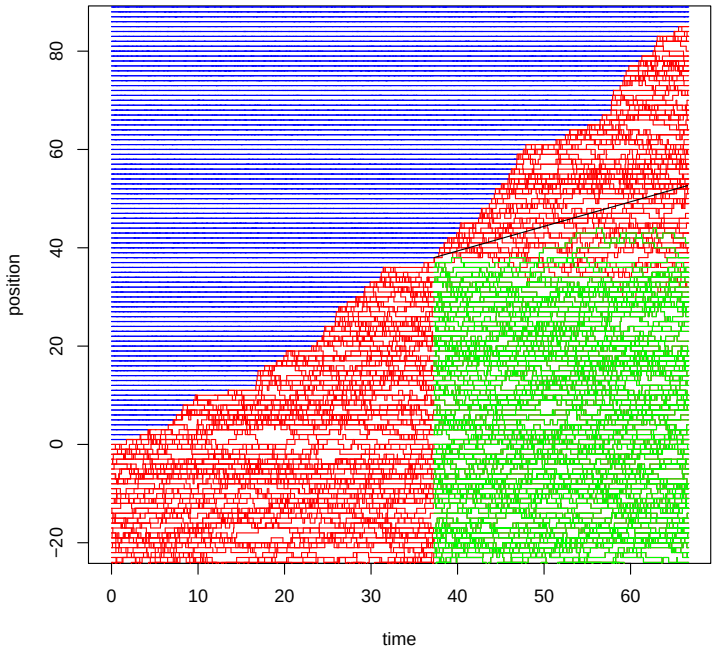


FIGURE 3. Same realization of the stochastic combustion model as in Fig. 2. Red (resp. blue) trajectories correspond to particles of type X (resp. Y), except that, posterior to the forward α time t, green is used instead of red to draw the trajectories of particles that lie below r_t at time t.

conditioning induced on the future of the trajectories of these particles by the definition of κ_n reduces[3] to the fact that, posterior to time κ_n, one must have $r_t \geq r_{\kappa_n} + \lfloor \alpha(t - \kappa_n) \rfloor$ for all t. As a consequence, one finally has that, for $n \geq 1$, almost surely

$$\mathbb{P}((\kappa_{n+1} - \kappa_n, r_{\kappa_{n+1}} - r_{\kappa_n}) \in \cdot | \mathcal{F}_{\kappa_n}) = \mathbb{P}_0((\kappa_1, r_{\kappa_1}) \in \cdot | A), \tag{33}$$

where $\mathcal{F}_{\kappa_n}$ denotes the $\sigma-$algebra generated by κ_n, r_{κ_n}, and the history of the process up to time κ_n, $\mathbb{P}_0$ denotes the probability describing the stochastic combustion model when one starts with $\mathfrak{a}$ particles of type Y at each $x \geq 1$, and $\mathfrak{a} + 1$ particles of type X at $x = 0$, and A is the event corresponding to the fact that $t = 0$ is a forward super-α time.

The next step is to prove tail bounds leading to the almost sure finiteness of the (κ_n)s, and the square-integrability property (iii). Although we do not give a detailed account of how these bounds are obtained, we briefly describe the two key ingredients that are used in the proofs.

The first one is the so-called *auxiliary front* $\tilde{r}_t$, which provides a ballistic lower bound for the position of the real front. The auxiliary front starts with

[3]This is not obvious. For instance, one has to check that the conditions contained in the definitions of $\kappa_1, \ldots, \kappa_{n-1}$, which bear upon the whole future of the process, reduce to this single condition as far as the behaviour of trajectories posterior to κ_n is concerned.

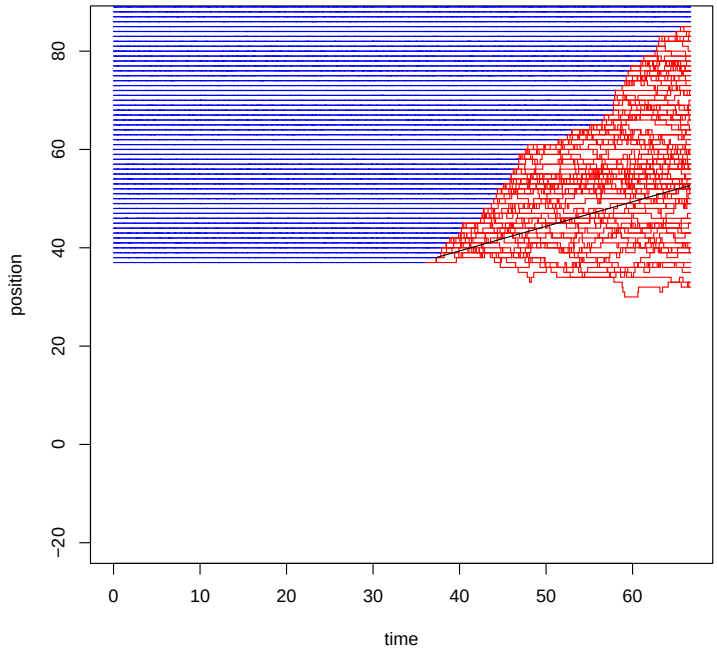


FIGURE 4. Compared to the realization of the stochastic combustion model depicted in Fig. 2 and 3, particles that lie below r_t at the forward α time t have been removed. Red (resp. blue) trajectories correspond to particles of type X (resp. Y), and one can check that the evolution of the front posterior to the α separation time t is unaffected by this removal.

$\tilde{r}_0 := 0$, and its subsequent evolution is defined through the random variables ν_k that count the time it takes for $\tilde{r}_t$ to climb from level k to level $k+1$. The definition of ν_k involves only the trajectories of the particles initially located at sites $k - M + 1 \leq x \leq k$, where $M \geq 1$ is an integer parameter. Specifically, consider the time it takes for a Y particle initially located at a site x, to first hit site $k+1$, counted from the instant it was turned into an X particle. One defines ν_k as the infimum of these hitting times over all the Y particles initially located at sites x satisfying $k - M + 1 \leq x \leq k$. Thanks to the fact that the auxiliary front involves only delayed trajectories of a subset of the X particles present in the actual stochastic combustion process, one has for all t that

$$\tilde{r}_t \leq r_t. \tag{34}$$

On the other hand, $\tilde{r}_t$ behaves mostly like a sum of i.i.d. random variables, since, for each $1 \leq j \leq M-1$, the random variables $(\nu_{Mk+j})_{k \geq 1}$ form an i.i.d.

sequence, whose tail[4] decays at least as fast as $t^{-\mathfrak{a}M/2}$, leading to a finite expectation when $M \geq 3$. As a consequence, provided that M is chosen ≥ 3, the law of large numbers implies that

$$\lim_{t\to\infty} \tilde{r}_t/t =: \alpha_{\text{aux}} > 0. \tag{35}$$

The second ingredient is the use of an exponential norm $\mathcal{M}_t$, in combination with martingale arguments. More precisely, one defines

$$\mathcal{M}_t := \sum_{x\leq r_t-1} \eta_t(x) e^{-\theta(x-r_t)}, \tag{36}$$

where $\eta_t(x)$ denotes the number of X particles at site x and time t. To each random walk $(W_t)_{t\geq 0}$ describing a particle trajectory, one can associate the exponential martingale

$$M_t := e^{\theta W_t - 2(\cosh(\theta)-1)t},$$

and these martingales can in turn be used to control the probability for particles that are currently located below the front, to hit in the future a straight line of slope α starting at the current location of the front. The norm $\mathcal{M}_t$ then appears as a key quantity in these martingale bounds. For instance, one has that the probability that any particle whose location at time 0 is $\leq r_0 - 1$, will lie above $r_0 - 1 + \alpha s$ at some later time $s \geq t$, is bounded above by $\mathcal{M}_0 e^{-\mu t}$, where

$$\mu := \alpha\theta - 2(\cosh\theta - 1). \tag{37}$$

The argument developed in **[53]** combines these two ingredients to produce tail estimates on the renewal structure. One chooses $\alpha < \alpha_{\text{aux}}$, so that the auxiliary front can be used to show that the front typically moves at a speed faster than α. One then chooses θ small enough so that $\mu > 0$, to have exponential decay in martingale bounds involving $\mathcal{M}_t$, and control the probability for particles below the front to hit straight lines of slope α in the future. At the core of the argument is a sequence of stopping times, which correspond to as many attempts at producing a forward α time. The above two ingredients are used to control the probability of success of each attempt, as well as the time between two consecutive attempts. We do not go into the details here, but a precise discussion of the corresponding problem in the context of the renewal structure defined for the KS infection model is given in Section 3.

2.3. Large deviations. We now describe how our results on large deviations are derived.

First, the existence of the limiting rate function in Theorem 9 essentially follows from a soft argument based on the sub-additivity of hitting times, yielding the fact that, for initial configurations consisting of a single X particle at the origin and $\mathfrak{a}$ particles of type Y at every site $x \geq 1$, for all $b \geq 0$, the limit

$$\lim_{t\to+\infty} t^{-1} \log \mathbb{P}(r_t \geq bt) \tag{38}$$

[4]The hitting time of a site by a single symmetric random walk has a tail decaying roughly as $t^{-1/2}$. Taking into account $\mathfrak{a}M$ independent such random walks yields a tail decaying as $t^{-\mathfrak{a}M/2}$.

exists.

The more difficult part of Theorem 9 is the characterisation of the zero set of the rate function, especially the proof that the rate function does not vanish on the interval $]v, +\infty[$. In fact, it is not difficult to show that, for all large enough b, $\mathbb{P}(r_t \geq bt)$ decays exponentially fast as t goes to infinity. However, showing this for b arbitrarily close to, but larger than, the speed v is a subtler problem. Indeed, one cannot apply standard large deviations theory to the regeneration structure, since e.g. the random variables κ_1 and r_{κ_1} fail to have finite exponential moments.

Instead, working with a homogeneous initial configuration containing exactly $\mathfrak{a}$ particles per site, we apply the renewal structure to a perturbation of the original model, in which the random walks have a small bias $\epsilon > 0$ to the right. Denoting by r_t^ϵ the position of the front for this perturbed model, one has again a law of large numbers:

$$\lim_{t\to\infty} t^{-1} r_t^\epsilon = v_\epsilon, \, \mathbb{P} - a.s. \text{ and in } L^1(\mathbb{P}). \tag{39}$$

The interest of introducing a bias to the right is that, reworking the estimates of **[53]** in this context, we can show that the random variables in the renewal structure do indeed have finite exponential moments (the key point being that the time it takes the auxiliary front to climb from level k to level $k+1$ now has an exponential tail, as opposed to polynomial for the original model). As a consequence, one can apply standard large deviations arguments to the regeneration structure, and prove that, for any $b > v_\epsilon$,

$$\limsup_{t\to+\infty} t^{-1} \log \mathbb{P}(r_t^\epsilon \geq bt) < 0. \tag{40}$$

On the other hand, biasing the random walks to the right cannot decrease the position of the front, so that at each time t, a comparison holds between the position of the front in the original model and in the model with a bias, so that for all t and x,

$$\mathbb{P}(r_t \geq x) \leq \mathbb{P}(r_t^\epsilon \geq x). \tag{41}$$

Combining (40) and (41) is enough to prove that the rate function I must be positive on every interval of the form $]v_\epsilon, +\infty[$, for $\epsilon > 0$. To prove that I is positive on the whole interval $]v, +\infty[$, noting that v_ϵ is a non-decreasing function of ϵ, we should prove in addition that

$$\lim_{\epsilon\to 0+} v_\epsilon = v. \tag{42}$$

It is indeed reasonable to expect such a continuity property to hold, but proving it seems to require substantial work. Indeed, write

$$v_\epsilon = \lim_{t\to+\infty} t^{-1}\mathbb{E}(r_t^\epsilon). \tag{43}$$

$$v = \lim_{t\to+\infty} t^{-1}\mathbb{E}(r_t).$$

For fixed t, it is possible (using the dominated convergence theorem) to prove that

$$\lim_{\epsilon\to 0+} \mathbb{E}(r_t^\epsilon) = \mathbb{E}(r_t). \tag{44}$$

Hence, to prove (42), it is enough to prove that

$$\lim_{\epsilon\to 0+}\lim_{t\to+\infty} t^{-1}\mathbb{E}(r_t^\epsilon) = \lim_{t\to+\infty}\lim_{\epsilon\to 0+} t^{-1}\mathbb{E}(r_t^\epsilon). \tag{45}$$

Our strategy for proving (45) is to prove that the convergence in (43) is uniform with respect to (small enough) ϵ, which implies that the limits with respect to $\epsilon \to 0+$ and to $t \to +\infty$ in (45)) can be exchanged. To prove this uniformity, we use the renewal structure once again, showing that the estimates on the second moments of the random variables in the renewal structure obtained in **[53]** lead to uniform upper bounds with respect to ϵ, which is enough to prove the required uniformity in (43).

We now give a brief sketch of the proof of Theorem 10 (which in particular implies that the rate function I vanishes on $[0, v]$). We start with Theorem 10 (c). The fact that $r_t = 0$ means that no particle in the initial configuration hits 1 before time t. Both the upper and lower bounds can then be understood heuristically as follows. Since we consider simple symmetric random walks, for large t, the constraint of not hitting 1 before time t has a cost only for particles within a distance of order $t^{1/2}$ of the origin. Now these particles perform independent random walks, and their number has an order of magnitude lying between $t^{u/2}$ and $t^{U/2}$. Turning this argument into a proper proof involves only elementary diffusive estimates and the reflection principle.

As for Theorem 10 (a), the idea of the proof when $s(\eta) = 1$ is to combine the following two arguments. First, for $b > 0$, it costs nothing to prevent all the particles in the initial condition from hitting $\lfloor bt \rfloor$ up to time t. Intuitively, this result comes from the fact that hitting $\lfloor bt \rfloor$ before time t has an exponential cost for any particle in the initial condition within distance $O(t)$ of the origin, and, due to (25), there is a subexponentially large number of such particles. Second, in the worst case where all the particles attached to sites $1 \le x \le bt$ are turned into X particles instantaneously at time zero, the cost of preventing all these particles from hitting bt up to time t is of order $\exp(-t^{1/2+o(1)})$, due to the lower bound in Theorem 10 (c) discussed above. The actual proof is in fact more complex since we want to consider probabilities of the form $\mathbb{P}(ct \le r_t \le bt)$, and not only $\mathbb{P}(r_t \le bt)$, and deal also with the case $s(\eta) < 1$.

By far the more difficult part of Theorem 10 is (b). The proof strategy is based on the sub-additivity property of the hitting times $T(u, v)$ stated in (31). Given $m \ge 1$, let $\chi_j := T(mj, m(j + 1))$. By sub-additivity, we have that

$$T(0, n) \le \sum_{j=0}^{\lfloor n/m \rfloor} \chi_j,$$

so that

$$\mathbb{P}(T(0, n) \ge cn) \le \mathbb{P}\left(\sum_{j=0}^{\lfloor n/m \rfloor} \chi_j \ge (mc)\lfloor n/m \rfloor\right). \tag{46}$$

Now, by translation invariance, for all $j \ge 0$, χ_j and $\chi_0 = T(0, m)$ have the same distribution, and it can be shown that

$$\lim_{m\to+\infty} m^{-1}\mathbb{E}(T(0, m)) = v^{-1}.$$

Hence, given $c > v^{-1}$ we can always find $m \geq 1$ such that $mc > \mathbb{E}(\chi_0)$, so that the r.h.s. of (46) is the probability of a large deviation above the mean for the sum $\sum_{j=0}^{\lfloor n/m \rfloor} \chi_j$. We then seek to apply large deviations bounds for i.i.d. variables in order to estimate this probability. Of course, the random variables $(\chi_j;\ j \geq 0)$ are *not* independent, but the dependency between $(\chi_j;\ j \leq j_1)$ and $(\chi_j;\ j \geq j_2)$ is weak when $j_2 - j_1$ is large. Indeed, for given j, χ_j mostly depends on the behavior of the random walks with initial locations close to mj. We implement this idea by using a technique already exploited in **[134]** in a similar context. Given $\ell \geq 1$, we define a family $(\chi'_j;\ j \geq 0)$ of hitting times as follows: χ'_j uses the same random walks as χ_j for particles initially located at sites x with $mj - m\ell < x < m(j+1)$, but uses fresh independent random walks for particles initially located at sites x with $x \leq mj - m\ell$. We can then prove that the following properties hold:

(a) For all $j \geq 0$, the family $(\chi'_{j+p(\ell+1)};\ p \geq 0)$ is i.i.d.;

(b) when ℓ is large, the probability that $\chi'_j = \chi_j$ is close to 1.

We can thus obtain estimates on the r.h.s. of (46) by estimating separately the probability that $\chi'_j = \chi_j$ for all $0 \leq j \leq \lfloor m/n \rfloor$, and the probability that $\sum_{j=0}^{\lfloor n/m \rfloor} \chi'_j \geq (mc)\lfloor n/m \rfloor$. Now, thanks to property (a) above, this last sum can be split evenly into $\ell + 1$ subsums of i.i.d. random variables distributed as $\chi_0 = T(0, m)$. Controlling the tail of $T(0, m)$ then allows us to apply large deviation bounds for i.i.d. variables separately to each of these subsums. One of the issues is that, for fixed m, the tail of $T(0, m)$ does not decay exponentially fast, but satisfies $\mathbb{P}(T(0, m) \geq t) = e^{-O(t^{1/2})}$ instead, so that non-standard large deviations estimates have to be used. Another issue is that, if one applies the above approach naively, optimizing the choice of ℓ as a function of n leads to a bound of order $e^{-n^{2/7+o(1)}}$. To achieve the $e^{-n^{1/3+o(1)}}$ bound stated in the Theorem 10, one has to use a slightly more subtle argument, involving a positive association property between the large deviation event we consider and a suitable piece of the event that $\chi'_j = \chi_j$ for all j.

3. Proofs: KS infection model ($D_X > 0$, $D_Y > 0$)

When both D_X and D_Y are non-zero, the KS infection model is much more complicated to deal with than the stochastic combustion model, one reason being the absence of an exact sub-additivity property comparable to the one that holds for the stochastic combustion model, which is stated in (31).

3.1. The shape theorem when $D_X = D_Y$. The proof of Theorem 11 by Kesten and Sidoravicius involves two major steps. The first one, developed in **[100]**, consists in establishing ballistic upper and lower bounds for $\mathscr{B}_t$.

The upper bound, quoted above as Theorem 13, and which is also valid when $D_X \neq D_Y$, is obtained through a kind of Peierls argument. To an X (infected) particle present at time t, one associates the so-called genealogical path describing the succession of contacts between X and Y particles that led to this particle being infected, all the way back to the initial configuration.

One then computes the expected number of genealogical paths that may lead to X particles outside $t[-C, C]^d$ at time t, for C large enough.

We now quote the lower bound:

THEOREM 15 (Kesten, Sidoravicius **[100]**). *Consider the KS infection model on $\mathbb{Z}$ with $D_X = D_Y$, starting with an initial configuration consisting of i.i.d. Poisson numbers of Y particles at each site, and a finite (positive) number of X particles. There exists a constant $c > 0$ such that for all $K > 0$, and all large enough t,*

$$\mathbb{P}\left(\llbracket t[-c, c]^d \rrbracket \not\subset \mathscr{B}_t\right) \leq t^{-K}.$$

Unlike the upper bound, the proof of this result is very involved, using, at its core, a multi-scale renormalisation argument. To explain what the problem is, let us consider the case $d = 1$. Define $\mathscr{R}_t$ as the position of the right-most particle of type X at time t, i.e.

$$\mathscr{R}_t := \sup\{x;\ x \text{ bears a type } X \text{ particle at time } t\}. \tag{47}$$

When there is a single particle located at $\mathscr{R}_t$ at time t, the instantaneous drift of $\mathscr{R}_t$ is zero, since the particle has an equal probability to jump to the right or to the left, while, when there is more than one particle, this drift is to the right. In fact, even if there is a single particle located at $\mathscr{R}_t$ at time t, the presence of particles (of either type) "close to" $\mathscr{R}_t$ is enough to ensure a drift to the right, since these have a positive probability to reach $\mathscr{R}_t$ within a fixed amount of time. Thus, to prove a ballistic lower bound on $\mathscr{R}_t$, one should prove that, with high probability, there are particles "close" to $\mathscr{R}_t$ for a positive fraction of time. The difficulty then lies in the complex interaction between $\mathscr{R}_t$ and the nearby particles, which destroys the homogeneity of the initial Poisson distribution of particles. The approach developed in **[100]** consists in studying not the specific path $(\mathscr{R}_t)_{t\geq 0}$ followed by the right-most X particle, but every possible path (with mild constraints), showing that, with high probability, any such path will have some particle close to it most of the time. This is where a percolation-type argument is used, based on a partitioning of $\mathbb{Z} \times [0, +\infty[$ into space-time blocks at various scales. The assumption $D_X = D_Y$ is important here, since it allows one to consider the path followed by a particle as a standard random-walk path with jump rate $D_X = D_Y$, regardless of the type of the particle.

The second step of the proof of the shape theorem, developed in **[103]**, is based on an approximate sub-additivity property of the so-called half-space processes. Such approximate sub-additivity is enough to prove that, for each unit vector $u \in \mathbb{R}^d$, there exists a constant $\lambda(u)$ such that, almost surely,

$$\lim_{t\to+\infty} t^{-1} H(t, u) = \lambda(u),$$

where $H(t, u)$ is (up to some modification of the initial configuration) the maximum of $x \cdot u >$ over all sites $x \in \mathscr{B}_t$. One then deduces the convergence of $t^{-1}\mathscr{B}_t$ towards an asymptotic shape.

3.2. Fluctuations of the front in dimension one when $D_X = D_Y$.

3.2.1. *Renewal structure.* We now discuss the proof of Theorem 12. As in **[53]**, the proof is based on a renewal structure, and we look for a.s. finite random times $0 =: \kappa_0 < \kappa_1 < \kappa_2 < \dots$ such that

(i) the r.v.s $(\kappa_{n+1} - \kappa_n, r_{\kappa_{n+1}} - r_{\kappa_n})_{n\geq 0}$ are independent,
(ii) the r.v.s $(\kappa_{n+1} - \kappa_n, r_{\kappa_{n+1}} - r_{\kappa_n})_{n\geq 1}$ are identically distributed,
(iii) $\mathbb{E}(\kappa_2 - \kappa_1)^2 < +\infty$ and $\mathbb{E}(r_{\kappa_2} - r_{\kappa_1})^2 < +\infty$.

However, it is not possible to use the same definition of $(\kappa_n)_{n\geq 0}$, since, due to the fact that both X and Y particles move, the distribution of Y particles located above r_t at a time where the front jumps, is not fixed, and depends upon the whole past of the process. To solve this problem, we have to consider a modified version of the renewal structure, where conditions on the past of the process are added to the conditions contained in the definition of a forward α time. In fact, we extend the random walk trajectories of the particles infinitely far into the past, which is possible thanks to the reversibility property of the initial Poisson distribution of particles with respect to systems of independent random walks. Thus, every particle in the process is considered to have a random walk trajectory $(W_s)_{s\in\mathbb{R}}$ (note that we do not attempt to define the infection dynamics for negative times).

We now introduce a bit of notation. For $t > 0$, introduce the set $X(t)$ formed by the trajectories of particles that are of type X at time[5] t. Similarly, introduce the set $Y(t)$ formed by the trajectories of the particles that are of type Y at time t. We extend the definition[6] by letting $X(0)$ (resp. $Y(0)$) denote the set of trajectories of particles initially located at sites $x \leq -1$ (resp. $x \geq 0$).

Then let us say that a time t at which the front $\mathscr{R}_t$ performs an upward (i.e. $+1$) jump is a *backward sub-α time* if $\mathscr{R}_t > \alpha t$ and if, for all $0 \leq s < t$, one has $\mathscr{R}_s < \mathscr{R}_t - \alpha(t - s)$. Say that t is a *backward super-α time* if, for any particle in $Y(t)$, the corresponding trajectory, denoted $(W_s)_{s\in\mathbb{R}}$, is such that, for all $s \in]-\infty, t]$, one has $W_s \geq \mathscr{R}_t - \alpha(t-s)$. If t is both a backward sub-α and super-α time, we say that t is a *backward α time*. We also redefine the notion of a forward super-α and sub-α time, adding extra conditions for merely technical reasons. In this section, we say that t is a *forward super-α time* if, for every $s \geq t$, one has $\mathscr{R}_s \geq \mathscr{R}_t + \lfloor \alpha(s-t) \rfloor$ and if, moreover, there exists a particle in $Y(t)$ such that $W_s = \mathscr{R}_t$ for all $s \in [t, t + \alpha^{-1}]$. We say that t is a *forward sub-α time* if, for any particle whose location at time t is $\leq \mathscr{R}_t - 1$, the corresponding random walk trajectory, denoted $(W_s)_s$, satisfies $W_s \leq \mathscr{R}_s - 1 + \alpha(s-t)$ for all $s \geq t$, and if, in addition, the trajectory, denoted W^*, of the particle that makes the front jump at time t, satisfies $W^*_s = r_t$ for $s \in [t, t + \alpha^{-1}]$, and then satisfies the inequality $W^*_s \leq r_t - 1 + \alpha(s-t)$ for all $s \geq t + \alpha^{-1}$. We extend the definition of a backward super-α time and of a forward super-α by allowing $t = 0$ in the above definitions. As before, we say that t is a *forward α time* if it is both a forward sub-α and super-α time. Finally, we say that t is an α separation time if t is both a forward

[5]By convention, we consider that a particle of type Y whose first contact with an X particle is at time t, turns into type X at time $t+$; this means that we exclude from $X(t)$ those particles that may be turned from type Y to type X precisely at time t.

[6]With this definition, particles initially at site 0 start to be considered X particles at time $0+$.

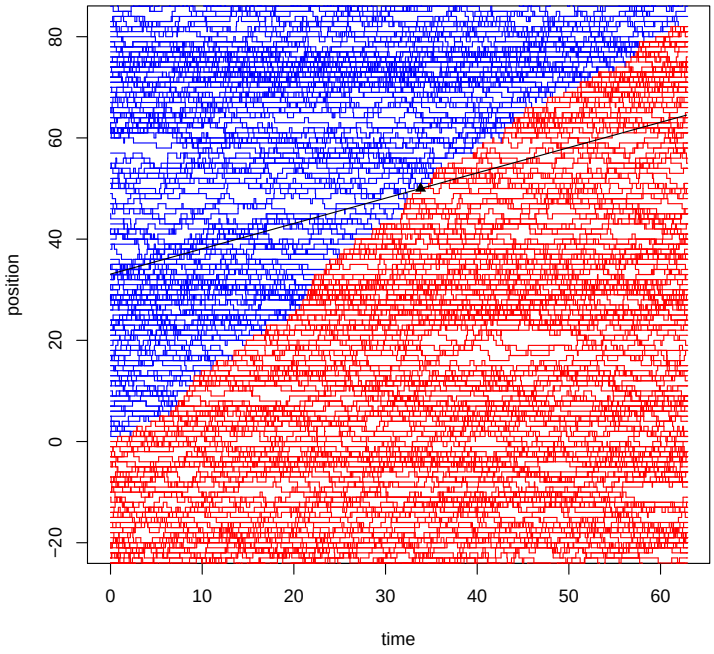


FIGURE 5. A realization of the KS infection model with an α separation time t. Red (resp. blue) trajectories correspond to particles of type X (resp. Y). The line of slope α associated with the forward α time is drawn in black, with a triangle showing the location of $(t, \mathscr{R}_t)$.

and backward α time. Note that, in the above definitions, we work with $(\mathscr{R}_s)_s$ instead of $(r_s)_s$, which is more convenient since it is the trajectory of $\mathscr{R}_t$ which defines the boundary that Y particles must avoid in order not to get infected (i.e. turned into X particles). As in the stochastic combustion case, what makes plausible the fact that α separation times exist is the fact that individual particles move diffusively (in both time directions) while the front moves ballistically. Figures 5 and 6 illustrate these definitions.

One can then define the renewal structure by $\kappa_0 := 0$ and

$$\kappa_{n+1} := \inf\{t > \kappa_n;\ t \text{ is an } \alpha \text{ separation time}\}. \tag{48}$$

The first observation is that, as in the stochastic combustion case, the definition of a forward α time implies that the particles in $X(\kappa_n)$ cannot have any influence upon the front posterior to κ_n, so that all the evolution of the front posterior to κ_n is due to the particles in $Y(\kappa_n)$. Moreover, the value of

$$(\kappa_{n+1} - \kappa_n, r_{\kappa_{n+1}} - r_{\kappa_n})$$

is identical to the value of (κ_1, r_{κ_1}) that would be obtained if we shifted the origin of space and time to (r_{κ_n}, κ_n), keeping only the particles in $Y(\kappa_n)$. Then, the key point leading to properties (i) and (ii) is that, *conditional upon the past history of the particles that are of type X at time κ_n*, the distribution

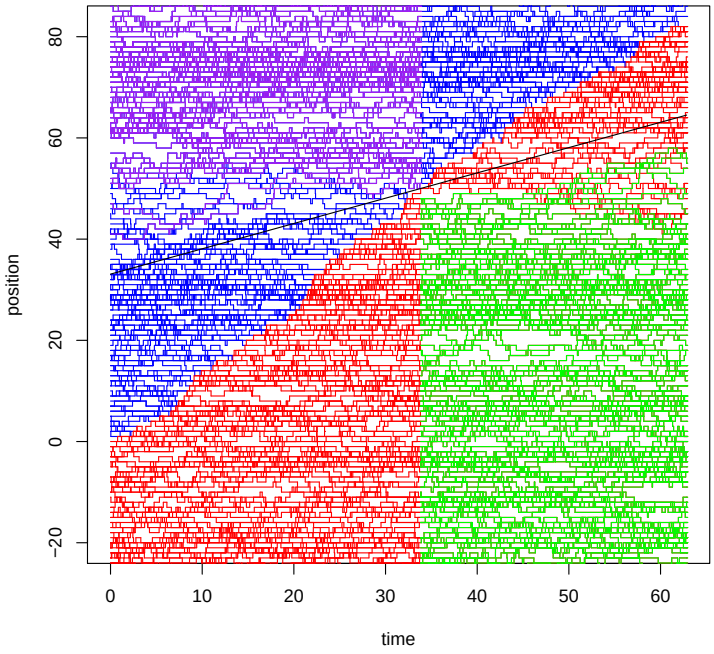


FIGURE 6. Same realization of the KS infection model as in Fig. 5. Red (resp. blue) trajectories correspond to particles of type X (resp. Y), except that, posterior (resp. prior) to the forward α time t, green (resp. purple) is used instead of red (resp. blue) to draw the trajectories of particles that lie below (resp. above) r_t at time t.

of the trajectories in $Y(\kappa_n)$ is completely determined, up to translation:

$$\mathbb{P}(\tau_{\kappa_n, r_{\kappa_n}}(Y(\kappa_n)) \in \cdot | \mathcal{F}^X_{\kappa_n}) = \mathbb{P}_0(Y(0) \in \cdot | A) \ a.s., \tag{49}$$

where $\tau_{t,x}$ denotes the space-time shift acting on trajectories, i.e. $\tau_{t,x}(W)_s := W_{s-t} - x$, $\mathcal{F}^X_{\kappa_n}$ denotes the $\sigma-$algebra generated by κ_n, r_{κ_n}, and the past history up to time κ_n of the particles that are of type X at time κ_n, $\mathbb{P}_0$ denotes the probability describing the KS infection model starting with i.i.d. numbers of Poisson particles at sites $x \geq 0$ and no particles at sites $x < 0$, while A denotes the event that $t = 0$ is a backward and forward super-α time.

The proof of (49) mainly relies on a time-reversal argument that we now sketch. Consider $t > 0$. Using the reversibility of the Poisson initial distribution of particles with respect to systems of independent random walks, we may consider building the trajectories of the particles in our model – regardless of the evolution of their type – by first putting i.i.d. Poisson numbers of particles at each site of $\mathbb{Z}$ at time $t-$, and then building random walk trajectories that extend in both time directions, starting from the positions of each of these particles (see Fig. 7). One can then construct the infection dynamics from these trajectories in the usual way (see Fig. 8).

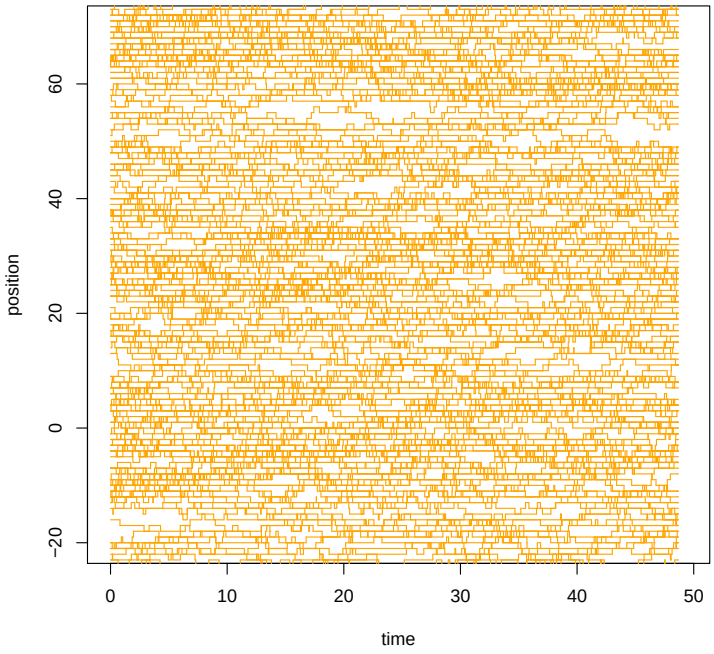


FIGURE 7. A stationary set of random walk trajectories on the interval $[0, t[$, with i.i.d. Poisson numbers of particles at each given time. One may obtain it by putting i.i.d. Poisson numbers of particles at time zero, then running independent random walks from time 0 to time t for each of these particles, or by putting i.i.d. Poisson numbers of particles at time $t-$, then running independent random walks in the reverse time direction, from time $t-$ to time 0.

We now want to understand the effect of conditioning this construction by the history up to time t, of the particles that are of type X at time t. Denote by $\mathscr{H}$ the specific[7] realization of the history under consideration, and by $(q_s)_{0\le s\le t}$ the corresponding history of the front (note that knowing $\mathscr{H}$ uniquely determines the history of the front up to time t). Assume moreover that t is an upward jump time[8] for the front, i.e. $q_t = q_{t-} + 1$. In our context, the particles that are of type X at time t are precisely those whose location at time $t-$ is $\le q_t - 1$, and their trajectories up to time t correspond exactly to $\mathscr{H}$, see Figures 9, 10. On the other hand, the particles whose location at time $t-$ is $\ge q_t$ coincide with the particles that are of type Y at time t, and

[7]For the sake of simplicity, we chose to ignore the technical difficulties associated with conditioning in the present context, and work as if the random variables involved were discrete.

[8]This is where the distinction between $t-$ and t becomes relevant in the discussion. $\mathscr{H}$ is equivalent to the history of particles up to time $t-$ except for the jump at time t of the particle that makes the front climb precisely at time t. This is the only particle whose location at time t differs from its location at time $t-$.

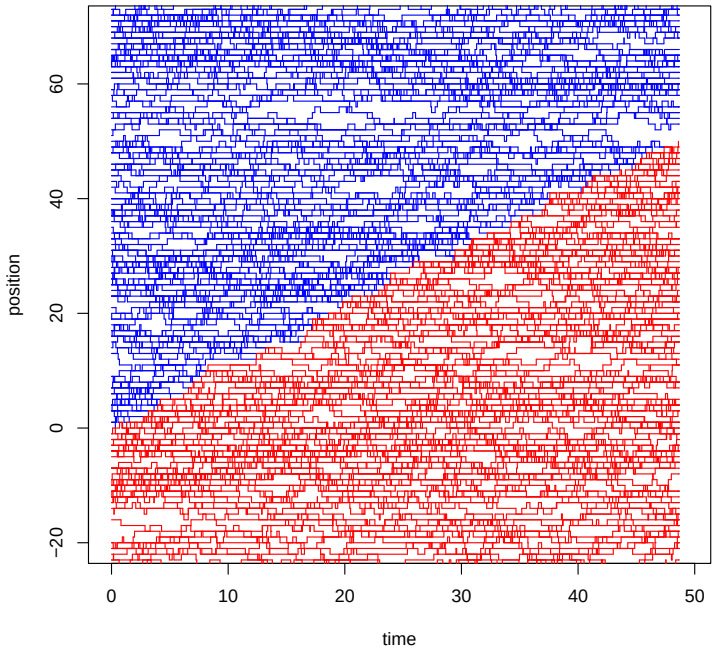


FIGURE 8. Realization of the KS infection model on the interval $[0, t]$, based on the trajectories shown on Fig. 7. Red (resp. blue) trajectories correspond to particles of type X (resp. Y).

their trajectories from time 0 to time t must avoid the front $(q_s)_{0 \le s \le t}$ and have a location at time 0 that is > 0 (for otherwise they would be of type X at time t). It turns out that these properties are enough to characterize the history being $\mathscr{H}$, i.e. the fact that $\mathscr{H}$ is the history up to time t of the particles that are of type X at time t is equivalent to having the intersection of the following two events:

(a) for every particle whose location at time $t-$ is $\ge q_t$, the corresponding trajectory W is such that $W_s > q_s$ for all $s < t$, and $W_0 > 0$;
(b) the history up to time t of the particles whose location at time $t-$ is $\le q_t - 1$, is given by $\mathscr{H}$.

Figure 12 illustrates what (a) means for particles whose location at time $t-$ is $\ge q_t$, while Figure 10 shows how (b) specifies the history up to time t of the particles whose location at time $t-$ is $\le q_t - 1$.

One key property is now that, prior to conditioning by (a) and (b), the two sets of trajectories whose locations at time $t-$ are $\le q_t - 1$ and $\ge q_t$ respectively, are independent. As a consequence, one sees that the conditional distribution of the trajectories of particles in $Y(q_t)$ given $\mathscr{H}$ corresponds to the one obtained by putting i.i.d. Poisson numbers of particles at each site $x \ge q_t$, building random walk trajectories starting from their positions, and then conditioning all these random walks W by $W_0 > 0$ and

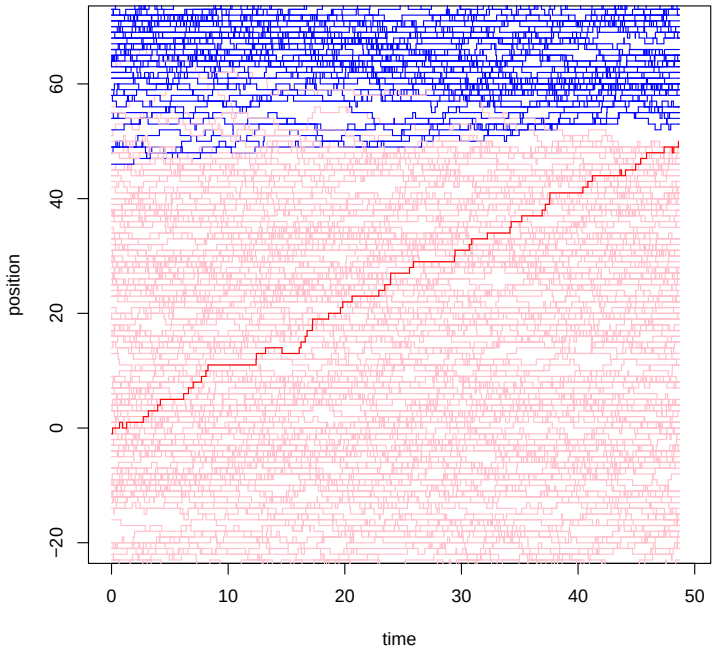


FIGURE 9. Same realization as in Fig. 8. Pink trajectories correspond to particles that are of type X *at time* t, while blue trajectories correspond to particles that are of type Y *at time* t. The trajectory of the front is drawn in red.

$W_s > q_s$ for all $s < t$. If t is a backward sub-α time (which can be told from $\mathscr{H}$), condition (a) is implied by the fact that t is a backward super-α time, since then, before time t, the trajectories of particles in $Y(t)$ must lie above a straight line which itself lies above $(q_s)_{0\le s\le t}$. One deduces that the conditional distribution of the trajectories of particles in $Y(q_t)$ given $\mathscr{H}$, and given the fact that t is a backward α time, corresponds to the one obtained by putting i.i.d. Poisson numbers of particles at each site $x \ge q_t$, building random walk trajectories starting from their positions, and then conditioning all these random walks W by $W_s > q_t - \alpha(t-s)$ for all $s < t$.

To complete the proof of (49), it remains to check that[9] as far as the trajectories of particles in $Y(\kappa_n)$ are concerned, the definition of κ_n does not induce extra conditions beyond being a backward and a forward super $\alpha-$time.

3.2.2. *Tail estimates on the renewal structure.* To prove condition (iii) (together with a.s. finiteness) and thus complete the proof of the central limit

[9]This is not immediate, since, as in the stochastic combustion case, one has to take care of the conditions on the indefinite future contained in the definitions of $\kappa_1, \dots, \kappa_n$. Here, we also have to deal with the fact that, for $1 \le i \le n-1$, κ_i contains conditions on the past of the trajectories in $Y(\kappa_i)$, which has a non-empty intersection with $Y(\kappa_n)$.

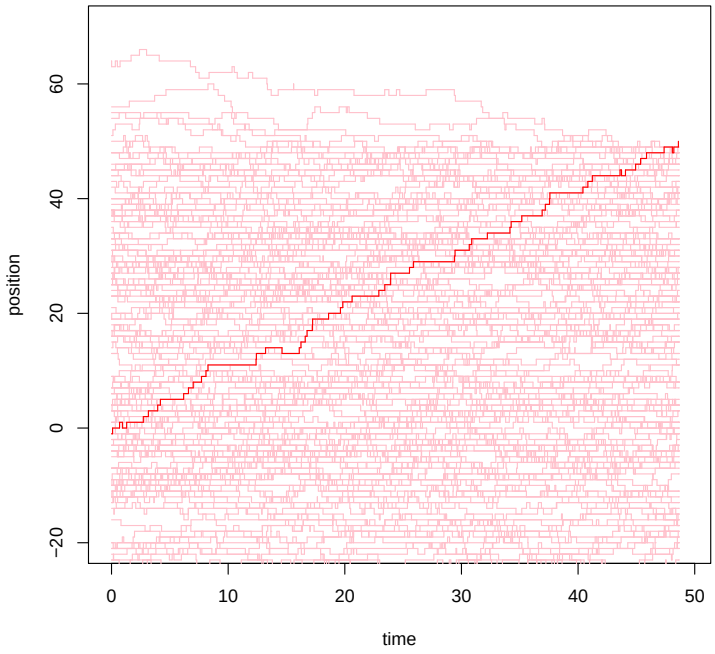


FIGURE 10. Same realization as in Fig. 8. Only the history $\mathscr{H}$ of particles that are of type X *at time* t is shown, with trajectories drawn in pink. The trajectory of the corresponding front is drawn in red.

theorem, one relies, as in **[53]** on a sequence of stopping times corresponding to successive attempts at producing α-separation times, that we now define[10].

We first introduce the following refinement of the notion of backward sub-α time: given $0 \leq s < t$, we say that t is an $(s, \alpha)-$*crossing time* if there exists $k \in \{1, 2, \ldots\}$ such that $\mathscr{R}_v < \mathscr{R}_s + k + \alpha(v - s)$ for all $v \in [s, t[$ and $\mathscr{R}_t \geq \mathscr{R}_s + k + \alpha(t - s)$. Note that if s is a backward sub-α time and if t is an $(s, \alpha)-$crossing time, then t is also a backward sub-α time.

We now define by induction the sequence of stopping times on which our estimates on the renewal structure are based. Besides α, the definition involves two integer parameters $\mathscr{C} \geq 1$ and $L \geq 1$. Let $D_0 := 0$ and $\Upsilon_0 := \emptyset$. For $n \geq 1$, assume that the random variables D_{n-1}, Υ_{n-1} have already been defined, and let S'_n be the infimum of the $t > D_{n-1}$ such that

- t is a backward sub-α time;
- $\Upsilon_{n-1} \subset X(t)$;
- $Y(t)$ contains at least $\mathscr{C}$ particles located at $\mathscr{R}_t$ at time t.

Then define S_n as the infimum of the $t > S'_n$ such that

- t is a backward sub-α time;

[10]The definition is slightly different from **[35]**, due to the use of labels for particle trajectories, which we do not discuss here.

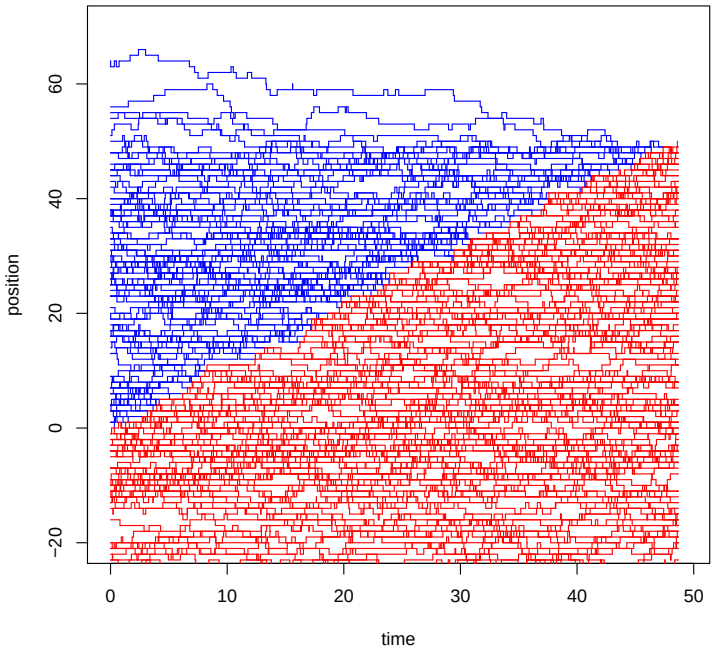


FIGURE 11. Same realization as in Fig. 8. The usual representation of the process (red for type X, blue for type Y), when only particles in $\mathscr{H}$ are present.

- $]S'_n, t[$ contains a number of $(S'_n, \alpha)-$crossing times at least equal to L;
- B_t contains at least $\mathscr{C}$ particles located at $\mathscr{R}_t$ at time t.

We use the notation W^{*n} for the trajectory of the particle that makes the front jump at time S_n, and define the subset $X(S_n)^* := X(S_n) \setminus \{W^{*n}\}$. If S_n is a backward super-α time, then $\Upsilon_n := \emptyset$ and D_n is defined as the infimum of the $t > S_n$ such that *at least one* of the following five conditions holds:

(1) $\mathscr{R}_t < \mathscr{R}_{S_n} + \lfloor \alpha(t - S_n) \rfloor$
(2) $t \leq S_n + \alpha^{-1}$ and there is no particle in $Y(S_n)$ whose trajectory W satisfies $W_{S_n} = \mathscr{R}_{S_n}$ and remains at $\mathscr{R}_{S_n}$ during $[S_n, S_n + t]$,
(3) $W_t > \mathscr{R}_{S_n} - 1 + \alpha(t - S_n)$ for some particle in $X(S_n)^*$,
(4) $t \leq S_n + \alpha^{-1}$ and $W^{*n}_t \neq \mathscr{R}_{S_n}$,
(5) $t > S_n + \alpha^{-1}$ and $W^{*n}_t > \mathscr{R}_{S_n} - 1 + \alpha(t - S_n)$,

On the other hand, if S_n is not a backward super-α time, consider the set of particles in $X(S_n)$ whose trajectories W are such that there exists a time $t < S_n$ for which $W_t < \mathscr{R}_{S_n} - \alpha(S_n - t)$. Among these particles, choose the one whose trajectory has the lowest coordinate at time S_n (breaking ties in an arbitrary manner), and, denoting its trajectory by $W^{(n)}$, let $\Upsilon_n := \{W^{(n)}\}$ and $D_n := S_n$.

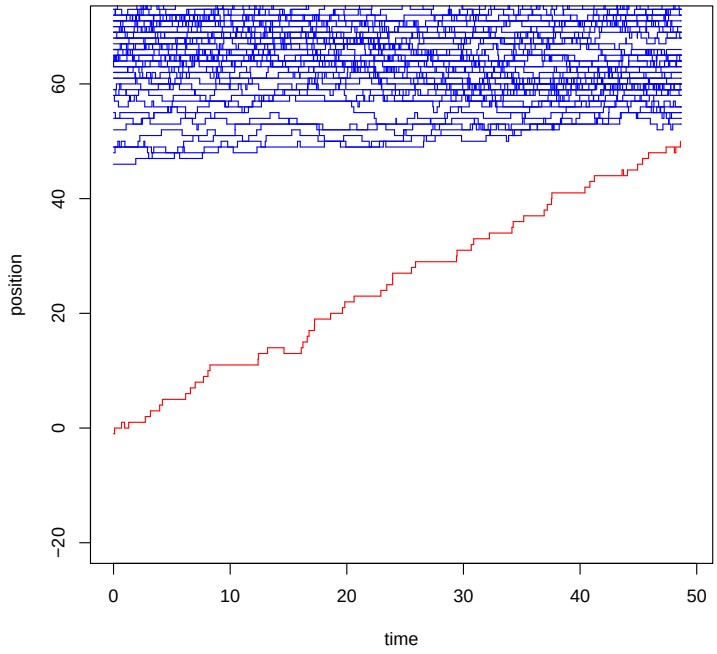


FIGURE 12. Same realization as in Fig. 8. The history of particles that are of type Y at time t, drawn in blue. The front is drawn in red.

The above set of definitions is a bit technical, so we now give a few explanations about the underlying ideas. The stopping times S_n defined above are our successive candidates to produce α separation times. Since by definition S_n is a backward sub-α time, one has to check whether, in addition, S_n is indeed a backward super-α time, a forward super-α time, and a forward sub-α time. When either of these properties fails, D_n is the earliest time at which such a failure can be detected. Indeed, $S_n = D_n$ when S_n fails to be a backward super-α time, this condition bearing only on the history of the process prior to time S_n. When S_n is a backward super-α time, conditions (1) and (2) in the definition of D_n detect the potential failure of S_n to be a forward super-α time, while (3)-(4)-(5) detect the potential failure of S_n to be a forward sub-α time. After time D_n, one has to wait until suitable conditions are met again, leading to the next candidate time S_{n+1}. These conditions include that S_{n+1} be a backward super-α time, but also that the number of particles located at $\mathscr{R}_{S_{n+1}}$ is large enough, and that a sufficient number of α crossings have been performed since time D_n. In addition, when S_n fails to be a backward super-α time, a suitably chosen "witness" $W^{(n)}$ has to be absorbed into the set $X(S_{n+1})$. Fig. 13, 14, 15, 16 illustrate these definitions in the main cases.

Introducing

$$\mathfrak{K} := \inf\{n \geq 1;\ D_n = +\infty\},$$

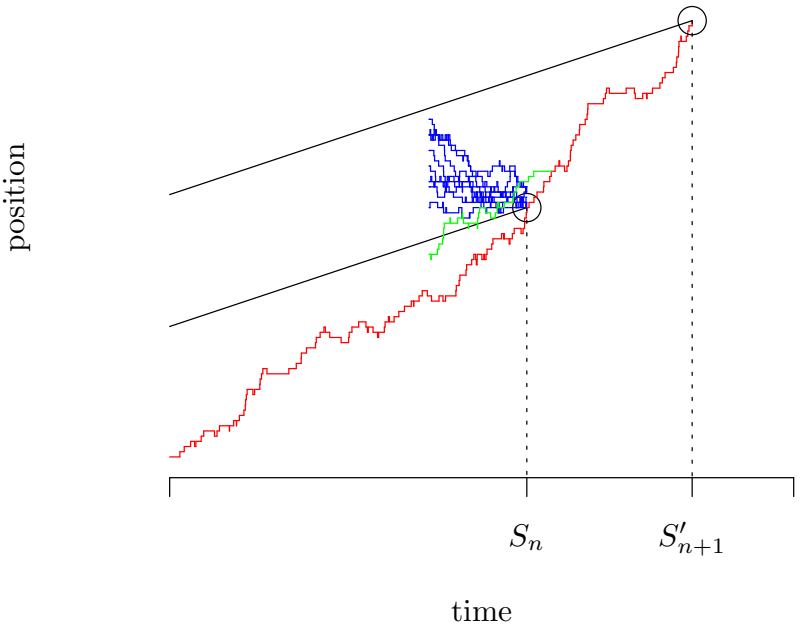


FIGURE 13. From S_n to S'_{n+1} when S_n fails to be a backward super-α time (in this case $D_n = S_n$). Only the most relevant portions of trajectories are shown. The trajectory of the front $\mathscr{R}_t$ is depicted in red, while blue is used for particles of type Y, except for the "witness" trajectory $W^{(n)}$, which is drawn in green. Circles are used at locations where the number of particles is assumed to be $\geq \mathscr{C}$.

we see that, when $\mathfrak{K} < +\infty$, $S_{\mathfrak{K}}$ is an α separation time, so the goal is to prove that $\mathfrak{K} < +\infty$ a.s., with also $\mathbb{E}(S^2_{\mathfrak{K}}) < +\infty$ and $\mathbb{E}(\mathscr{R}^2_{S_{\mathfrak{K}}}) < +\infty$, in order to prove the desired estimates on the random variables κ_n. We do not go into the technical details here, but at least give a overview of the proof strategy.

Our estimates make use of three main ingredients that we now describe.

The first ingredient is a description of the conditional distribution of $Y(S_n)$ with respect to[11] $\mathcal{F}^X_{S_n}$, based on the fact that S_n is a backward sub-α time, and obtained by exploiting a time-reversal argument similar to the one

[11]Here $\mathcal{F}^X_{S_n}$ is the $\sigma-$algebra generated by S_n, $\mathscr{R}_{S_n}$, and the trajectories up to time S_n of the particles that are of type X at time S_n

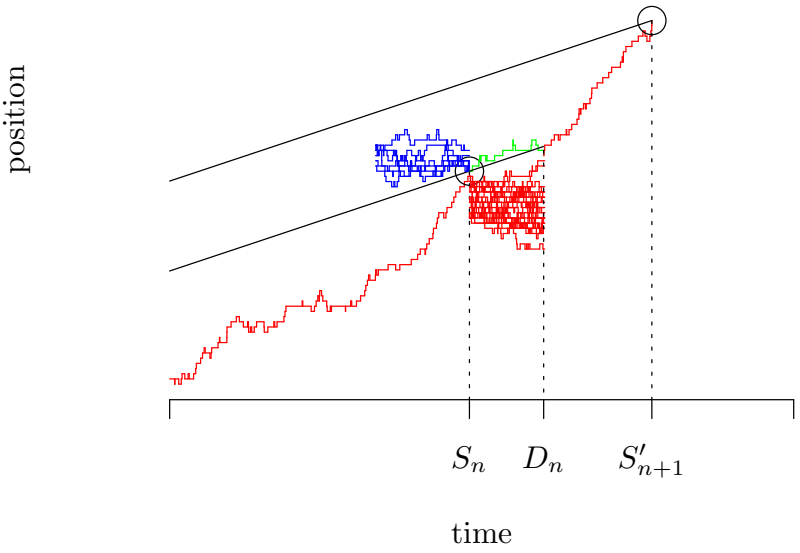


FIGURE 14. From S_n to S'_{n+1} when S_n is a backward super-α time but condition (1) is realized first. Only the most relevant portions of trajectories are shown. The trajectory of the front $\mathscr{R}_t$ is depicted in red, except for the part causing condition (1), which is drawn in green. Otherwise, blue (resp. red) is used for particles of type Y (resp. X). Circles are used at locations where the number of particles is assumed to be $\geq \mathscr{C}$.

leading to the proof of (49). Specifically, we have an identity of the form

$$\mathbb{P}(\tau_{S_n, r_{S_n}}(Y(S_n)) \in \cdot | \mathcal{F}^X_{\kappa_n}) = \mathbb{P}_0(Y(0) \in \cdot | B), \tag{50}$$

where B is a random event depending on $\mathcal{F}^X_{\kappa_n}$ satisfying

$$B \supset B_0 := \{t = 0 \text{ is a backward super } \alpha \text{ time}\} \cap \{\Xi_0 = 1\}, \tag{51}$$

where Ξ_0 is the indicator function of the event that there are at least $\mathscr{C}$ particles at site 0 at time $t = 0$. In words, (50) and (51) show that the conditional distribution of $Y(S_n)$ with respect to $\mathcal{F}^X_{S_n}$ can be compared (up to a translation) to the distribution obtained by putting i.i.d. Poisson numbers of particles at sites $x \geq 0$, conditioned by the event B_0, whose probability with respect to $\mathbb{P}_0$ is strictly positive. Note that this is to ensure such a property that the sets Υ_k have been introduced. Indeed, for each $k \leq n - 1$

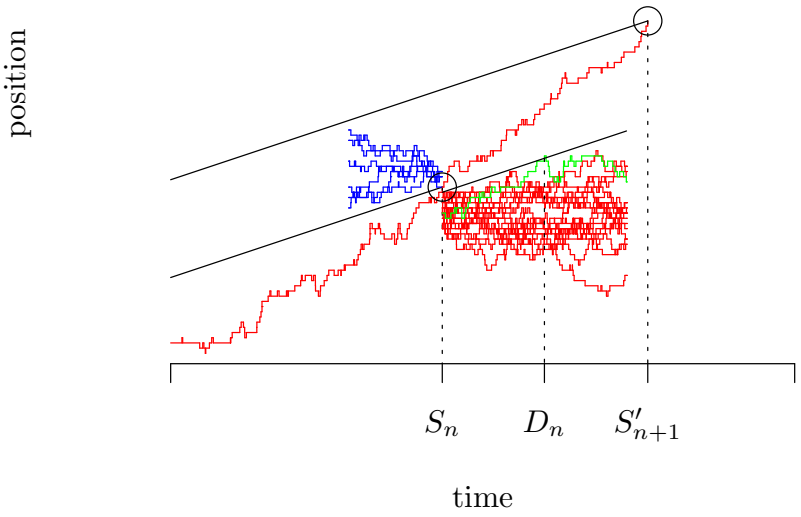


FIGURE 15. From S_n to S'_{n+1} when S_n is a backward super-α time but condition (3) is realized first. The trajectory of the front $\mathscr{R}_t$ is depicted in red, which is also used for particles of type X, except the trajectory causing condition (3), which is drawn in green. Otherwise, blue is used for particles of type Y. Circles are used at locations where the number of particles is assumed to be $\geq \mathscr{C}$.

such that S_k fails to be a backward super-α time, Υ_k contains a "witness" of this failure that is later absorbed in the set $X(S_n)$, so that, using only the information available in $\mathcal{F}^X_{S_n}$, we can e.g. tell which of the S_ks are backward super-α times, and which are not.

Another ingredient is a ballistic lower bound for the front after time S_n. Remember that, in the stochastic combustion case, one could rely on the auxiliary front to provide such bounds, while in the present case, the only available bound is the one by Kesten and Sidoravicius (quoted earlier as Theorem 15). Despite the fact that this bound is stated only for a homogeneous initial configuration comprising i.i.d. Poisson numbers of particles both at the left and at the right of the origin, it is possible to derive from it (using coupling and a symmetrization trick) a version that works for the kind of initial condition we have here, i.e. with a control bearing only on the

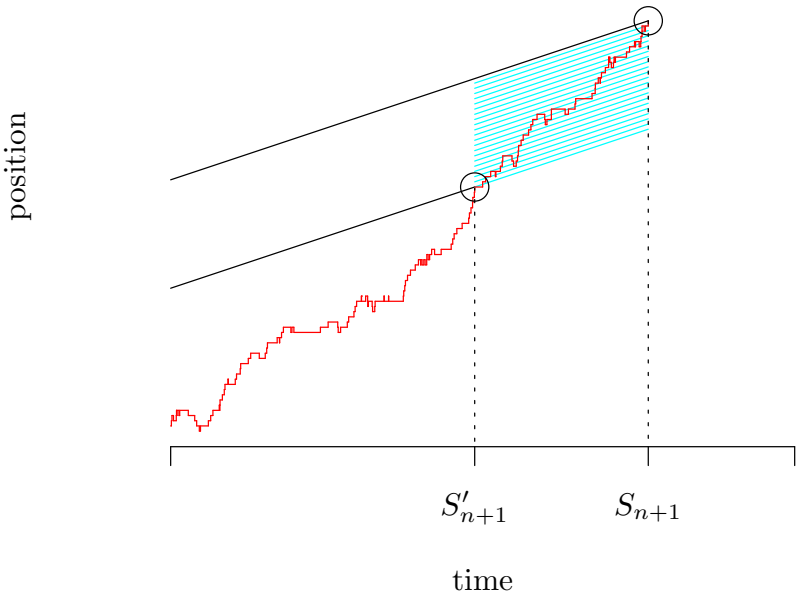


FIGURE 16. From S'_{n+1} to S_{n+1} when S_n fails to be an α separation time. The trajectory of the front $\mathscr{R}_t$ is depicted in red, and light blue lines illustrate the successive α-crossings performed by the front. Circles are used at locations where the number of particles is assumed to be $\geq \mathscr{C}$.

distribution of particles at the right of the origin (given by (50) and (51)). This is where the condition that there are at least $\mathscr{C}$ particles plays a role, and, specifically, we can prove that there exists $\beta > 0$, and $c_1, c_2 > 0$, where c_1 depends on $\mathscr{C}$, such that, for every $t > 0$,

$$\mathbb{P}(\mathscr{R}_{S_n+t} \leq \mathscr{R}_{S_n} + \beta t | \mathcal{F}^X_{S_n}) \leq c_1 t^{-c_2 \cdot \mathscr{C}} \ a.s. \tag{52}$$

Finally, as in **[53]**, an important tool to control the behaviour of particles located to the left the front, is the exponential norm

$$\mathcal{M}_t := \sum_{x \leq \mathscr{R}_t - 1} \eta_t(x) e^{-\theta(x - \mathscr{R}_t)}, \tag{53}$$

where $\eta_t(x)$ denotes the number of particles (which must be of type X) located below $\mathscr{R}_t$ at time t, used in combination with martingale estimates. Although we do not enter into the details, let us point out that the way we handle $\mathcal{M}_t$ differs substantially from **[53]**, where a very involved inductive scheme is used to control the time it takes for $\mathcal{M}_t$ to get below a specific

threshold. Using a softer and hopefully more transparent argument (where (50) and (52) play a key role), we are able to prove an estimate of the form

$$\mathbb{E}_\nu(\mathcal{M}_{S_{n+1}}\mathbf{1}(D_n < +\infty)|\mathcal{F}^X_{S_n}) \leq c_3 e^{-\theta L}\mathcal{M}_{S_n} + c_4, \tag{54}$$

where c_3 is a strictly positive constant depending on $\mathscr{C}$, and c_4 is a strictly positive constant depending on $\mathscr{C}$ and L. This is where the choice of a large enough L plays a role, since we can then use (54) to obtain a uniform bound on the expectation of $\mathcal{M}_{S_n}$ conditional upon the fact that $\mathfrak{K} \geq n$.

With the three above ingredients, one can then prove that $\mathfrak{K} < +\infty$ a.s., with also $\mathbb{E}(S^2_{\mathfrak{K}}) < +\infty$ and $\mathbb{E}(\mathscr{R}^2_{S_{\mathfrak{K}}}) < +\infty$, leading to the desired estimates for the random variables κ_n. Broadly speaking, (50), (52), and (54) respectively help bounding from below the probability that, conditional upon $\mathfrak{K} \geq n$, S_n is a backward super-α time, a forward super-α time, and a forward sub-α time. When either of these properties fail, they also lead to bounds on the tail of $W^{(n)}_{S_n} - \mathscr{R}_{S_n}$ (when S_n fails to be a backward super-α time), or $D_n - S_n$ (when S_n fails to be a forward α time), from which bounds on the tail of $S'_{n+1} - S_n$, then $S_{n+1} - S_n$ follow, using ballisticity of the front.

3.3. Extension to the case $D_X > D_Y$. The proof of Theorem 14 consists in an extension of the arguments leading to the proof of Theorem 12. To this end, we consider a construction of the dynamics with $D_X > D_Y$ that uses random walk trajectories with a constant jump rate equal to D_Y. As long as a particle is of type Y, it follows the corresponding trajectory in the usual way, while, as soon as it is turned into a particle type X, it starts following the trajectory with a speed multiplied by a factor D_X/D_Y. As before, we denote by $(W_s)_s$ the actual trajectory followed by a particle, while the trajectory with constant jump rate equal to D_Y from which this trajectory is constructed is denoted $(\mathcal{W}_s)_s$. Figures 17 and 18 illustrate this construction.

Remember that the first key element of the proof in the case $D_X = D_Y$ is a description of the conditional distribution of the trajectories of particles that are of type Y at time t, given the past history of the particles that are of type X at time t, which is itself based on a time-reversal argument. We now explain how a similar point can be made in the case where $D_X > D_Y$. Let us redo the time-reversal argument, putting i.i.d. Poisson numbers of particles on the sites of $\mathbb{Z}$ at time $t-$, running for each particle an independent random walk trajectory with constant jump rate equal to D_Y, extending in both time directions. Again we want to understand the impact on this construction of conditioning by the past history up to time t of the particles that are of type X at time t. Let us consider an even more general problem, that of conditioning by the full history (i.e. $(W_t)_{t\in]-\infty,+\infty[}$) of the particles that are of type X at time t. Note that it is equivalent to condition by the full history of the trajectories $(\mathcal{W}_t)_{t\in]-\infty,+\infty[}$ associated with the same particles, since we can infer the history of the front up to time t, hence of the individual jump rates of the particles, starting from any of these two sets of trajectories. Call $\mathscr{H}'$ the specific history of the $\mathcal{W}$ trajectories by which we condition, and $(q_s)_{0\leq s\leq t}$ the corresponding history of the front. We assume not only that $q_t = q_{t-} + 1$, but also that t is in fact a record time for the front, i.e. $q_t > \sup_{0\leq s<t} q_s$. By definition, particles that are of type X at time t are

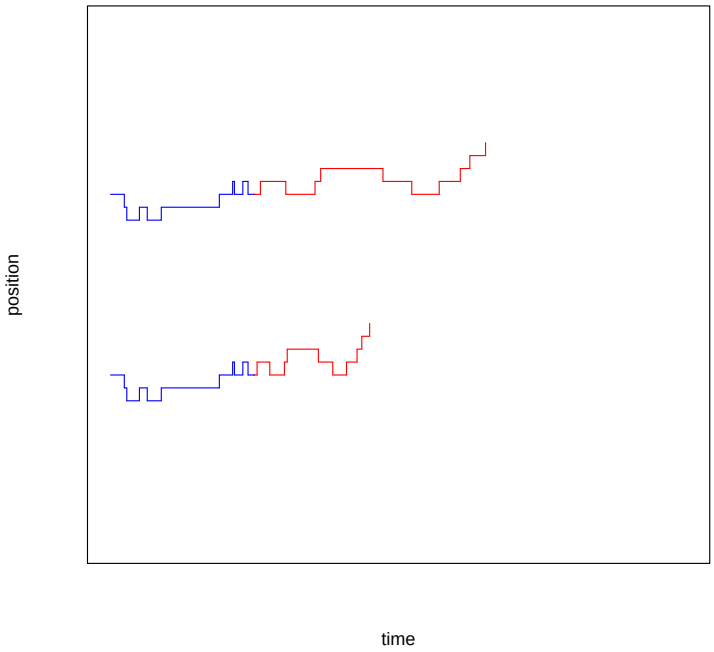
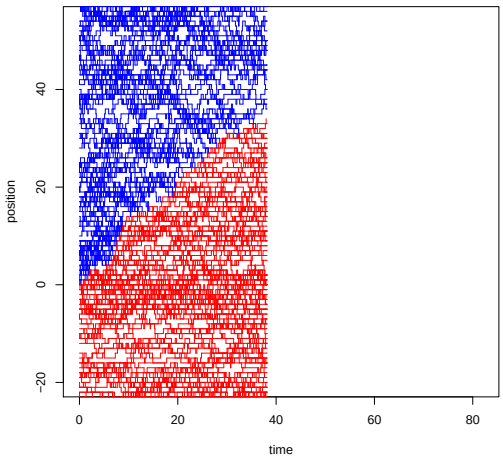


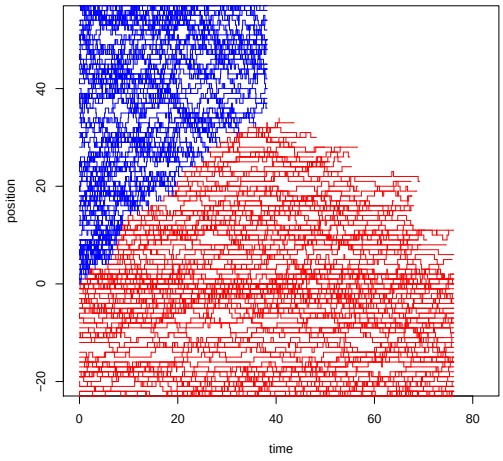
FIGURE 17. Comparison of $(\mathcal{W}_s)$ (above) and (W_s) (below). Blue (resp. red) correspond to type Y (resp. X).

those for which $W_{t-} < q_t$, while particles that are of type Y at time t are those for which $W_{t-} \geq q_t$. For particles that are of type Y at time t, one has by definition $W_t = \mathcal{W}_t$, so that all particles that are of type Y at time t satisfy $\mathcal{W}_{t-} \geq q_t$. It turns out that it is also true that particles that are of type X at time t satisfy $\mathcal{W}_{t-} < q_t$, despite the fact that in general $\mathcal{W}_t \neq W_t$ for these particles. Indeed, particles that are of type X at time t must have been turned into X particles strictly prior to time t, so that, for each such particle, $\mathcal{W}_{t-} = W_s$ for some s that is both $< t$ and posterior to the time at which the particle was turned into an X particle, whence $\mathcal{W}_{t-} \leq \mathscr{R}_s < \mathscr{R}_t$, where we have used the fact that t is assumed to be a record time for the front. With these observations, we can redo exactly the same argument as in the case $D_X = D_Y$, and conclude that the conditional distribution of the trajectories of particles in $Y(q_t)$ given $\mathscr{H}'$ and given the fact that t is a backward α time, corresponds to the one obtained by putting i.i.d. Poisson numbers of particles at each site $x \geq q_t$, building random walk trajectories starting from their positions, and then conditioning all these random walks W by $W_s \geq q_t - \alpha(t-s)$ for all $s \leq t$.

Another key element of the proof in the case $D_X = D_Y$ is the ballistic lower bound, deduced from that obtained in [**100**]. Unfortunately, we are not able to prove such a bound for the original KS infection model with $D_X > D_Y$. However, for the remanent version of the model, it is easily checked that $\mathscr{R}_t \geq \check{\mathscr{R}}_t$, where $\check{\mathscr{R}}_t$ is the position of the right-most X particle



(a) (W_s) trajectories



(b) $(\mathcal{W}_s)$ trajectories

FIGURE 18. Realization of the KS infection model with $D_X = 2$ and $D_Y = 1$. The actual evolution of the process is shown in (a), where the trajectories (W_s) are drawn with the usual convention that red (resp. blue) is for type X (resp. type Y). The corresponding evolution of the trajectories $(\mathcal{W}_s)_s$ is shown in (b).

for the infection dynamics whose particle trajectories are given by $(\mathcal{W}_s)_s$ instead of (W_s), which corresponds to the original KS infection dynamics in which both particle types have jump rate D_Y. The ballistic lower bound of **[100]** then applies to $\check{\mathscr{R}}_t$, yielding the required ballisticity bound on $\mathscr{R}_t$.

The proof of Theorem 14 then follows that of Theorem 12, with rather minor changes.

4. Discussion

Condition (25) in Theorem 9 is sharp in the sense (see Theorem 3 in **[36]**) that, if it fails for some $\theta > 0$, one cannot have $I \equiv 0$ on $[0, v]$, while, if it fails for a sufficiently large θ, the law of large numbers with limiting speed v breaks down in such a way that one cannot have $I > 0$ on $]v, +\infty[$. Anyhow, every "physically reasonable" initial configuration of X particles should satisfy condition (25).

On the other hand, the assumption that the initial configuration of Y particles consists of a fixed number $\mathfrak{a}$ of Y particles per site seems restrictive. In fact, most available results for the stochastic combustion model do assume such an initial configuration (one exception is **[2]**, which proves a shape theorem for general random i.i.d. initial configurations, in the discrete-time case). Nevertheless, it is quite likely that the approach of **[53]**, based on a renewal structure, can be extended without too many difficulties to random initial configurations, say containing i.i.d. Poisson numbers of Y particles at each site, which would lead to a central limit theorem for the front comparable to Theorem 8. The study of large deviations probabilities for the front starting with random initial configurations may however reveal interesting new phenomena, since one then has to take into account the contribution of very unlikely but very inhomogeneous initial configurations.

For the stochastic combustion model, the exact order of magnitude of slowdown large deviations probabilities, i.e. probabilities that are of the form $\mathbb{P}\left[\frac{r_t}{t} \leq b\right]$, where $0 \leq b < v$, is not known. Indeed, Theorem 10 shows that, for homogeneous initial configurations, this probability lies between $\exp(-t^{1/2+o(1)})$ and $\exp(-t^{1/3+o(1)})$ for large t, but it is unclear whether one of these bounds is sharp, or whether the true order of magnitude lies strictly between them. On the one hand, for extreme slowdown events, i.e. $r_t \leq 0$, Theorem 10 shows that $\exp(-t^{1/2+o(1)})$ is the correct order of magnitude. On the other hand, the $\exp(-t^{1/3+o(1)})$ order of magnitude for slowdown probabilities is reminiscent of the behaviour observed for one-dimensional random walk in random environment with positive or zero drift in the annealed case, see **[130, 58]**, which clearly has some similarities with the stochastic combustion model (in our model, the position of the rightmost X particle has positive or zero drift, depending on whether there is more than a single particle at its current location). One may note that the proof strategy consisting in bounding large deviations of $T(0, n)$ by large deviations of $\sum_{j=0}^{\lfloor n/m \rfloor} T(mj, m(j+1))$ (thanks to sub-additivity) for fixed m, cannot in any case deliver a smaller order of magnitude than $\exp(-t^{1/3+o(1)})$.

Theorems 12 and 14 are established for an initial condition consisting of i.i.d. Poisson particles, of type Y (resp. X) to the right (resp. to the left) of the origin, and it is natural to ask whether more general conditions can be considered. However, as in **[100, 103]**, the i.i.d. Poisson distribution of particles plays a central role in the arguments, so it is unclear whether the result can be generalized beyond easy extensions, such as conditioning the configuration of X particles by a non-zero probability event.

As noted in Section 3, the missing element that would be needed to apply the proof of Theorem 14 to the original (non-remanent) KS infection

model, is a ballistic lower bound comparable to Theorem 15. Unfortunately, we do not see at the moment how to obtain such a bound. One frustrating aspect of the problem is that bounds that are just slightly sub-ballistic are available, showing that with large probability one must have $r_t \geq t(\log t)^{-p}$ for some $p > 0$. This indicates that the front moves much faster than the particles surrounding it at any given time, so that some sort of renewal should occur for the configuration of particles surrounding the front, from which a truly ballistic lower bound should follow. Despite the intuitive appeal of this idea (see also Chapter 3), we could not adapt the definition of the renewal structure to accommodate for sub-ballistic lower bounds: the use of straight lines seems hard to circumvent, and straight lines (if only in the definition of a backward sub-α time) make it necessary to use *a priori* ballistic bounds, not just slightly sub-ballistic ones. On the other hand, in the $D_X = D_Y$ case, one may note that, since the proof of Theorem 14 uses only the upper and lower bounds proved in **[100]** (Theorems 13 and 15), not the full asymptotic shape theorem (Theorem 11) proved in **[103]**, our approach gives an alternative way to derive the $d = 1$ case of Theorem 11 from the results of **[100]**.

Clearly one of the most important questions surrounding $X + Y \to 2Y$ models is the behaviour of fluctuations in dimensions $d \geq 2$. One might speculate that, for $d = 2$, these fluctuations are described by the KPZ equation, which would lead to an order of magnitude of $t^{-1/3}$, instead of the $t^{-1/2}$ observed in dimension $d = 1$, but we know of no reasonably detailed heuristic argument supporting such a conjecture. We have not yet tried to apply the renewal structure idea to dimensions $d \geq 2$, but clearly this cannot be done through a simple generalization of what has been developed in the $d = 1$ case. The problem is that, in contrast with the $d = 1$ case, and in contrast also with multi-dimensional situations where renewal methods have proved useful, such as the study of random walk in random environments or self-interacting random walks, one does not have to keep track of just a single trajectory (that of the front for $X+Y \to 2X$ models when $d = 1$, that of the particle for random walk models), but of the whole interface between X and Y particles. Note that, for a different growth model based on random walk trajectories on $\mathbb{Z}^d$, namely the internal diffusion-limited aggregation (IDLA) model, several results on the fluctuations around the asymptotic shape have been obtained when $d \geq 2$, see **[6, 8, 7, 94]**.

Variations upon the stochastic combustion model have been considered, which contain additional interaction rules between particles. For instance, **[93]** considers the case where X particles no longer move independently of each other, but according to a simple exclusion process, and use the renewal structure approach to establish a central limit theorem for the initial position of the front (see also **[52]**, where another kind of saturation mechanism is considered). It would be interesting to extend the results obtained for the KS infection model with $D_X = D_Y$ to such situations (e.g. by imposing exclusion or zero-range interaction between particles).

A combination of theoretical arguments and numerical evidence, see **[118]** suggests that, for a given density of Y particles in the initial condition, the asymptotic speed of propagation of the front (whose very existence is not proved mathematically) in the KS infection model should have the simple

expression

$$v = c \cdot D_X \tag{55}$$

for some constant c, which has the remarkable feature of not depending on D_Y. Proving such a result seems out of reach (at the moment) of renewal methods, which succeed in proving the existence of an asymptotic speed, but do not provide an explicit formula for it, leading merely to the fact that

$$v = \frac{\mathbb{E}_0(r_{\kappa_1}|A)}{\mathbb{E}_0(\kappa_1|A)}. \tag{56}$$

Still, (56) may provide interesting insights into the dependence of v with respect to the model parameters, and such fascinating conjectures as (55), even if only approximately true, certainly make a case for devoting more efforts towards a rigorous understanding of the value of v.

One key object in the study of random walks in random environment and tagged particle processes, is the environment viewed from the particle. For one-dimensional $X+Y \to 2Y$ models, the natural analog is the environment viewed from the front, i.e. the Markov process (Z_t) defined by

$$Z_t := (\eta_t(x - r_t))_{x\in\mathbb{Z}}. \tag{57}$$

Note that the renewal structure can be used to prove e.g. the ergodicity of (Z_t) (see **[53]** for the stochastic combustion case). On the other hand, it is natural to ask whether it is possible to bypass the renewal structure and directly analyse the environment viewed from the front, to prove e.g. the central limit theorem via a Kipnis-Varadhan type approach (see **[104, 106]**). To our knowledge, this approach has not been brought to fruition, one first obvious problem being that no explicit form of the invariant distribution is available.

For the one-dimensional modified DLA model, which is a slight variant of the case $D_X = 0$, $D_Y > 0$ aimed at producing a spatially growing cluster of X particles, a phenomenon which is not expected to hold when both $D_X > 0$ and $D_Y > 0$ appears, namely, a phase transition with respect to the density of particles in the initial configuration. Indeed, Kesten and Sidoravicius proved in **[102]** that, if one starts with i.i.d. Poisson numbers of Y particles of mean $0 < \mu < 1$, one does not have that $r_t \propto t$ for large t, but rather that $r_t \propto t^{1/2}$. When $\mu > 1$, it is expected, based on theoretical arguments and numerical evidence, that $r_t \propto t$ if $\mu > 1$. More generally, in the whole parameter space where $D_X > 0$ and $D_Y > 0$, it is expected that a behaviour similar to the case $D_X = D_Y > 0$ holds, i.e. positive asymptotic speed, and gaussian $t^{-1/2}$ fluctuations. Under the remanent assumption, Theorem 14 shows that it is indeed the case when $D_X > D_Y$. However, we have no idea of how to attack the case where $D_X < D_Y$.

In the context of random walks in a random environment created by an interacting particle system, general laws of large numbers have recently been obtained, see **[9, 59]**, using renewal techniques which bear some resemblance to the ones we used (see also **[136]** and the references therein). It is unclear whether an approach similar to the one developed in **[9, 59]** can be applied to the context we have studied, due to the rather poor mixing properties of the particle environment.

We conclude with a variation of the KS infection model for which a different kind of question has been investigated. The variation consists in adding to the reaction $X + Y \to 2X$ a recovery reaction (from infected to healthy) $X \to Y$ with constant rate $\lambda > 0$. In the case $D_X = D_Y$, [**101**] show that there is a critical value $0 < \lambda_c < +\infty$ such that X particles disappear when $\lambda > \lambda_c$, while they survive if $\lambda < \lambda_c$. In the stochastic combustion case, the situation is more complex, since it depends on the density μ of Y particles in the initial condition. Indeed, [**101**] proves that for large enough μ (depending only on the dimension), X particles do survive regardless of the value of λ. In the one-dimensional case, [**138**] more precisely prove that, given λ, there exists $\mu_c \in [\lambda/(1+\lambda), 1]$ such that X particles survive locally when $\mu > \mu_c$, while they locally disappear when $\mu < \mu_c$ (i.e. any finite subset of $\mathbb{Z}$ a.s. does not contain X particles after a certain time).

CHAPTER 3

Excited random walks

1. Introduction

1.1. Model(s). A discrete-time homogeneous Markov chain $(X_n)_{n\geq 0}$ on a discrete set V is characterized by the fact that, given the history of the chain up to time n – i.e. given the sequence $X_0, \ldots, X_n$ –, the step leading from X_n to X_{n+1} is described by a probability distribution on V of the form

$$\mathbb{P}(X_{n+1} = \cdot | X_0, \ldots, X_n) = \omega(X_n, \cdot), \tag{58}$$

which depends exclusively on X_n. *Excited random walks*, also known as *cookie random walks*, are a class of self-interacting random walks in which, in addition to the dependence on the current location of the walk X_n, the probability distribution of the next step is allowed to depend on the number of times the current location of the walk has been visited in the past. In other words, (58) is replaced by

$$\mathbb{P}(X_{n+1} = \cdot | X_0, \ldots, X_n) = \omega(X_n, L_n(X_n), \cdot), \tag{59}$$

where L_n denotes the local time of the walk up to time n, i.e.

$$L_n(x) := \sum_{k=0}^{n} \mathbf{1}(X_n = x). \tag{60}$$

In the sequel, we deal almost exclusively with nearest-neighbour random walks on $\mathbb{Z}^d$, with $d \geq 1$, so we can use the additive structure of $\mathbb{Z}^d$ to label the random walk steps. Indeed, denoting by $(e_i)_{1\leq i\leq d}$ the canonical basis of $\mathbb{Z}^d$, and by $\mathcal{E}$ the family of unit vectors of $\mathbb{Z}^d$, i.e. $\mathcal{E} := \{\pm e_i;\ 1 \leq i \leq d\}$, we can write the transition probabilities of the walk under the form

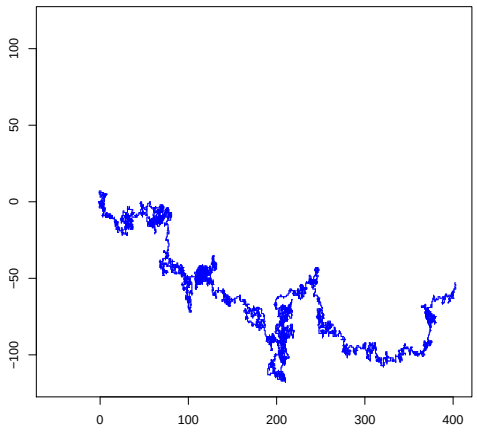
$$(\omega(x, \ell, e);\ x \in \mathbb{Z}^d,\ \ell \in \{0, 1, 2, \ldots\},\ e \in \mathcal{E}),$$

with the following version of (59):

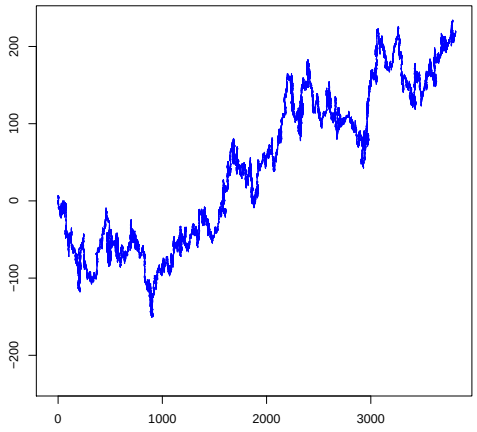
$$\mathbb{P}(X_{n+1} = x + e | X_0, \ldots, X_n) = \omega(X_n, L_n(X_n), e). \tag{61}$$

The original excited random walk model, introduced by Benjamini and Wilson in **[19]**, corresponds to the case where the random walk has a fixed positive bias (excitation) in the e_1 direction when it first hits a site, but no bias in the other directions, while it behaves like a simple symmetric random walk on $\mathbb{Z}^d$ when it hits a previously visited site. In other words, there is a parameter $0 < p \leq 1/2$ such that

$$\omega(x, \ell, e) = \begin{cases} \frac{1+p}{2d} \text{ if } \ell = 1 \text{ and } e = +e_1, \\ \frac{1-p}{2d} \text{ if } \ell = 1 \text{ and } e = -e_1, \\ \frac{1}{2d} \text{ otherwise.} \end{cases}$$



(a) Excited random walk, first 10^4 steps



(b) Excited random walk, first 10^5 steps

FIGURE 1. Realization of the original excited random walk model on $\mathbb{Z}^2$ with bias parameter $p = 0.2$.

To allow for a more general type of directional bias, consider a non-zero vector $u \in \mathbb{R}^d$, and say that $\omega(x, \ell, \cdot)$ is:

- $u-$positive if $\sum_{e\in\mathcal{E}} \omega(x, \ell, e)e \cdot u \geq 0$,
- $u-$strictly positive if $\sum_{e\in\mathcal{E}} \omega(x, \ell, e)e \cdot u > 0$,
- balanced if $\omega(x, \ell, e) = \omega(x, \ell, -e)$ for all $e \in \mathcal{E}$.

For instance, the multi-dimensional excited random walk considered in by Menshikov, Popov, Ramírez and Vachkovskaia in [**120**] corresponds to the case where there exists a vector u such that, for all $x \in \mathbb{Z}$, $\omega(x, 1, \cdot)$ is $u-$strictly positive, while $\omega(x, \ell, \cdot)$ is balanced, for all $\ell \geq 2$.

The term "cookie random walk" was introduced by Zerner [**156**] in the context of excited random walks on $\mathbb{Z}$ for which $\omega(x,\ell,+1) \geq 1/2$ (i.e. $\omega(x,\ell,\cdot)$ is 1-positive) for all x,ℓ. This rather enigmatic name is explained by a colourful description of the corresponding model, which goes as follows. Initially, each site $x \in \mathbb{Z}$ bears a (possibly) infinite stack of cookies. Then, each time the random walk hits a site x, and if the corresponding stack is non-empty, the walker eats the cookie lying on the top of the stack, which provides her/him with a bias to the right for the next step. In the absence of cookies at site x, a simple symmetric random walk step is performed. Note that, by allowing cookies to produce negative or zero bias as well, one may always assume that each site bears a countably infinite stack of cookies.

Using this cookie terminology (in dimension $d \geq 2$), the original excited random walk model of Benjamini and Wilson corresponds to the case where there is one cookie per site $x \in \mathbb{Z}^d$, each cookie producing a fixed positive bias in the e_1 direction. In the sequel, we use the term *positive cookies random walk* to refer to the case where, for some fixed u, $\omega(x,\ell,\cdot)$ is $u-$positive for all x,ℓ, while we use the term *boundedly many cookies random walk* for the case where there exists an $M \geq 1$ such that, for all $\ell \geq M+1$, $\omega(x,\ell,\cdot)$ is balanced. The family $(\omega(x,\ell,e);\ x \in \mathbb{Z}^d,\ \ell \in \{0,1,2,\ldots\},\ e \in \mathcal{E})$ is called the *cookie environment*, and we consider the case where the cookie environment is either deterministic and translation-invariant, i.e. $\omega(x,\ell,\cdot) = \omega(0,\ell,\cdot)$, or drawn at random, under the assumption that the cookie environments at distinct sites form an i.i.d. family. When we speak of positive cookies random walk or boundedly many cookies random walk in random environment, we always assume that one can choose u and M deterministically in such a way that the corresponding properties hold for almost every environment.

1.2. Results. Most available results have been obtained for nearest-neighbour excited random walks on $\mathbb{Z}$, thanks to the existence of an alternative representation of the walk in terms of branching processes with immigration. Our own work deals with the case of excited random walks on $\mathbb{Z}^d$ for $d \geq 2$, which, due to the unavailability of the branching process representation, turns out to be rather remote from the $d = 1$ case. As a consequence, our emphasis in the sequel is on the multi-dimensional case, and we content ourselves with a brief overview of the one-dimensional case, which includes only partial results and does not do justice to the large body of work devoted to the subject. Also, we do not discuss walks on other types of graphs such as trees (see [**122, 152, 14**]), strips (see [**157, 63**]), nor excited Brownian motions (see [**133, 132**]). For a more thorough exposition, we refer to the very nice recent review paper by Kosygina and Zerner [**108**].

1.2.1. *Case* $d = 1$. Over the years, a fairly detailed picture of the behaviour of the boundedly many cookies random walk on $\mathbb{Z}$ in either deterministic or i.i.d. environment has emerged, and we quote the main results below. Note that the theorems quoted in the sequel under a general form have often been established first in special cases, e.g. under the assumption of positive cookies, or with respect to a deterministic environment. Since we do not discuss the details of these successive improvements, we quote the authors of the partial results along those of the final one.

The key quantity to characterize the behaviour of the model is the average drift over the whole sequence of cookies at a site, i.e.

$$\delta := \mathbb{E}\left(\sum_{\ell \geq 1} \omega(0,\ell,1) - \omega(0,\ell,-1)\right). \tag{62}$$

where $\mathbb{E}$ refers to the joint probability measure describing both the possibly random environment and the random evolution of the walk conditional upon the realization of the environment. We also introduce the following weak ellipticity assumption:

$$\mathbb{P}(\forall \ell \geq 0,\ \omega(0,\ell,1) > 0) > 0 \tag{63}$$

We start with a recurrence/transience criterion. Here, recurrence means that each site $x \in \mathbb{Z}$ is visited $\mathbb{P}-$almost surely an infinite number of times, while transience to the right (resp. to the left) means that $\mathbb{P}-$almost surely, $\lim_{n\to+\infty} X_n = +\infty$ (resp. $-\infty$).

THEOREM 16 (Zerner [**156**], Kosygina and Zerner [**109**]). *For the boundedly many cookies case in i.i.d. random environment, under the ellipticity assumption* (63)*, the random walk is*

- *recurrent if* $-1 \leq \delta \leq 1$,
- *transient to the right if* $\delta > 1$,
- *transient to the left if* $\delta < -1$.

Then one has the following law of large numbers.

THEOREM 17 (Zerner [**156**], Mountford, Pimentel, Valle [**122**], Basdevant and Singh [**12**], Kosygina and Zerner [**109**]). *For the boundedly many cookies case in i.i.d. random environment, under the ellipticity assumption* (63)*, one has a law of large numbers with deterministic speed* v*:*

$$\lim_{n\to+\infty} n^{-1} X_n = v,\ \mathbb{P}-a.s.$$

Moreover, one has that

- $v = 0$ *if* $-2 \leq \delta \leq 2$,
- $v > 0$ *if* $\delta > 2$,
- $v < 0$ *if* $\delta < -2$.

We then describe the remarkably rich range of possible behaviours for the fluctuations of the model, mentioning only the order of magnitude of the fluctuations and the type of the limiting distribution.

THEOREM 18. *For the boundedly many cookies case in i.i.d. random environment, under the ellipticity assumption* (63)*, the various fluctuations regimes of the random walk are described by Table 1 below.*

1.2.2. *Case* $d \geq 2$. We now discuss the (comparatively) far less complete results available in the $d \geq 2$ case, starting with the recurrence/transience question.

THEOREM 19 (Benjamini, Wilson [**19**]). *For the original excited random walk model in dimension* $d \geq 2$*, the walk is transient in the* $+e_1$ *direction, i.e.*

$$\lim_{n\to+\infty} X_n \cdot e_1 = +\infty\ \mathbb{P}-a.s.$$

Range	*Order of magnitude*	*Limiting distribution*	*Reference(s)*
$0 \leq \delta < 1$	$X_n \propto n^{1/2}$	*Brownian motion perturbed at extrema*	**[63, 62]**
$\delta = 1$	$X_n \propto n^{1/2} \log n$	*running max. of Brownian motion*	**[62]**
$1 < \delta < 2$	$X_n \propto n^{\delta/2}$	*stable,* $\alpha = \delta/2$	**[13, 107]**
$\delta = 2$	$X_n - \gamma_n \propto n(\log n)^{-2}$, $\gamma_n \propto \log n$	*stable,* $\alpha = 1$	**[13, 107, 108]**
$2 < \delta < 4$	$X_n - vn \propto n^{2/\delta}$	*stable,* $\alpha = \delta/2$	**[107]**
$\delta = 4$	$X_n - vn \propto (n \log n)^{1/2}$	*stable,* $\alpha = 2$ *(Gaussian)*	**[107]**
$\delta > 4$	$X_n - vn \propto n^{1/2}$	*stable,* $\alpha = 2$ *(Gaussian)*	**[109]**

TABLE 1. Fluctuation regime of the one-dimensional excited random walk as a function of δ.

To state the next result, introduce the analog of the quantity δ defined in the case $d = 1$, which is the average drift vector defined as

$$\delta := \mathbb{E}\left(\sum_{e \in \mathcal{E}} \omega(0, 1, e)e\right). \tag{64}$$

Introduce also the uniform ellipticity assumption that, for some $\epsilon > 0$,

$$\mathbb{P}\left(\inf_{e \in \mathcal{E}} \omega(0, 1, e) \geq \epsilon\right) = 1. \tag{65}$$

THEOREM 20 (Zerner **[157]**). *Consider the positive cookies model in i.i.d. random environment, under the uniform ellipticity assumption* (65)*. Then the fact that* $\delta \cdot u > 0$ *implies that the walk is transient in direction* u*, i.e.*

$$\lim_{n \to +\infty} X_n \cdot u = +\infty \ \mathbb{P} - a.s.$$

An important consequence of Theorem 20 (and of Theorem **[19]** in the less general case of the original excited random walk model) is the existence, when $\delta \neq 0$, of an almost surely finite renewal structure for the walk, similar to the one used in the context of random walks in random environments (see **[144]**). A precise definition of the renewal structure in the present context is given in Subsection 2.3. In turn, the very existence of this renewal structure is enough (see e.g. **[155]**) to prove a law of large numbers of the form

$$\lim_{n \to +\infty} n^{-1} X_n = v, \ \mathbb{P} - a.s. \tag{66}$$

and the relevant question is then whether $v \neq 0$, i.e. whether the random walk is ballistic.

For the original excited random walk model in dimension $d \geq 4$, ballisticity is easily proved, as noted by Benjamini and Wilson in **[19]**. Proofs of ballisticity for $d = 3$, then $d = 2$, were then given by Kozma in two so far unpublished manuscripts **[110, 111]**. Using a different approach, we succeeded **[34]** in obtaining (stretched exponential, see (74)) tail estimates for

the renewal structure in all dimensions $d \geq 2$, leading to an alternative proof of ballisticity.

THEOREM 21 (Kozma **[110, 111]**, B. and Ramírez **[34]**). *For $d = 2$ and $d = 3$, the original excited random walk model is ballistic in the e_1 direction.*

Another consequence of the tail bounds proved in **[34]** on the renewal structure, is the following central limit theorem.

THEOREM 22 (B., Ramírez **[34]**). *Consider the original excited random walk model. For all $d \geq 2$, there exists a non-degenerate $d \times d$ covariance matrix C, such that*

$$t \mapsto n^{-1/2}(X_{\lfloor nt \rfloor} - v\lfloor nt \rfloor),$$

converges in law as $n \to +\infty$ to a Brownian motion with covariance matrix C, with respect to the Skorohod topology on the space of càdlàg functions.

Note that, using a completely different approach based on the lace expansion technique, van der Hofstad and Holmes **[147]** also proved a central limit theorem for the original excited random walk model, valid for dimensions $d \geq 8$ and sufficiently small excitation parameter p (depending on the dimension). One of the interests of the lace expansion approach is that it allows one, see **[146]**, to prove that the asymptotic speed in the e_1 direction, i.e. $v \cdot e_1$, is a monotonic increasing function of the excitation parameter p, provided that $d \geq 9$.

Recently, Menshikov, Ramírez, Popov and Vachkovskaia **[120]** found a way of extending the estimates on the renewal structure to a more general class of excited random walks. Specifically, they consider an i.i.d. cookie environment for which the first cookie satisfies an assumption of uniform $u-$strict positivity, i.e. there exist u and $\epsilon > 0$ such that

$$\mathbb{P}\left(\sum_{e \in \mathcal{E}} \omega(0,1,e) e \cdot u \geq \epsilon\right) = 1. \tag{67}$$

The subsequent cookies are assumed to be balanced, i.e.

$$\mathbb{P}\left(\omega(0,\ell,\cdot) \text{ is balanced}\right) = 1 \text{ for all } \ell \geq 2. \tag{68}$$

THEOREM 23 (Menshikov, Popov, Ramírez and Vachkovskaia **[120]**). *Assuming* (67), (68), *and the uniform ellipticity assumption* (65), *one has that, for all $d \geq 2$, the excited random walk is ballistic, with asymptotic speed $v \neq 0$, and there exists a non-degenerate $d \times d$ covariance matrix C, such that, with respect to $\mathbb{P}$,*

$$t \mapsto n^{-1/2}(X_{\lfloor nt \rfloor} - v\lfloor nt \rfloor),$$

converges in law as $n \to +\infty$ to a Brownian motion with covariance matrix B, with respect to the Skorohod topology on the space of càdlàg functions.

So far, we have only discussed the positive cookies case. The general case where cookies may produce both positive and negative drift in a given direction is not well understood, especially in low dimensions.

The following example, given in **[109]**, illustrates that the situation is more complex than in the one-dimensional case, where a simple classification of the possible behaviours can be given, based solely on the value of δ. In the

example, each site bears two cookies, the first one with bias $\epsilon > 0$ in the e_1 direction, the second one with an opposite bias $-\epsilon$ in the same direction. For this example, $\delta = 0$, but it can be shown with a specific argument that, for $d \geq 4$, a law of large numbers holds with a limiting speed satisfying $v \cdot e_1 > 0$. Moreover, by symmetry, reversing the order of the two cookies has the effect of turning v into $-v$, while keeping $\delta = 0$.

For large values of d, it is possible to adapt the argument – based on the so-called cut-times of the simple symmetric random walk – given by Bolthausen, Sznitman and Zeitouni [**39**] in the context of random walks in random environment. A law of large numbers for the excited random walk can thus be proved in high enough dimensions (at least 6), without $u-$positivity assumptions, see [**89**]. Combined with lace expansion estimates, this approach was used in [**90**] to show that, in dimension $d \geq 9$, a sufficiently large drift in the e_1 direction provided by the first cookie cannot be offset by drifts in the opposite direction provided by later cookies, no matter how large their drift.

Finally, let us mention [**3**], where in dimension $d = 3$, the excitation provides a drift towards a reflecting "wall" formed by a plane, making the walk recurrent, and [**18**], where, in dimension $d = 4$, the effect of the cookies is not to produce bias, but a balanced step restricted to dimensions e_1, e_2, while, in the absence of cookies, the steps of the walks are restricted to dimensions e_3, e_4 (and balanced too), making the walk transient.

2. Proofs

2.1. The approach of Benjamini and Wilson [19]. In [**19**], Benjamini and Wilson observed that a key quantity in the study of the excited random walk model is the number of distinct sites visited by the walk before time n, i.e.

$$R_n := \#\{X_k;\ 0 \leq k \leq n-1\}.$$

Indeed, call $I_n^{(1)}$ the sum of the R_n steps performed before time n by the walk, which start from a site visited for the first time. Call $I_n^{(2)}$ the sum of the remaining $n - R_n$ steps, so that one has

$$X_n = I_n^{(1)} + I_n^{(2)}.$$

Since the steps that define $I_n^{(1)}$ are positively biased in the e_1 direction, the law of large numbers shows that, when R_n is large,

$$I_n^{(1)} \cdot e_1 \sim cR_n,$$

where c is a positive constant. On the other hand, the steps defining $I_n^{(2)}$ are unbiased, and their number is at most n, so that one has that[1], at most,

$$I_n^{(2)} \cdot e_1 \propto n^{1/2}.$$

[1]We are not cheating here. On the one hand, the event that the first visit of X_k by the walk happens precisely at time k, is measurable with respect to the past history of the walk up to time k. On the other hand, conditional upon this past history, the step leading from X_k to X_{k+1} is either biased or unbiased, depending on whether X_k has been visited prior to time k or not.

As a consequence, if one can show that, for large n, one typically has that

$$R_n >> n^{1/2},$$

the contribution of $I_n^{(2)}$ to $X_n \cdot e_1$ is negligible compared to that of $I_n^{(1)}$, so that

$$X_n \cdot e_1 \sim cR_n.$$

In dimension $d \geq 3$, one can use the fact that the projection of the walk on the directions $(e_2, \ldots, e_d)$ is[2] a simple symmetric random walk on $\mathbb{Z}^{d-1}$ to provide lower bounds on the order of magnitude of R_n. Indeed, the number R'_n of distinct sites visited by the projection up to time n, is a lower bound for R_n. Moreover, the large n behaviour of the number of distinct sites visited before time n by a $(d-1)-$dimensional simple symmetric random walk is a well-studied object.

Specifically, we have that $R'_n \propto n$ when $d \geq 4$, while $R'_n \propto n(\log n)^{-1}$ when $d \geq 3$. This is enough to show ballisticity when $d \geq 4$, and transience in the e_1 direction (but not ballisticity) when $d = 3$. On the other hand, when $d = 2$, one has that $R'_n \propto n^{1/2}$, so the order of magnitude of R'_n is comparable to that of $I_n^{(2)}$, and the above argument does not lead to a definite conclusion regarding X_n.

To deal with the $d = 2$ case, Benjamini and Wilson **[19]** devised a clever argument based on a natural coupling of the excited random walk with a simple symmetric random walk, and the notion of *tan points*. The coupled random walk (Z_n) performs exactly the same steps as the excited random walk (X_n) in direction e_2, while, for steps in direction e_1 it only satisfies the inequality $(Z_{n+1} - Z_n) \cdot e_1 \leq (X_{n+1} - X_n) \cdot e_1$. Such a coupling is possible since the excited random walks steps in direction e_1 always have zero or positive bias. An illustration of the coupling is given in Figure 2.

It seems rather reasonable to expect that $(X_n)_n$ tends to visit more sites than $(Z_n)_n$, since the occasional biased steps experienced by $(X_n)_n$ should help it spread its trajectory into space. Unfortunately, it is not clear how to make this into a rigorous argument. For instance, it is not true that, if the first visit of site Z_n by the coupled simple random walk occurs at time n, then the first visit of site X_n by the excited random walk occurs at time n. However, the corresponding property is true if one restricts attention to the so-called *tan points*, that we now define.

Say that m is a tan point for $(Z_n)_n$ if

$$Z_m \cdot e_1 > Z_k \cdot e_1$$

for all $0 \leq k \leq m-1$ such that

$$Z_k \cdot e_2 = Z_m \cdot e_2.$$

An illustration of the definition is given in Fig. 3. One similarly defines the notion of tan point for $(X_n)_n$. The terminology is explained by the following picture: imagine the Sun being placed infinitely far in the $+e_1$ direction. Then, if m is a tan point, the past history of the random walk up to time m

[2]This is true up to a non-problematic time-change.

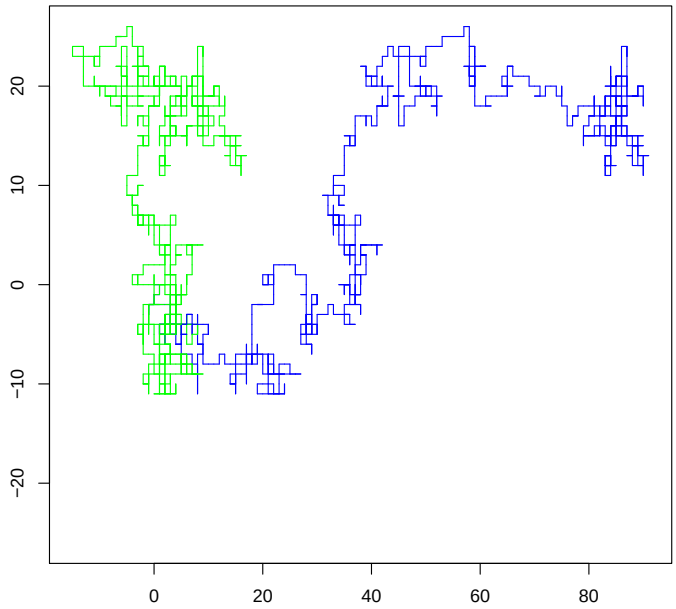


FIGURE 2. Joint realization of the first 1000 steps of an excited random walk with $p = 0.4$ (blue) and the coupled simple random walk (green).

does not shield the point Z_m from sunbeams parallel to the e_1 axis. Since by construction the sequence

$$(X_n \cdot e_1 - Z_n \cdot e_1)_{n \geq 0}$$

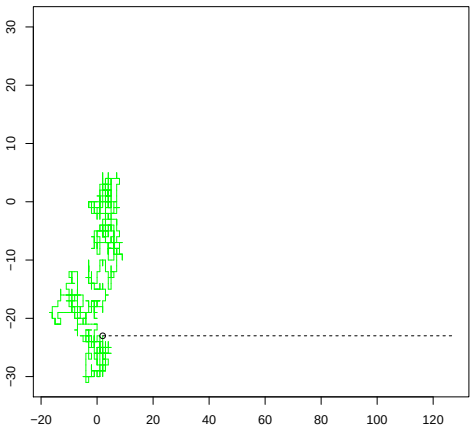
is non-decreasing, it is easily checked that a tan point for (Z_n) must also be a tan point for (X_n), see Fig. 4. Calling N_n the number of tan points of $(Z_k)_k$ that appear before time $\leq n$, one thus has that

$$R_n \geq N_n.$$

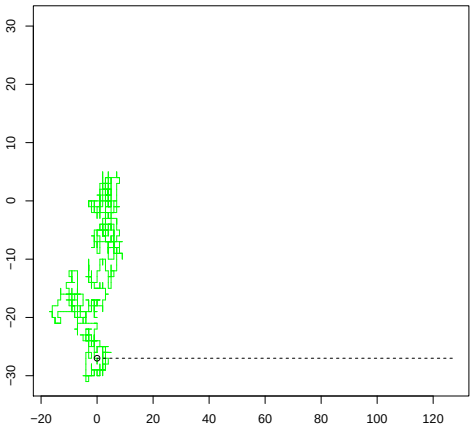
To obtain a lower bound on the order of magnitude of N_n for large n, Benjamini and Wilson exploited a combinatorial result by Bousquet-Mélou and Schaeffer [**40**]. Their conclusion is that the number of tan points of a simple random walk on $\mathbb{Z}^2$ between its first entrance to, and its first exit from, a horizontal strip of height r, has a probability of being of order $r^{4/3}$ that is bounded from below when r is large. Since typically $Z_n \cdot e_2 \propto n^{1/2}$, the number of disjoint such strips crossed by the walk up to time n is of order $n^{1/2}$, whence an overall number of tan points of order

$$\left(n^{1/2}\right)^{4/3} = n^{2/3} >> n^{1/2}.$$

As in the $d = 3$ case, this is enough to prove transience in the e_1 direction, but not ballisticity.



(a) The end-point is a tan-point.



(b) The end-point is not a tan-point

FIGURE 3. Two pieces of trajectories of a simple random walk illustrating the notion of a tan point. The end-point of each trajectory is circled.

2.2. Transience for the general positive cookies walk (Zerner [157]). In [**157**], Zerner developed a nice approach to the proof of the transience of the positive cookies excited random walk when $\delta \cdot u > 0$, quoted as Theorem 20. Unlike the approach of Benjamini and Wilson [**19**], which involves precise estimates and an explicit connection to the simple random walk, the approach of [**157**] is exclusively based on soft arguments combining martingale techniques and properties of the environment viewed from the particle. Here is an admittedly very rough sketch of the corresponding

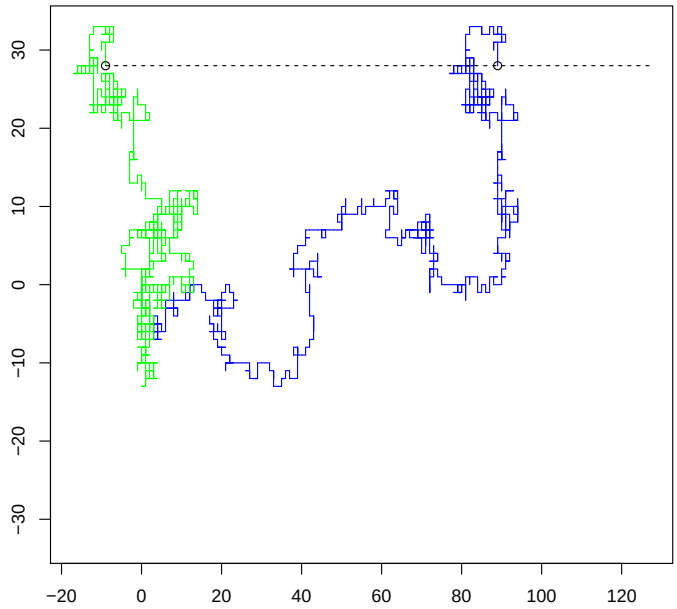


FIGURE 4. Joint realization of an excited random walk with $p = 0.4$ (blue) and the coupled simple random walk (green), where the end-point of the simple random walk is a tan point. The end-points of both trajectories are circled.

argument. Introduce the drift accumulated by the walk up to time n, i.e.

$$D_n := \sum_{k=0}^{n-1} \sum_{e \in \mathcal{E}} \omega(X_k, e, L_k(X_k))e, \tag{69}$$

and introduce the (vector-valued) martingale $(M_n)_{n \geq 0}$ defined by

$$M_n := X_n - D_n.$$

A key object in the proof is the (scalar) martingale $(M_n \cdot u)_{n \geq 0}$, which is especially useful here since, due to the assumption of $u-$positivity of the cookies, one always has that $D_n \cdot u$ is non-negative. Introducing

$$T_x := \inf\{n \geq 0;\ X_n \cdot u \geq x\},$$

one can show with the help of this martingale that $\mathbb{P}(T_x < +\infty) = 1$ for all $x \geq 0$.

One then "stationarizes" the problem by showing, using only minimal regularity properties of the model, that a stationary distribution exists for the Markov chain describing the cookie environment viewed from the particle at the successive times $T_1, T_2, \ldots$, with the additional property $(\mathscr{P})$ that, starting from this distribution, the cookie environment viewed from the particle at time T_1 restricted to the half space $\{x \in \mathbb{Z}^d;\ x \cdot u \geq 0\}$ has the same

distribution than when starting with the original distribution of the cookie environment.

Using again the $u-$positivity of cookies in combination with the martingale $(M_n)_n$ and stationarity, one shows that $\tilde{\mathbb{P}}(A_u) = 1$, where $A_u := \{\lim_{n\to+\infty} X_n \cdot u = +\infty\}$ and $\tilde{\mathbb{P}}$ is obtained by using the stationary cookie environment. An elementary argument then implies that $\tilde{\mathbb{P}}(A_u \cap B_u) > 0$, where $B_u := \{\forall n \geq 1\ X_n \cdot u > 0\}$. Thanks to property $(\mathscr{P})$, $\mathbb{P}(A_u \cap B_u) = \tilde{P}(A_u \cap B_u)$, so we have that $\tilde{\mathbb{P}}(A_u \cap B_u) > 0$. Finally, a classical argument shows that $\mathbb{P}(A_u)$ can only be equal to 0 or 1, whence the conclusion that $\mathbb{P}(A_u) = 1$.

2.3. The renewal structure. We now give the precise definition of the renewal structure used for the excited random walk. Note that the definition is the exact counterpart of the one used for multi-dimensional random walks in random environment, see **[144]** (which itself is a generalization of the one appearing in the one-dimensional case **[97, 96]**), and here, as opposed to the case discussed in Chapter 2 for interacting particle systems, finding a definition with the right structural properties is not difficult. We give the definition in the general case of a vector u for which $\delta \cdot u > 0$. For the original excited random walk model, one takes $u := e_1$.

We say that $m \geq 1$ is a *renewal time* for the walk if the following condition is satisfied

$$\sup_{0\leq k\leq m-1} X_k \cdot u < X_m \cdot u \leq \inf_{k\geq m+1} X_k \cdot u, \tag{70}$$

and define the sequence $(\kappa_n)_{n\geq 0}$ by $\kappa_0 := 0$ and

$$\kappa_{n+1} := \inf\{m > \kappa_n;\ m \text{ is a renewal time}\}. \tag{71}$$

The almost sure finiteness of the κ_ns is a direct consequence (see e.g. **[157]**) of transience in direction u, and one then has that

(i) the r.v.s $(\kappa_{n+1} - \kappa_n, r_{\kappa_{n+1}} - r_{\kappa_n})_{n\geq 0}$ are independent,
(ii) the r.v.s $(\kappa_{n+1} - \kappa_n, r_{\kappa_{n+1}} - r_{\kappa_n})_{n\geq 1}$ are identically distributed,

To prove tail estimates on κ, one introduces the following sequence of stopping times, starting with $D_0 := 0$. For $n \geq 1$,

$$S_n := \inf\{m > D_{n-1};\ X_m \cdot u > \sup_{0\leq k\leq m-1} X_k \cdot u\}, \tag{72}$$

$$D_n := \inf\{m > S_n;\ X_m \cdot u < X_{S_n} \cdot u\}. \tag{73}$$

One then lets

$$K := \inf\{n \geq 1;\ D_n = +\infty\},$$

and observe that $\kappa_1 = S_K$ (the really important point is that S_K is a renewal time so that $\kappa_1 \leq S_K$).

These definitions are illustrated in Fig. 5 and 6.

2.4. Estimates on the renewal structure I: original excited random walk model (B., Ramírez [34]). We now describe the argument developed in **[34]** to prove tail estimates on the renewal structure, leading to the following bound, which is (more than) sufficient to establish finiteness of

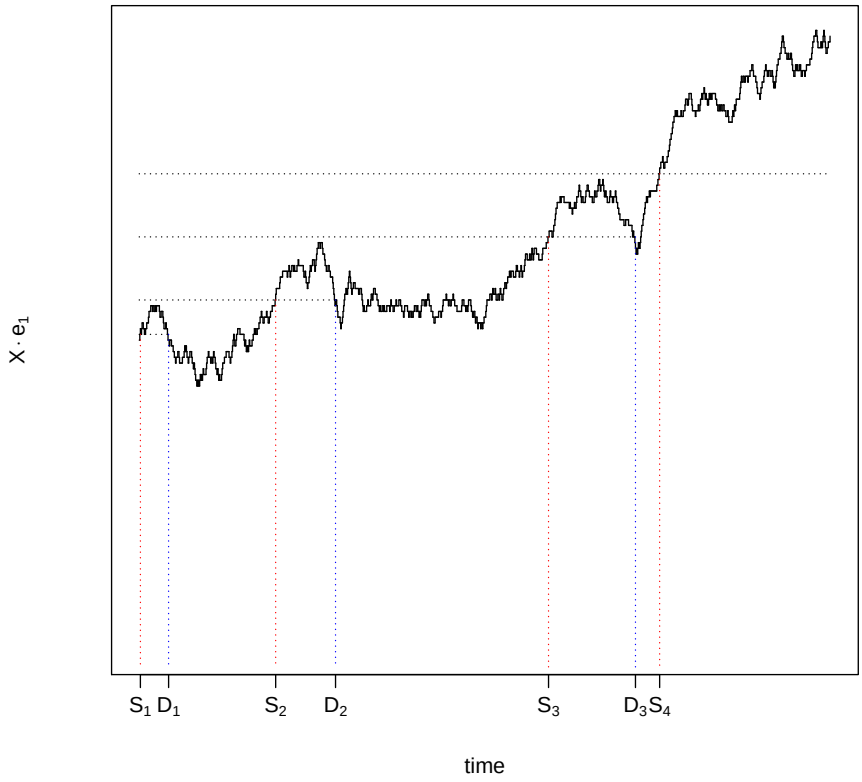


FIGURE 5. Position on the e_1 axis vs. time plot of an excited random walk on $\mathbb{Z}^2$ with $p = 0.1$. The sequence of stopping times $S_1, D_1, \ldots$ is shown on the horizontal axis, while the successive values of $X_{S_i} \cdot e_1$ are shown on the vertical axis.

the second moment and thus derive the law of large numbers and the central limit theorem (Theorem 22):

$$\mathbb{P}(\kappa_1 \geq t) = e^{-t^{-1/19+o(1)}}. \tag{74}$$

The key idea is to exploit the directional super-diffusive lower bounds of the type obtained[3] in **[19]**, which show that $X_n \cdot e_1 \gtrsim cn^a$, for some $a > 1/2$ and $c > 0$. Specifically, a general argument shows that, as soon as one has that, for some $\psi > 0$,

$$\mathbb{P}(X_n \cdot e_1 \leq n^a) \leq \exp\left(-n^{\psi+o(1)}\right), \tag{75}$$

which is the form under which these bounds are proved and used in **[34]**, a bound similar to (74) follows.

[3]We have seen that such bounds can be obtained by counting the number of distinct sites visited by the projected walk when $d \geq 3$, or by using estimates on the number of tan points when $d = 2$.

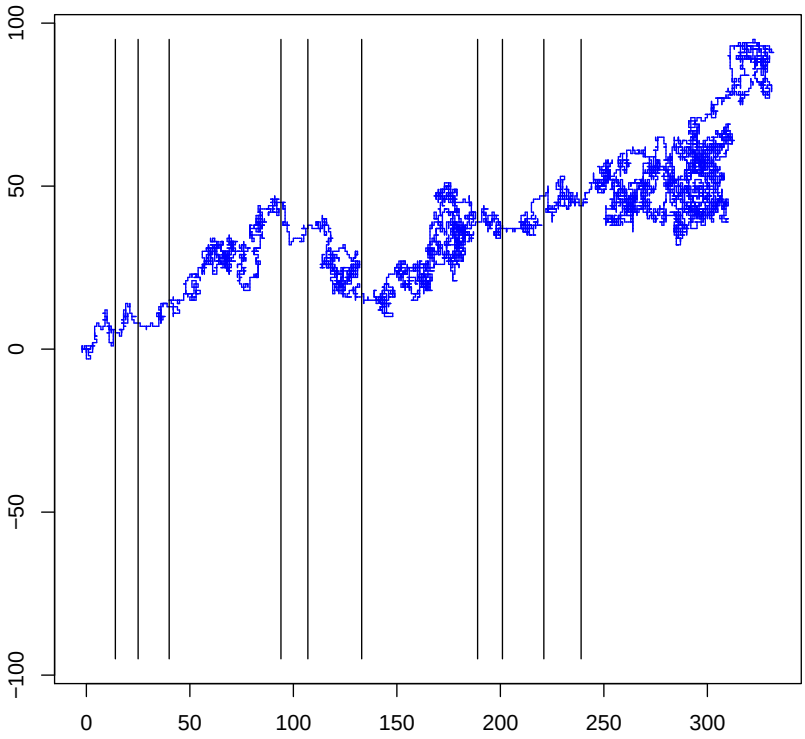


FIGURE 6. Trajectory of an excited random walk on $\mathbb{Z}^2$. Some (but not all, for the sake of readability) of the renewal times κ_i are shown by means of vertical straight lines with e_1 coordinate $X_{\kappa_i} \cdot e_1$.

Here is a description of the three ingredients used in the proof, in addition to the super-diffusive lower bound (75).

The first ingredient is a bound on $X_{S_n} \cdot e_1$ in terms of the time intervals $D_k - S_k$ for $1 \leq k \leq n$. Start with the following decomposition:

$$X_{S_n} \cdot e_1 = X_{S_1} \cdot e_1 + \sum_{k=1}^{n-1} (X_{S_{k+1}} \cdot e_1 - X_{S_k} \cdot e_1) \tag{76}$$

Then, by definition of the random variables S_n and D_n (see Fig. 6), we have that $X_{S_1} \cdot e_1 = 1$ and that

$$X_{S_{k+1}} \cdot e_1 = \max_{S_k \leq i \leq D_k} X_i \cdot e_1 + 1 \leq X_{S_k} \cdot e_1 + (D_k - S_k + 1), \tag{77}$$

where the inequality in (77) is due to the fact that the walk has nearest-neighbour steps. Plugging this inequality in (76), one has that

$$X_{S_n} \cdot e_1 \leq 1 + \sum_{k=1}^{n-1} (D_k - S_k + 1). \tag{78}$$

The second ingredient is a bound on the tail of the random variables $D_k - S_k$ appearing in (78). By definition of S_k and D_k, the cookie environment seen by the walk between time S_k and time $D_k - 1$ consists of exactly one cookie per site, so that, conditional on S_k being finite, and on the past of the walk up to time S_k, the distribution of $D_k - S_k$ is exactly the distribution of

$$D := \inf\{k \geq 0;\ X_k \cdot e_1 = -1\}$$

with respect to $\mathbb{P}$. We can then use the directional super-diffusive lower bound (75) to control the tail of D, since $\mathbb{P}(D = k) \leq \mathbb{P}(X_k \cdot e_1 = -1)$, leading to the bound

$$\mathbb{P}(k \leq D < +\infty) \leq \sum_{m=k}^{+\infty} \mathbb{P}(X_m \cdot e_1 = -1) \leq \exp\left(-k^{\psi+o(1)}\right). \tag{79}$$

The third ingredient just consists in the observation that, since the successive finite values of S_n are as many attempts at producing a renewal time, each of which has a conditional probability of success given the past equal to $\mathbb{P}(\forall k \geq 0\ X_k \cdot e_1 \geq 0) > 0$, the tail of K is geometric.

We now want to bound the tail of κ_1. To this end, write $t = vw$, for a pair of integers $v, w \geq 1$, and observe that, for κ_1 to be larger than t, at least one of the following three events must happen

$$A := \{X_t \cdot e_1 \leq vw\},\ B := \{K > v\},\ C := \bigcup_{i=1}^{v} \{w < D_i - S_i + 1 < +\infty\}.$$

Indeed, if, for instance, neither B or C holds, (78) shows that $X_{\kappa_1} \cdot e_1 = X_{S_K} \cdot e_1$ must be $\leq vw$, so that κ_1 cannot be larger than t unless $X_t \cdot e_1 \leq vw$, i.e. A, holds. Then the individual probabilities of the events A, B, C can be controlled, respectively with another use of the directional super-diffusive lower bound (75) for A, the geometric bound on the tail of K for B, and the bound (79) on the tail of $D_k - S_k$ (with a basic union bound) for C. Choosing $v \sim t^\alpha$ and $w \sim t^\beta$ with $\alpha, \beta > 0$ such that $\alpha + \beta < a$ (where a is the exponent in (79)), one obtains a bound of the form

$$\mathbb{P}(\kappa_1 \geq t) \leq \exp(-n^{\psi'}), \tag{80}$$

with $\psi' > 0$. The precise value of ψ' obtained in **[34]** after a little optimization over the various bounds is $\psi' = 1/19$, but any bound such as (75) with a positive value of ψ would lead to a tail bound like (80) with a positive ψ'.

In our opinion, one of the nice features of the above argument is that it allows one to lift a directional super-diffusive but sub-ballistic lower bound into a proof of ballistic behaviour (and more), thanks to the exploitation of the space-homogeneity of the model, which is reflected by the existence of the renewal structure. In fact, few special properties of the excited random walk model are used in the above argument: beyond the directional super-diffusive bound (75), all that is needed[4] is that the initial environment is space-homogeneous and that the interaction between the walk and its environment is local (which here corresponds to the fact that the cookie

[4]It seems that we also need the fact that $\mathbb{P}(\forall k \geq 0\ X_k \cdot e_1 \geq 0) > 0$, but this also can be obtained as a consequence of (75).

environment at a site is modified only when the walk hits this site), along with bounded steps. A (shamelessly vague) formulation could be stated[5] as:

directional super-diffusivity+local interaction+homogeneous environment

$$\Big\Downarrow \text{ renewal}$$

ballisticity, CLT, etc.

2.5. Estimates on the renewal structure II: general multidimensional model (Menshikov, Popov, Ramírez, Vachkovskaia [120]). To apply the argument described in the previous subsection, a directional lower bound comparable to (75) is needed, and the methods used to prove such estimates for the original excited random walk seem difficult to adapt to more general models. In **[120]**, a completely different approach was developed, leading to super-diffusive lower bounds in a much more general setting.

The key result proved in **[120]** is the following (Proposition 4.1 in **[120]** is actually more general, but we quote only the version corresponding to the context discussed here). For the excited random walk with one $u-$positive cookie per site in dimension $d \geq 2$, under the uniform ellipticity assumption (65), there exists $b > 1/2$ and $\phi > 0$ such that

$$\mathbb{P}(R_n \leq n^b) \leq \exp\left(-n^{\phi+o(1)}\right), \tag{81}$$

where R_n denotes the number of distinct points visited by the walk up to time n. This is enough to derive a result analogous to (75) once uniform strict $u-$positivity of the cookies is added, which in turn implies a tail bound on the renewal times thanks to the argument described in the previous subsection. Note that it is not necessary to assume $u-$strictly positive cookies for (81) to hold.

At the heart of the proof of (81) is a combination of martingale arguments, whose main steps we now briefly sketch. Remember the definitions of the accumulated drift D_n and martingale M_n from (69) and (70):

$$D_n := \sum_{k=0}^{n-1} \sum_{e \in \mathcal{E}} \omega(X_k, e, L_k(X_k))e, \;\; M_n := X_n - D_n,$$

and the fact that, from $u-$positivity, one always has that $D_n \cdot u \geq 0$.

A first estimate controls the tail of the time spent by the walk in a given strip of the form $S_m := \{x \in \mathbb{Z}^d; \; m \leq x \cdot u < m+1\}$ up to time n. Broadly speaking, each time the walk lies in S_m, ellipticity combined with a gambler's ruin type estimate obtained via M_n, yields a lower bound on the probability that the walk will hit a strip S_q with $q >> m$ before hitting S_x again. An upper bound on the probability for the walk to go back from S_q to S_m before time m is then provided by applying Azuma's inequality for martingales with bounded increments.

On the other hand, if one restricts attention to time intervals $[n_1, n_2]$ during which the walk does not visit any new site, $(X_n)_n$ itself behaves as a martingale, and an argument similar in spirit to the one above, using

[5] In fact, it is not really necessary to have $a > 1/2$ in a bound of the type of (75), any $a > 0$ would do.

$||X_n - X_{n_1}||^b$ for a suitably chosen $b < 1$ instead of $M_n \cdot u$, allows one to show that, if $n_2 - n_1$ is large, any given set of points which covers a sufficiently large fraction of a large ball centered at X_{n_1}, will be visited with probability close to 1.

The argument then goes as follows. Divide $\mathbb{Z}^d$ into disjoint strips of the form $H_j := \{x \in \mathbb{Z}^d;\ (j-1)h \leq x \cdot u < jh\}$. Strips into which the walk visits more than r sites up to time n are called traps. When there are many traps, the number of visited sites is large, which is precisely what one wants to prove, so one may restrict attention to situations where the number of traps is small. Then the first estimate shows that the total time spent by the walk in traps cannot be too large. On the other hand, the second estimate shows that, whenever the walk is not in a trap, it must soon hit a previously unvisited site, since the walk is surrounded by a large proportion of them.

3. Perspectives

As mentioned above, we only have a limited understanding of multi-dimensional excited random walks where cookies can produce both positive and negative bias in a given direction. Finding reasonably general criteria for recurrence/transience or ballisticity is thus a challenging open problem.

More generally, one may combine the idea of an excited random walk with more general processes than classical random walks, such as random walks in a random potential, or persistent random walks.

One more specific interesting open question about excited random walks in dimension $d \geq 2$ is that of large deviations, even in the context of the original model (see **[129]** for a treatment of the large deviations in the case $d = 1$). A general large deviations argument due to Rassoul-Agha **[135]**, extending previous work by Varadhan **[149]** on random walks in random environment, shows that a large deviations principle indeed holds for X_n, with a convex lower semi-continuous rate function. However, this result is not explicit, and, in particular, it does not allow one to determine the zero set of the rate function.

The following elementary argument shows that slowdown probabilities for the walk are actually on a subexponential scale. Consider the event that $|X_k| \leq \lambda$ for all $0 \leq k \leq n$. If the walk were an ordinary simple random walk, the corresponding probability would be (roughly) of order $e^{-n\lambda^{-1/2}}$. Now one can ask the cookies in $[-\lambda, \lambda]^d$ to have no effect on the walk. To make this idea precise, assume that, when the random walk hits site x for the first time and chooses to move in the e_1 direction, the corresponding step Δ_x is specified by a uniform random variable U_x according to

$$\Delta_x = 2\mathbf{1}(U_x < (1+p)/2).$$

Conditioned upon the event that $U_x > p$ for all $x \in [-\lambda, \lambda]^d$, the random variables Δ_x are symmetric, so the evolution of the walk is identical to that of a simple random walk as long as it remains in $[-\lambda, \lambda]^d$. We deduce that the probability of having $X_k \in [-\lambda, \lambda]^d$ for all $0 \leq k \leq n$ is at least of order $e^{-\lambda^d - n\lambda^{-1/2}}$. Optimizing over λ, we see that we can achieve a probability of order $e^{-n^{\frac{d}{d+1/2}}}$. Finding the exact order of magnitude of the slowdown large deviations is an open question.

As far as speedup large deviations are concerned, one might expect that the cost is always non-zero on an exponential scale (it is trivially so for speed values larger than p/d by comparison with a simple random walk with bias p in the e_1) direction. If it were possible to couple two versions of the excited random walk in a non-decreasing way with respect to the cookie environment[6], a rather direct argument based on sub-additivity would do the job. Unfortunately, except in dimension 1, where it is e.g. a consequence of the branching process representation, it is not clear that such a coupling exists, even though it seems intuitively clear that putting more cookies in the environment tends to push X_n further in the e_1 direction.

[6]More precisely, starting from two cookie environments ω_1 and ω_2 such that, at each site, ω_2 has a cookie whenever ω_1 has one, we would like to find a coupling between two versions $(X_n^{(1)})_n$ and $(X_n^{(2)})_n$ of the excited random walk, with respective initial cookie environments ω_1 and ω_2, such that $X_n^{(1)} \cdot e_1 \leq X_n^{(2)} \cdot e_1$ almost surely.

CHAPTER 4

Branching-Selection dynamics

Natural selection is believed to be one of the fundamental processes shaping the evolution of life on Earth. It is an extremely complex process, due to the interplay of a very large number of factors over a huge variety of scales of time and space. The mathematical modeling of selection processes started with the work of the founding fathers of population genetics, such as Fisher [**74**], Haldane [**87**] and Wright [**154**], and is still an active field of research (see e.g. [**140**]). Most often, models focus on a few specific aspects, making rather drastic simplifying assumptions on the other features of the process, and the models we discuss here are no exception.

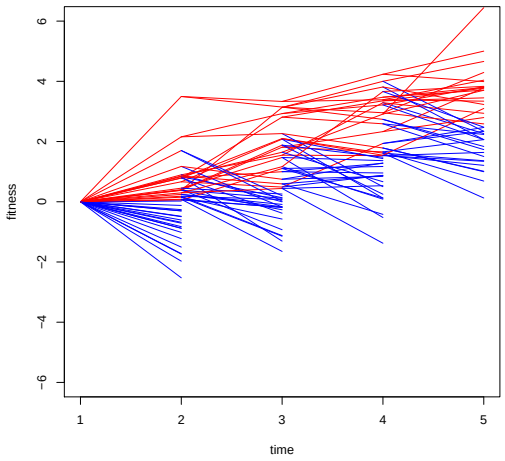
1. Model(s)

In the models we consider, the main focus is the joint effect of the population size and the distribution of mutations, on the speed of evolution. One first assumption we make is that of a *constant population size* N. In other words, the number of individuals in the population under study is kept to a constant value N over the generations. Another assumption is that individuals can be described by a *single numerical fitness value*, regardless of the complex type differences that may exist between them. Finally, we assume an *asexual* population, in which single individuals give birth to children, and we make the assumption that the effect of mutations simply consists in shifting the fitness value of a child from the fitness value of its parent by a random amount. To simplify the terminology, we speak of the location of an individual on the real line to refer to the value of its fitness.

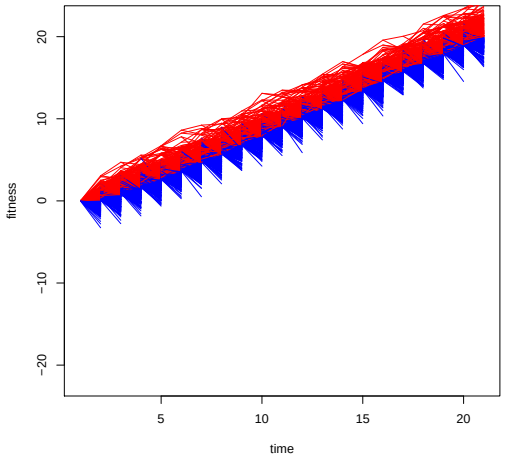
Various approaches can be used to model the gradual replacement of individuals by their children, and the effect of selection. The first model we consider is the discrete-time $N-$branching random walk, abbreviated *discrete time* $N-BRW$ in the sequel. At each discrete time $n = 0, 1, \ldots$, we have a population of N particles with fitness-values in $\mathbb{R}$, representing the $n-$th generation. The population of particles evolves through the repeated application of branching and selection steps defined as follows (see Fig. 1):

- Branching: each of the N particles is replaced by two new children particles, whose positions are shifted from that of the original particle by independently performing two random walk steps, according to a prescribed distribution μ;
- Selection: only the N right-most particles, i.e. those with the N highest fitness values, are kept among the $2N$ particles obtained at the branching step, to form the new population.

An alternative model is the *continuous-time* $N-BRW$, that describes a population of N particles in which the replacement of individuals by their



(a) $N = 20$



(b) $N = 200$

FIGURE 1. Realization of the discrete-time $N-$BRW model with $\mu = \mathcal{N}(0, 1)$. Red (resp. blue) lines connect parents with selected (resp. eliminated) children.

children occurs in continuous – rather than discrete – time. At rate 1, each particle produces a new particle, whose position is shifted from its parent by performing one random walk step according to the distribution μ. Immediately after such a branching event, only the N right-most particles are kept among the $N + 1$ particles present in the population.

Yet another model is the $N-$branching Brownian motion, abbreviated in the sequel as the $N-BBM$. As in the $N-$BRW, each particle branches at rate 1, but particles positions evolve in continuous-time according to independent Brownian motions.

Note that one may imagine plenty of other ways of introducing selection in the model. The selection mechanism used above has the advantage (as far as the theoretical analysis of the models is concerned) of not introducing additional randomness beyond that due to mutations. In other, less unrealistic models, selection would be modeled by the fact that an individual produces a random number of children, whose expected value increases in proportion of the individual's fitness. For instance, in a population described by a list of N (positive) fitness values $(x_1, \dots, x_N)$, the distribution of the numbers of children of individuals $1, \dots, N$, denoted $(n_1, \dots, n_N)$, would be multinomial with parameters N and $(p_1, \dots, p_N)$, where

$$p_i := \frac{x_i}{\sum_{j=1}^N x_j}.$$

2. Results

Our primary motivation for studying branching-selection dynamics is that they provide (extremely simplified) models of the process of natural selection. However, it turns out that the specific models we consider are also related to a more general theory of stochastic fronts, developed by theoretical physicists Brunet and Derrida, with applications in other fields (e.g. stochastic models of polymers). Before stating precise mathematical results, we give a brief description of the main predictions of this theory.

2.1. Brunet-Derrida theory. The F-KPP equation, named after Fisher **[75]** and Kolmogorov, Petrovsky and Piscounov **[105]**, is one of the classical PDE models of front propagation, whose salient feature is that is leads to traveling wave solutions. In its simplest form, the equation reads

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} + u(1-u),$$

where $u = u(x,t)$, $x \in \mathbb{R}$, $t \geq 0$.

The equation possesses one stable equilibrium $u \equiv 1$ and one unstable equilibrium $u \equiv 0$. Typically, one considers an initial condition satisfying $\lim_{x\to-\infty} u(x,0) = 1$ and $\lim_{x\to+\infty} u(x,0) = 0$, and, under generic assumptions, the corresponding solution u converges, for large times, towards a traveling wave solution $\tilde{u}$ describing the invasion of the 0 phase by the 1 phase, at a constant velocity. More precisely, $\tilde{u}$ is of the form

$$\tilde{u}(x,t) = g(x - vt),$$

where $v \in \mathbb{R}$ is the wave speed, while $g(x),\ x \in \mathbb{R}$ describes the wave shape, and satisfies $\lim_{x\to-\infty} g(x) = 1$ and $\lim_{x\to+\infty} g(x) = 0$. Such traveling wave solutions exist for every v above a critical speed v_*. For *localized initial conditions*, i.e. $u(x,0) = 1$ for all $x \leq a$ and $u(x,0) = 0$ for all $x \geq b$, u is attracted towards the traveling wave solution associated with the minimal speed value v_*.

The results of Brunet and Derrida deal with systems described by F-KPP like equations perturbed by small stochastic noise terms. One example is the stochastic F-KPP equation

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} + u(1-u) + \sqrt{\epsilon u(1-u)}\dot{W}, \tag{82}$$

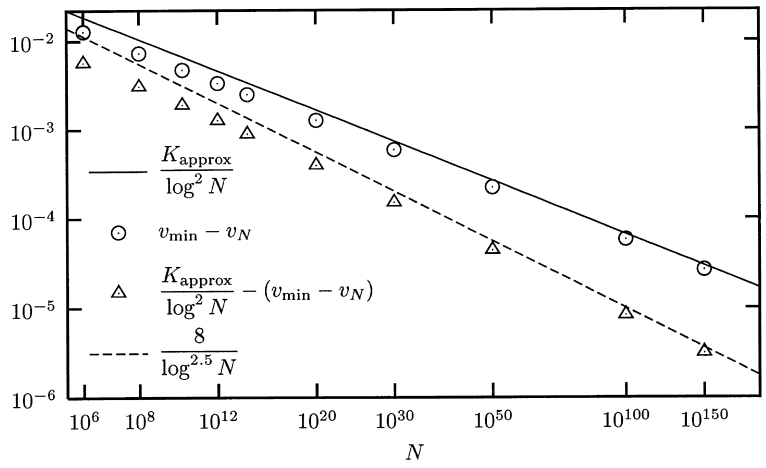


FIGURE 2. Numerical simulations from Brunet and Derrida [**46**]. Circles show the estimated values of $v_* - v_N$ for values of N ranging from 10^6 to 10^{150}, while the solid line shows the value of $\frac{C}{(\log N)^2}$. (Here v_* is $v_{\min}$ and C is K_{approx}.)

where $\dot{W}$ is a standard space-time white-noise.

For this model, Brunet and Derrida predicted that, to the first order, the effect of the noise term is to shift the velocity of the limiting traveling wave solution by an amount which, in the limit where $\epsilon \to 0$, goes to zero at an extremely slow rate. Specifically, starting with a localized initial condition, the solutions of (82) converge to traveling waves with speed v_ϵ, with

$$v_* - v_\epsilon \sim \frac{\pi^2}{(\log \epsilon)^2}. \tag{83}$$

For $N-$branching-selection processes, a prediction similar to (83) holds:

$$v_* - v_N \sim \frac{C}{(\log N)^2}, \tag{84}$$

where v_N is the large-time asymptotic velocity of the particle system with N particles, and v_* is the $N \to +\infty$ limit of the speed, which coincides with the maximum speed of the corresponding branching model (BRW or BBM) without selection.

To explain the connection of $N-$branching selection processes with the F-KPP equation, denote by $\mathfrak{X}_N(t)$ the population of particles in the process at time t, and, for $x \in \mathbb{R}$, define

$$F_N(x,t) := \frac{\text{number of particles in } \mathfrak{X}_N(t) \text{ whose position is } > x}{N},$$

and

$$F(x,t) := \lim_{N \to +\infty} F_N(x,t).$$

For the $N-$BRW in continuous time, Durrett and Remenik rigorously proved, see [**69**], that, under generic regularity assumptions, $F(x,t)$ is well-defined and satisfies the following equation

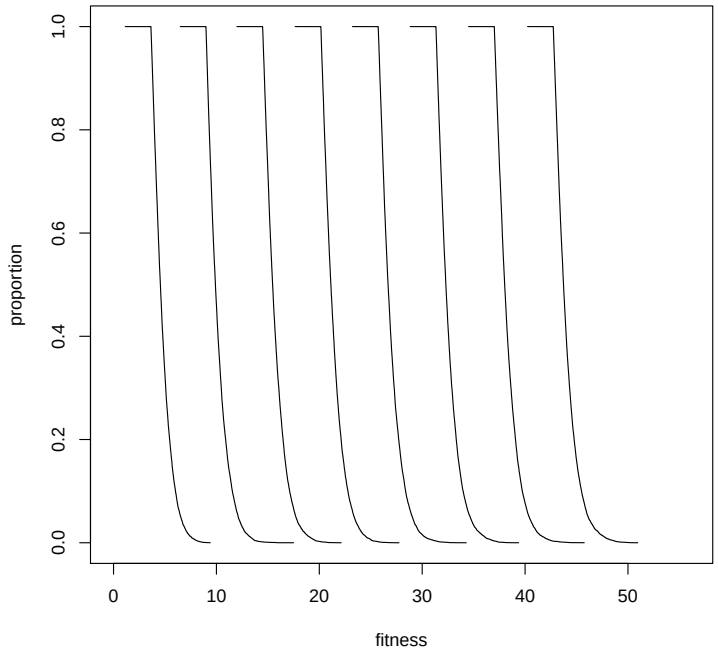


FIGURE 3. The successive graphs of $x \mapsto F_N(x,t)$ for a realization of the discrete-time $N-$BRW with $\mu = \mathcal{N}(0,1)$, $N = 10000$, and $t = 5, 10, 15, \ldots, 40$. The traveling wave behaviour is clearly visible.

$$\begin{cases} \dfrac{\partial F}{\partial t} = \left(F(\cdot,t) \star \mu\right)(x), \ x > \gamma(t), \\ F(x,t) = 1, \ x \leq \gamma(t). \end{cases} \tag{85}$$

where γ is a continuous increasing function[1]. To emphasize the analogy with the F-KPP equation, one may rewrite the r.h.s. of (85) as

$$\left(F(\cdot,t) \star \mu\right)(x) = \underbrace{\left(F(\cdot,t) \star \mu\right)(x) - F(x,t)}_{\rightsquigarrow \frac{\partial^2 F}{\partial x^2}} + F(x,t).$$

Thus, (85) means that, above $\gamma(t)$, F satisfies a linear equation similar to the F-KPP equation without the saturation term $(1-u)$, while a saturation mechanism is added below $\gamma(t)$ to ensure that F does not exceed 1. Similar equations can be written for the $N-$BRW in discrete time, or the $N-$BBM, although the corresponding results have not been formally proved. Fig. 3 illustrates the corresponding traveling-wave behaviour in the $N-$BRW case.

[1]Note that γ is not specified *a priori*, so that finding γ is part of solving the equation. In this sense, (85) constitutes a free boundary problem.

Note that F describes the model in the infinite-population limit $N \to +\infty$. The dynamics of the actual finite-population particle system $\mathfrak{X}_N$ is described by F_N, which only satisfies an approximation of (85), where stochastic fluctuations have to be added to account for the fact that the population size N is large but finite. For instance, F_N cannot take any intermediate value between 0 and $1/N$, so that F_N cannot be described by (85) with a higher "resolution" than $1/N$. Remarkably, Brunet and Derrida could produce an argument successfully predicting (83) or (84) based on barely more that this rather qualitative remark.

A very rough sketch of their argument is as follows. For solutions u of the stochastic F-KPP equation perturbed by stochastic noise such as (82), at every time $t > 0$, $x \mapsto u(x,t)$ continuously connects 1 at $x = -\infty$ to 0 at some random $x = r(t)$ defining the position of the front, right of which $u(\cdot, t)$ is identically zero. Looking at the equation (82), one can see that stochastic effects due to the noise term counterbalance the $u(1-u)$ creation term when u is of order ϵ. To find the asymptotic speed of propagation of the front, one should thus look for traveling waves obeying the F-KPP equation at the left of the front, taking values of order ϵ near the front, and which are identically equal to zero at the right of the front. To study these traveling waves, one replaces the F-KPP equation by a linear approximation, for which explicit solutions can be found – these solutions should be approximately valid for the original equation, thanks to the fact that the values of u are small near the front. One can then check that the speed of these traveling waves must satisfy (83). The same kind of argument is used to derive (83) for the branching-selection particle system, with ϵ replaced by $1/N$ – at a heuristic level, the only relevant properties are that the equation is similar to the linearized F-KPP equation for small values of u, has a saturation mechanism preventing the occurrence of large values of u, and that fluctuations lead to a cut-off for values of u of order ϵ. This heuristic picture describing the first-order correction to the limiting velocity was developed (and also compared with numerical simulations, see Fig. 2) in **[44, 45, 46]**.

Refining this approach, Brunet and Derrida (also with Mueller and Munier) could give a much more detailed description of the behaviour of the corresponding stochastic models, see **[42, 41, 43]**. For $N-$branching-selection processes, their finding is that the relevant time-scale to study the process is $(\log N)^3$. Broadly speaking, the heuristic picture with $1/N$ cut-off used to study the first-order correction to the velocity, describes a meta-stable state in which the process spends most of its time. However, on the $(\log N)^3$ time scale, perturbations due to the appearance of particles far to the right of the front appear, whose impact is to shift the position of the front by an amount of order $\log \log N$, before it returns to the next (suitably shifted) meta-stable state. This leads, among other things, to the second-order correction for the velocity shift:

$$v_* - v_N - \frac{C}{(\log N)^2} \sim -C' \frac{\log \log N}{(\log N)^3}, \tag{86}$$

but also to estimates on the asymptotic diffusion constant of the process, and to the conclusion that, on the $(\log N)^3$ time scale, the genealogy of the process is described by a Bolthausen-Sznitman coalescent, as opposed to

the classical Kingman coalescent appearing in population genetics models without selection.

2.2. First-order correction to the velocity shift. We now describe rigorous mathematical results that confirm Brunet and Derrida's predictions concerning the first-order velocity shift for various models.

The following result shows the correctness of Brunet and Derrida's conjecture for the stochastic F-KPP equation with noise (see also the earlier work **[54]**, where only partial results were obtained).

THEOREM 24 (Mueller, Mytnik, Quastel **[124, 123]**). *For the stochastic F-KPP equation with space-time white noise*

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} + u(1-u) + \sqrt{\epsilon u(1-u)}\dot{W},$$

starting with a localized initial condition, the asymptotic velocity of the front v_ϵ *satisfies*

$$v_* - v_\epsilon \sim \frac{\pi^2}{(\log \epsilon)^2}.$$

Note that Theorem 24 holds under more general assumptions (on the form of the equation and on the initial condition) that we do not quote here. Also, the conclusion can be strengthened by showing that

$$v_* - v_\epsilon - \frac{\pi^2}{(\log \epsilon)^2} = O\left(\frac{\log|\log \epsilon|}{|\log \epsilon|^3}\right), \tag{87}$$

which gives an upper bound on the same order of magnitude as the conjectured behaviour (86).

We now describe assumptions under which we have obtained the corresponding result for the $N-$BRW in discrete time.

We consider binary branching with i.i.d. random walk steps whose common distribution is denoted μ. Introduce

$$\Lambda(t) := \log \int \exp(tx) d\mu(x).$$

Here are the assumptions on μ:

(1) There exist $\sigma, \zeta > 0$ such that $\Lambda(t) < +\infty$ for all $t \in [-\sigma, \zeta]$.
(2) There exists $t^* \in]0, \zeta[$ such that $t^*\Lambda'(t^*) - \Lambda(t^*) = \log 2$.

Under these assumptions, it can be shown (see **[27]**) that, for all $N \geq 1$, there exists an asymptotic speed for the interacting particle system in the sense that, almost surely, one has that

$$\lim_{t\to+\infty} t^{-1} \max(\mathfrak{X}_N(t)) = \lim_{t\to+\infty} t^{-1} \min(\mathfrak{X}_N(t)) = v_N \in \mathbb{R}.$$

The main result on the velocity shift is that

THEOREM 25 (B., Gouéré **[27]**). *Under assumptions (1)-(2) above, one has that, as* $N \to +\infty$,

$$v_* - v_N \sim \frac{C}{(\log N)^2}, \tag{88}$$

where

$$C := \tfrac{\pi^2}{2} t^* \Lambda''(t^*), \ v_* := \Lambda'(t^*).$$

Note that, under assumptions (1)-(2), v_* is the asymptotic speed of the right-most position of a binary branching random walk with step distribution μ, i.e. almost surely

$$v_* = \lim_{t\to+\infty} t^{-1} \max \mathfrak{Y}(t),$$

where $\mathfrak{Y}(t)$ denotes the population of particles in the $t-$th generation of the branching random walk. In fact, there exists a close relation between the velocity shift described by Theorem 25 and the survival probability of the branching random walk killed below a linear space-time boundary, and estimates on this survival probability are crucial to the proof of Theorem 25.

We now describe this model in more detail. Consider a fixed speed $v \in \mathbb{R}$, and for each time t, kill every particle in $\mathfrak{Y}(t)$ whose location lies strictly below vt (here, killing means that not only the particle, but also all its descendants, are removed from the process). Then define $\rho(v)$ as the probability that the branching random walk survives, i.e. that, after killing, the process still contains particles for every time $t \geq 0$. When $v \geq v_*$, one has that $\rho(v) = 0$, while $\rho(v) > 0$ when $v < v_*$. The following theorem characterizes the speed at which $\rho(v)$ goes to zero when v approaches the critical speed v_*, starting from a single particle at the origin at time 0.

THEOREM 26 (Gantert, Hu, Shi [**80**]). *Assume (1)-(2). For $v < v^*$, one has the following asymptotic behavior as $v \nearrow v^*$:*

$$\log \rho(v) \sim -\pi \sqrt{\frac{\Lambda''(t^*)t^*}{2(v^* - v)}}. \tag{89}$$

Using a completely different approach, we obtained an alternative proof of Theorem 26, under more stringent assumptions on μ, and with a strengthening of the control upon the error term.

THEOREM 27 (B., Gouéré [**28**]). *Under the assumption that μ has bounded support and that assumption (2) above holds, one has that, as $v \nearrow v^*$,*

$$\log \rho(v) = -\pi \sqrt{\frac{\Lambda''(t^*)t^*}{2(v^* - v)}} + O(\log(v^* - v)).$$

Combined with the approach of [**27**], this strengthening leads to an improved bound for the velocity shift that matches (87), i.e.

$$v_* - v_N - \frac{C}{(\log N)^2} = O\left(\frac{\log\log N}{(\log N)^3}\right). \tag{90}$$

Finally, we mention [**65, 17, 16**], where a Brunet-Derrida velocity shift is proved for for F-KPP type equations with small deterministic cut-off.

2.3. Higher-order results. In this section, we quickly mention results connected with the higher-order description of branching selection systems (as opposed to the first-order correction results discussed above).

In [**37**], Berestycki, Berestycki and Schweinsberg studied the BBM with killing at the nearly critical speed

$$\gamma_N := \sqrt{2 - \frac{\pi^2}{a_N^2}},$$

where

$$a_N := \frac{1}{\sqrt{2}}(\log N + 3\log\log N).$$

Remarkably, they succeeded in giving, in the framework of this model, a rigorous content to the heuristic picture obtained by Brunet and Derrida, with perturbations of the metastable state of the system occurring on a $(\log N)^3$ time scale and at a distance of order $\log\log N$ of the front. In particular, this led to the proof that the genealogy of this model is described, on the $(\log N)^3$ time scale, by the Bolthausen-Sznitman coalescent.

In **[38]**, the same approach was used to obtained refined estimates on the survival probability of the BBM killed at a nearly critical speed, yielding precise higher-order asymptotics that refine the analog for BBM of Theorems 26 and 27.

Finally, very recently, Maillard (**[119]**) succeeded in adapting the approach of **[37]** to the $N-$BBM (by making use of killing at a suitably defined random barrier instead of a linear deterministic one), leading to a very precise description of the $N-$BBM on $(\log N)^3$ time scales that we now quote.

To precisely locate the population of particles in $\mathfrak{X}_N(t)$, one introduces the median position of the particles defined by

$$m_N(t) := \inf\{x \in \mathbb{R};\ \mathfrak{X}_N(t) \cap [x, +\infty[\text{ contains less than } N/2 \text{ particles}\}.$$

THEOREM 28 (Maillard, **[119]**). *Assume that the initial distribution consists of N particles drawn in an i.i.d. way according to the density defined (up to normalization) by $\sin(\pi x/a_N)e^{-\sqrt{2}x}\mathbf{1}_{[0,a_N]}(x)$. Then the finite dimensional distributions of the process*

$$(m_N((\log N)^3 t) - a_N(\log N)^3 t)_{t\geq 0}$$

converge, as $N \to +\infty$ to those of a Lévy process $(L_t + x_{1/2})_{t\geq 0}$, where $L_0 := 0$ and

$$\mathbb{E}(e^{i\lambda L_1}) = i\lambda c + 2\pi^2 \int_0^{+\infty} \left(e^{i\lambda x} - 1 - i\lambda x \mathbf{1}(x \leq 1)\right) d\Lambda(x),$$

where c is a constant and Λ is the image of the measure $x^{-2}\mathbf{1}(x > 0)dx$ by the map $x \mapsto \frac{1}{\sqrt{2}}\log(1+x)$.

2.4. Proportional selection scheme. Here, we give a very brief description of results on branching-selection models obtained in our PhD thesis. These results deal with discrete-time models using a proportional selection scheme. Specifically, from a population of N individuals at time t $\mathfrak{X}_N(t) := (X^1(t), \ldots, X^N(t))$, a new population of N individuals is created by letting each individual yield $n_i(t)$ copies of itself, where $(n_1(t), \ldots, n_N(t))$ follows a multinomial distribution with parameters N and $(p_1(t), \ldots, p_N(t))$, with

$$p_i(t) := \frac{X^i(t)}{\sum_{j=1}^N X^j(t)}.$$

Then the fitness of each individual in the new population is shifted by a random amount with distribution $\mu := \frac{1}{2}\delta_{-1} + \frac{1}{2}\delta_{+1}$. A truncation mechanism (e.g. reflection at zero) is added so as to keep fitness values positive.

Our main result is the following:

THEOREM 29 (B., Bienvenüe [23]). *For all $N \geq 1$, one has the following convergence in distribution on the Skorohod space:*

$$\left(\frac{\mathfrak{X}_N(\lfloor sT \rfloor)}{\sqrt{T}}\right)_{s\geq 0} \xrightarrow[T\to+\infty]{} (1,\dots,1)(Z_s)_{s\geq 0},$$

where $(Z_s)_{s\geq 0}$ is a Bessel process of dimension $2N-1$.

Additional results (e.g. a description of the asymptotic distribution of fitness differences within the population) are also available on this model, see **[23]**.

Note that, as opposed to the $N-$BRW case, the effect of mutations becomes smaller and smaller as the overall fitness of the population increases, explaining the $T^{1/2}$ scaling of the speed of evolution, as opposed to the linear speed observed in the $N-$BRW case.

An extension of Theorem 29 was obtained in **[22]**, to deal with the case of inhomogeneous mutations. There, an individual with fitness x leads to a child with fitness $x+\zeta x^\alpha$, where $0<\alpha<1$ and where ζ is a centered random variable with finite moments of all order. (A truncation mechanism again forces fitness values to remain positive). In this case, the proper normalization is T^β, where $\beta := \frac{1}{2(1-\alpha)}$, and the convergence is to the $2\beta-$th power of a Bessel process of dimension $(2N-3)(1-\alpha)^{-1}+2$. One interest of this extension is that, in the absence of selection, the model experiences a phase transition from positive recurrence when $\alpha<1/2$ to null-recurrence/transience when $\alpha \geq 1/2$.

For the sake of completeness, let us mention that, in **[26]**, we obtained results on a different kind of branching-selection dynamics in random environment, where the fitness landscape is either given by a supercritical Galton-Watson tree conditioned upon non-extinction, or a branching random walk on the binary tree. Although related to the branching-selection models described here, the perspective is rather different, and we do not elaborate on these results in this manuscript.

3. Proofs

3.1. Proof of Theorem 25. The main tool used in **[27]** to prove (88) is a comparison of the $N-$BRW with the BRW killed below a near-critical space-time boundary. Indeed, there is a natural coupling between the $N-$BRW and a system of N independent branching random walks: if we suppress the selection steps in the definition of the $N-$BRW and let every particle survive and branch at each step, we obtain a system of N independent classical BRWs. The idea is to compare the $N-$BRW with this system of BRWs killed below a space-time boundary moving at a speed w such that

$$\rho(w) \approx \frac{1}{N},$$

which, from (89), must satisfy

$$v^* - w \sim \frac{C}{(\log N)^2}.$$

Two distinct comparison arguments are used to prove that

$$v^* - v_N \lesssim \frac{C}{(\log N)^2} \tag{91}$$

and

$$v^* - v_N \gtrsim \frac{C}{(\log N)^2}. \tag{92}$$

Let us start with (91). Choosing a small $\phi_1 > 0$ and

$$w_1 := v^* - \frac{C + \phi_1}{(\log N)^2}, \tag{93}$$

we see that, by (89), one has

$$\rho(w_1) >> \frac{1}{N}.$$

As a consequence, if we use speed $w := w_1$ to kill the BRWs, there is a probability close to 1 that, among our system of N independent BRWs, at least one survives. Now consider

$$w_2 := v^* - \frac{C + \phi_2}{(\log N)^2}, \tag{94}$$

where $\phi_2 > \phi_1$, and assume that $\min \mathfrak{X}_N(t)$ grows at a speed $\leq w_2$, which we want to disprove. After a sufficiently long time, say t, (t can be chosen to be on the $(\log N)^3$ time scale), the survival of at least one of the killed BRWs leads to the presence of particles in $\mathfrak{X}_N(t)$ so far to the right of $w_2 \cdot t$ that such a particle can then branch without being affected by selection for more than $\log_2 N$ steps, leading to a population in the $N-$BRW comprising more than N particles, a contradiction.

The above argument is an oversimplification (in particular, we have argued as if $\min \mathfrak{X}_N(t)$ had an exactly linear growth over the time scales we consider), but the proof in **[27]** goes along this line. As a result, v_N cannot be smaller than any w_2 of the form (93), and one obtains a proof of (91).

For the proof of (92), one chooses a w_1 of the form

$$w_1 := v^* - \frac{C - \phi_1}{(\log N)^2}, \tag{95}$$

leading to the fact that

$$\rho(w_1) << \frac{1}{N}, \tag{96}$$

and accordingly

$$w_2 := v^* - \frac{C - \phi_2}{(\log N)^2}, \tag{97}$$

where $\phi_2 > \phi_1$.

The idea is then to show that, if $\min \mathfrak{X}_N(t)$ grows at a speed $\geq w_2$, at least one of the N independent BRWs must survive killing below a space-time boundary moving at speed w_1. In view of (96), this event has a small probability, so that v_N cannot be larger than any w_2 of the form (97), and we obtain a proof of (92). One has to take care of the fact that ρ is the probability of survival of the killed BRW up to an infinite time-horizon, while the survival events appearing in the bound involve a finite time-horizon. Still, it can be shown that the survival probability over a large enough time scale

(which again can be chosen to be of order $(\log N)^3$), leads to a good enough approximation of (96).

Note that it is not unfair to say that our main contribution in **[27]** was to find an appropriate way of putting together several pieces of arguments that had been developed by other authors. Indeed, the key estimate is Theorem 26 by Gantert, Hu and Shi, while the proof strategy of (92) owes much to the paper **[128]** by Pemantle, which deals with complexity bounds for algorithms seeking near optimal paths in branching random walks. Also, at the heuristic level, the existence of a link between the Brunet-Derrida behavior of a branching-selection particle system such as the one studied here, and the asymptotics of the survival probability for branching random walks killed below a linear space-time barrier, was already suggested in the papers **[60, 142]** by B. Derrida and D. Simon. Finally, let us mention that a preliminary version of **[27]**, see **[21]**, was completed by one of the authors (B.) before the results in **[80]** became publicly available. In **[21]**, only the $(\log N)^{-2}$ order of magnitude of the difference $v_* - v_N$ was established, in the special case where the step-distribution is Bernoulli.

3.2. Proof of Theorem 27. The proof strategy used by Gantert, Hu and Shi **[80]** to prove Theorem 26 is probabilistic in nature, and relies among other things, on a first-second moment argument, using a change-of-measure technique combined with refined "small deviations" estimates for random walk paths, and exploiting some ideas developed in **[98]** in the context of branching Brownian motion.

On the other hand, our proof relies on the characterization of the survival probability of the branching random walk as the solution of a non-linear convolution equation. Indeed, for $x \in \mathbb{R}$ and $t \in \mathbb{N}$, let $q_v(x,t)$ denote the survival probability for the t first steps, of the BRW starting with one particle at site x at time 0, when killing below a straight-line of slope v is applied. We also use the notation $q_v(x,\infty)$ to denote the probability of survival up to an infinite time horizon.

Analysis of the first step performed by the walk leads to the following equation

$$\begin{cases} q_v(x,t+1) = 2(q_v(\cdot,t) \star \mu)(x-v) - (q_v(\cdot,t) \star \mu)(x-v)^2, \ x \geq 0, \\ q_v(x,t+1) = 0, \ x < 0. \end{cases} \tag{98}$$

A purely analytical treatment of the above equation, making use of monotonicity properties, shows that $q_v(x,\infty)$ is then uniquely characterized, among a suitable class of functions, by being a stationary (with respect to time) solution of the equation, i.e.

$$\begin{cases} q_v(x,\infty) = 2(q_v(\cdot,\infty) \star \mu)(x-v) - (q_v(\cdot,\infty) \star \mu)(x-v)^2, \ x \geq 0, \\ q_v(x,\infty) = 0, \ x < 0. \end{cases} \tag{99}$$

In turn, this equation is (once more !) analogous to the F-KPP equation, as is apparent when rewritten in terms of $u(x,t) := q_v(-x+tv,\infty)$,

$$\begin{cases} \underbrace{u(x,t+1)-u(x,t)}_{\rightsquigarrow \frac{\partial u}{\partial t}} = \underbrace{(\tilde{u}(x,t)-u(x,t))}_{\rightsquigarrow \frac{\partial^2 u}{\partial x^2}} + \underbrace{\tilde{u}(x,t)-\tilde{u}(x,t)^2}_{\rightsquigarrow u-u^2}, \ x \le v(t+1), \\ u(x,t) = 0, \ x > vt, \end{cases} \tag{100}$$

where $\tilde{u}(x,t) := (u(\cdot,t) \star \tilde{\mu})(x)$ and $\tilde{\mu}$ is the image of μ by the map $x \mapsto -x$.

The idea is then to adapt the (non-rigorous methods) developed by Brunet and Derrida to study stochastic front propagation models, to treat this equation for v close to v_*. At the heuristic level, this approach was used by Derrida and Simon in **[60, 142]**. Our approach is inspired by the (rigorous) treatment of Mueller, Mytnik and Quastel of a continuous-time version of (100), which appears as a key intermediate step in their proof of Theorem 24 (see **[123, 124]**). The idea is to compare the solutions of the original non-linear equation to solutions of suitably adjusted linear approximations of it, for which explicit solutions are available. In the framework of **[123, 124]**, the corresponding equation is a second-order non-linear o.d.e., for which specific techniques (such as phase-plane analysis) can be applied, while such tools are not available in our discrete-time setting. Still, the monotonicity properties of (98) allow us to compare sub- and super- solutions to (99) to $q_v(\cdot,\infty)$.

In our opinion, one of the interests of the present proof is that, combined with the comparison approach used in **[27]**, it provides a justification of the velocity shift asymptotics (88) along the lines of the original analytic argument of Brunet and Derrida. Although this argument is based on an analysis of perturbations of (85) to which we were unable to give rigorous content, making a detour via the survival probability, which satisfies a dual version of the equation like (100), allowed us to obtain a rigorous proof.

4. Discussion

Theorem 25 is established for models with binary branching under assumptions (1) and (2) on the step distribution μ. The assumption of binary branching was made for the sake of simplicity, and it should not be difficult to generalize the results to models with supercritical stochastic branching with suitable tail decay. Similarly, the part of assumption (1) dealing with the left-tail of μ is made only for technical reasons. On the other hand, a slower than exponential decay of the right-tail for μ, or the absence of assumption (2), are expected to alter the validity of the theorem. Indeed, a sub-exponential tail for μ can lead to an infinite maximal speed for the corresponding branching random walk. On the other hand, in the Bernoulli case where $\mu = p\delta_1 + (1-p)\delta_0$, (2) breaks down when $p \ge 1/2$, and the conclusion of Theorem 25 does not hold (see e.g. the Appendix of **[28]** for a short discussion of the meaning of (2), and **[55]** for a detailed discussion of the $p > 1/2$ case).

For Theorem 27 too, the assumption of binary branching is made to simplify the exposition, but can be relaxed to allow stochastic branching

mechanisms. On the other hand, it is not clear whether an extension to general step distributions with unbounded support but e.g. exponential decay of the tail is possible.

Note that, even within the rather limited scope considered here, where reproduction is asexual, and individuals are identified with a single numerical fitness value, a wide variety of models can be considered. For instance, the effect of fitness on the population composition can be modeled in many distinct ways. Also, the specific assumptions on the frequency of mutations, and on the way they affect individual fitness values, can have a large impact upon the model's behaviour. This is also true of the kind of limit (with respect to population size, time and fitness scales, mutation rates) that is investigated. As a result, we are very far from having a complete picture of how such simple evolution models behave, and a quite rich set of open questions may be asked about them. One specific problem we have started to study with P. Maillard is the behaviour of the $N-$BRW model when the step distribution μ has a heavy tail.

Even though much progress has been made in turning Brunet and Derrida's predictions into mathematical theorems, substantial work remains to be done before a proper mathematical understanding of most of their results is achieved. For instance, in the $N-$BBM case, it is not yet proved that the second-order correction to the velocity is indeed given by (86), or that the genealogy is indeed described by the Bolthausen-Sznitman coalescent. Also, the mathematical developments do not match (for the moment !) the unity of the theoretical physics' approach which is able to treat in the same way such diverse objects as interacting particle systems and noisy PDEs.

One aspect of Brunet and Derrida's work we have not mentioned is the discovery of special families of branching selection models for which an exact computation of some quantities is possible, see **[47, 43]**. For instance, in the so-called *exponential model*, an individual with fitness value x gives birth to an infinite number of children whose fitness values form a Poisson process on $\mathbb{R}$ with intensity $e^{-(y-x)}dy$, and, starting from a population of N individuals, one keeps only the N children with the largest fitness values[2]. Although the asymptotic behaviour of this model is quite different from the binary branching $N-$BRWs considered above, one can match the asymptotics of these two models in a coherent way. Recently, Comets, Quastel and Ramírez **[51]** studied variants of one of these exactly solvable models, establishing precise asymptotic results, and showing a form of robustness of the behaviour of the model.

[2]This makes sense despite the infinite number of children, since $\mathbb{R}_+$ always contains a finite number of children.

Bibliography

[1] O. S. M. Alves, F. P. Machado, and S. Yu. Popov. The shape theorem for the frog model. *Ann. Appl. Probab.*, 12(2):533–546, 2002.

[2] O. S. M. Alves, F. P. Machado, S. Yu. Popov, and K. Ravishankar. The shape theorem for the frog model with random initial configuration. *Markov Process. Related Fields*, 7(4):525–539, 2001.

[3] Gideon Amir, Itai Benjamini, and Gady Kozma. Excited random walk against a wall. *Probab. Theory Related Fields*, 140(1-2):83–102, 2008.

[4] Christophe Andrieu and Dan Crisan, editors. *Conference Oxford sur les méthodes de Monte Carlo séquentielles*, volume 19 of *ESAIM Proceedings*. EDP Sciences, Les Ulis, 2007.

[5] P. F. Arndt, C. B. Burge, and T. Hwa. DNA sequence evolution with neighbor-dependent mutation. *J. Comput. Biol.*, 10(3–4):313–322, 2003.

[6] A. Asselah and A. Gaudillière. From logarithmic to subdiffusive polynomial fluctuations for internal DLA and related growth models. arXiv:math.PR/1009.2838.

[7] A. Asselah and A. Gaudillière. Lower bounds on fluctuations for internal DLA. arXiv:math.PR/1111.4233.

[8] A. Asselah and A. Gaudillière. Sub-logarithmic fluctuations for internal DLA. arXiv:math.PR/1011.4592.

[9] L. Avena, F. den Hollander, and F. Redig. Law of large numbers for a class of random walks in dynamic random environments. *Electron. J. Probab.*, 16:no. 21, 587–617, 2011.

[10] G. Baele, Y. Van de Peer, and D. Vansteelandt. Using non-reversible context-dependent evolutionary models to study substitution patterns in primate non-coding sequences. *J. Mol. Evol.*, 71(1):34–50, 2010.

[11] Márton Balázs, Firas Rassoul-Agha, and Timo Seppäläinen. The random average process and random walk in a space-time random environment in one dimension. *Comm. Math. Phys.*, 266(2):499–545, 2006.

[12] Anne-Laure Basdevant and Arvind Singh. On the speed of a cookie random walk. *Probab. Theory Related Fields*, 141(3-4):625–645, 2008.

[13] Anne-Laure Basdevant and Arvind Singh. Rate of growth of a transient cookie random walk. *Electron. J. Probab.*, 13:no. 26, 811–851, 2008.

[14] Anne-Laure Basdevant and Arvind Singh. Recurrence and transience of a multi-excited random walk on a regular tree. *Electron. J. Probab.*, 14:no. 55, 1628–1669, 2009.

[15] S. Behrens and M. Falconnet. Accurate estimations of evolutionary times in the context of strong CpG hypermutability. *J. Comput. Biol.*, 19(5):519–531, 2012.

[16] R. Benguria and M. C. Depassier. On the speed of pulled fronts with a cutoff. *Phys. Rev. E*, 75(5), 2007.

[17] R. Benguria, M. C. Depassier, and M. Loss. Upper and lower bounds for the speed of pulled fronts with a cut-off. *Europ. Phys. J. B*, 61(3):331–334, 2008.

[18] Itaı Benjamini, Gady Kozma, and Bruno Schapira. A balanced excited random walk. *C. R. Math. Acad. Sci. Paris*, 349(7-8):459–462, 2011.

[19] Itai Benjamini and David B. Wilson. Excited random walk. *Electron. Comm. Probab.*, 8:86–92 (electronic), 2003.

[20] J. Bérard. The empirical collision probability of a population of interacting ants. Unpublished manuscript, available at `http://math.univ-lyon1.fr/~jberard/modele-fourmis-boug.pdf`.

[21] J. Bérard. An example of Brunet-Derrida behavior for a branching-selection particle system on Z. arXiv:0810.5567.

[22] J. Bérard. Contribution à l'étude probabiliste des algorithmes d'évolution. *PhD thesis, Univ. Claude Bernard*, 2001.

[23] J. Bérard and A. Bienvenüe. Sharp asymptotic results for simplified mutation-selection algorithms. *Ann. Appl. Probab.*, 13(4):1534–1568, 2003.

[24] Jean Bérard. Asymptotics of a two-dimensional sticky random walk. arXiv:math/0405451.

[25] Jean Bérard. The almost sure central limit theorem for one-dimensional nearest-neighbour random walks in a space-time random environment. *J. Appl. Probab.*, 41(1):83–92, 2004.

[26] Jean Bérard. Genetic algorithms in random environments: two examples. *Probab. Theory Related Fields*, 133(1):123–140, 2005.

[27] Jean Bérard and Jean-Baptiste Gouéré. Brunet-Derrida behavior of branching-selection particle systems on the line. *Comm. Math. Phys.*, 298(2):323–342, 2010.

[28] Jean Bérard and Jean-Baptiste Gouéré. Survival probability of the branching random walk killed below a linear boundary. *Electron. J. Probab.*, 16:no. 14, 396–418, 2011.

[29] Jean Bérard, Jean-Baptiste Gouéré, and Didier Piau. Solvable models of neighbor-dependent substitution processes. *Math. Biosci.*, 211(1):56–88, 2008.

[30] Jean Bérard and Laurent Guéguen. Accurate estimation of substitution rates with neighbor-dependent models in a phylogenetic context. *Syst. Biol.*, 61(3):510–521, 2012.

[31] Jean Bérard and Alexis Huet. Sequential Monte Carlo methods for nucleotide substitution models. *In preparation.*

[32] Jean Bérard and Didier Piau. Coupling from the past times with ambiguities and perturbations of interacting particle systems. arXiv:math.PR/1206.4983.

[33] Jean Bérard and Didier Piau. Coupling times with ambiguities for particle systems and applications to context-dependent DNA substitution models. arXiv:math.PR/0712.0072.

[34] Jean Bérard and Alejandro Ramírez. Central limit theorem for the excited random walk in dimension $D \geq 2$. *Electron. Comm. Probab.*, 12:303–314 (electronic), 2007.

[35] Jean Bérard and Alejandro F. Ramírez. Fluctuations of the front in a one dimensional model for the spread of an infection. *In preparation.*

[36] Jean Bérard and Alejandro F. Ramírez. Large deviations of the front in a one-dimensional model of $X + Y \to 2X$. *Ann. Probab.*, 38(3):955–1018, 2010.

[37] J. Berestycki, N. Berestycki, and J. Schweinsberg. The genealogy of branching Brownian motion with absorption. *Ann. Probab.*, to appear.

[38] Julien Berestycki, Nathanaël Berestycki, and Jason Schweinsberg. Survival of near-critical branching Brownian motion. *J. Stat. Phys.*, 143(5):833–854, 2011.

[39] Erwin Bolthausen, Alain-Sol Sznitman, and Ofer Zeitouni. Cut points and diffusive random walks in random environment. *Ann. Inst. H. Poincaré Probab. Statist.*, 39(3):527–555, 2003.

[40] Mireille Bousquet-Mélou and Gilles Schaeffer. Walks on the slit plane. *Probab. Theory Related Fields*, 124(3):305–344, 2002.

[41] E. Brunet, B. Derrida, A. H. Mueller, and S. Munier. Noisy traveling waves: effect of selection on genealogies. *Europhys. Lett.*, 76(1):1–7, 2006.

[42] E. Brunet, B. Derrida, A. H. Mueller, and S. Munier. Phenomenological theory giving the full statistics of the position of fluctuating pulled fronts. *Phys. Rev. E*, 73(5):056126, May 2006.

[43] É. Brunet, B. Derrida, A. H. Mueller, and S. Munier. Effect of selection on ancestry: an exactly soluble case and its phenomenological generalization. *Phys. Rev. E (3)*, 76(4):041104, 20, 2007.

[44] Eric Brunet and Bernard Derrida. Shift in the velocity of a front due to a cutoff. *Phys. Rev. E (3)*, 56(3, part A):2597–2604, 1997.

[45] Éric Brunet and Bernard Derrida. Microscopic models of traveling wave equations. *Computer Physics Communications*, 121-122:376–381, 1999.

[46] Éric Brunet and Bernard Derrida. Effect of microscopic noise on front propagation. *J. Statist. Phys.*, 103(1-2):269–282, 2001.

[47] Eric Brunet and Bernard Derrida. Exactly soluble noisy traveling-wave equation appearing in the problem of directed polymers in a random medium. *Phys. Rev. E*, 70:016106, 2004.

[48] Olivier Cappé, Eric Moulines, and Tobias Rydén. *Inference in hidden Markov models*. Springer Series in Statistics. Springer, New York, 2005. With Randal Douc's contributions to Chapter 9 and Christian P. Robert's to Chapters 6, 7 and 13, With Chapter 14 by Gersende Fort, Philippe Soulier and Moulines, and Chapter 15 by Stéphane Boucheron and Elisabeth Gassiat.

[49] R. Chachick and A. Tanay. Inferring divergence of context-dependent substitution rates in Drosophila genomes with applications to comparative genomics Phylogenetic estimation of context-dependent substitution rates by maximum likelihood. *Mol. Biol. Evol.*, 29(7):1769–1780, 2012.

[50] O. F. Christensen, A. Hobolth, and J. L. Jensen. Pseudo-likelihood analysis of codon substitution models with neighbor dependent rates. *J. Comput. Biol.*, 12(9):1166–1182, 2005.

[51] Francis Comets, Jeremy Quastel, and Alejandro F. Ramírez. Last Passage Percolation and Traveling Waves. arXiv:1203.2368.

[52] Francis Comets, Jeremy Quastel, and Alejandro F. Ramírez. Fluctuations of the front in a stochastic combustion model. *Ann. Inst. H. Poincaré Probab. Statist.*, 43(2):147–162, 2007.

[53] Francis Comets, Jeremy Quastel, and Alejandro F. Ramírez. Fluctuations of the front in a one dimensional model of $X + Y \to 2X$. *Trans. Amer. Math. Soc.*, 361(11):6165–6189, 2009.

[54] Joseph G. Conlon and Charles R. Doering. On travelling waves for the stochastic Fisher-Kolmogorov-Petrovsky-Piscunov equation. *J. Stat. Phys.*, 120(3-4):421–477, 2005.

[55] Olivier Couronné and Lucas Gerin. A branching-selection process related to censored Galton-Watson processes. *Ann. Inst. Henri Poincaré Probab. Stat.*, to appear.

[56] Emilio De Santis and Mauro Piccioni. Exact simulation for discrete time spin systems and unilateral fields. *Methodol. Comput. Appl. Probab.*, 10(1):105–120, 2008.

[57] Pierre Del Moral. *Feynman-Kac formulae*. Probability and its Applications (New York). Springer-Verlag, New York, 2004. Genealogical and interacting particle systems with applications.

[58] Amir Dembo, Yuval Peres, and Ofer Zeitouni. Tail estimates for one-dimensional random walk in random environment. *Comm. Math. Phys.*, 181(3):667–683, 1996.

[59] F. den Hollander, R. dos Santos, and V. Sidoravicius. Law of large numbers for non-elliptic random walks in dynamic random environments. *Stochastic Process. Appl.*, to appear.

[60] B. Derrida and D. Simon. The survival probability of a branching random walk in presence of an absorbing wall. *Europhys. Lett. EPL*, 78(6):Art. 60006, 6, 2007.

[61] R.L. Dobrushin, V.I. Kryukov, and A.L. Toom, editors. *Stochastic Cellular Systems: Ergodicity, Memory, Morphogenesis*. Manchester University Press, Chichester, 1990.

[62] D. Dolgopyat and E. Kosygina. Scaling limits of recurrent excited random walks on integers. arXiv:math.PR/1201.0379.

[63] Dmitry Dolgopyat. Central limit theorem for excited random walk in the recurrent regime. *ALEA Lat. Am. J. Probab. Math. Stat.*, 8:259–268, 2011.

[64] Arnaud Doucet, Nando de Freitas, and Neil Gordon, editors. *Sequential Monte Carlo methods in practice*. Statistics for Engineering and Information Science. Springer-Verlag, New York, 2001.

[65] Freddy Dumortier, Nikola Popović, and Tasso J. Kaper. The critical wave speed for the Fisher-Kolmogorov-Petrowskii-Piscounov equation with cut-off. *Nonlinearity*, 20(4):855–877, 2007.

[66] L. Duret and P.F. Arndt. The impact of recombination on nucleotide substitutions in the human genome. *PLoS Genet.*, 4(5), 2008.

[67] L. Duret and N. Galtier. The covariation between TpA deficiency, CpG deficiency, and G+C content of human isochores is due to a mathematical artifact. *Mol. Biol. Evol.*, 17(11):1620–1625, 2000.

[68] Richard Durrett. *Probability models for DNA sequence evolution*. Probability and its Applications (New York). Springer, New York, second edition, 2008.

[69] Rick Durrett and Daniel Remenik. Brunet-Derrida particle systems, free boundary problems and Wiener-Hopf equations. *Ann. Probab.*, 39(6):2043–2078, 2011.

[70] Mikael Falconnet. Phylogenetic distances for neighbour dependent substitution processes. *Math. Biosci.*, 224(2):101–108, 2010.

[71] J. Felsenstein. *Inferring Phylogenies*. Sinauer, 2004.

[72] Pablo A. Ferrari. Ergodicity for spin systems with stirrings. *Ann. Probab.*, 18(4):1523–1538, 1990.

[73] Pablo A. Ferrari, Roberto Fernández, and Nancy L. Garcia. Perfect simulation for interacting point processes, loss networks and Ising models. *Stochastic Process. Appl.*, 102(1):63–88, 2002.

[74] R. A. Fisher. *The Genetical Theory of Natural Selection*. Clarendon Press, Oxford, 1930.

[75] R. A. Fisher. The wave of advance of advantageous genes. *Ann. Eugenics*, 7:355–369, 1937.

[76] Peter Gács. Reliable cellular automata with self-organization. *J. Statist. Phys.*, 103(1-2):45–267, 2001.

[77] A. Galves, N. Garcia, and E. Löcherbach. Perfect simulation and finitary coding for multicolor systems with interactions of infinite range. 2008, arXiv:0809.3494.

[78] A. Galves, N. Garcia, E. Löcherbach, and E. Orlandi. Kalikow-type decomposition for multicolor infinite range particle systems. 2010, arXiv:1008.2740.

[79] A. Galves, E. Löcherbach, and E. Orlandi. Perfect simulation of infinite range Gibbs measures and coupling with their finite range approximations. *J. Stat. Phys.*, 138(1-3):476–495, 2010.

[80] Nina Gantert, Yueyun Hu, and Zhan Shi. Asymptotics for the survival probability in a killed branching random walk. *Ann. Inst. Henri Poincaré Probab. Stat.*, 47(1):111–129, 2011.

[81] Olivier Gascuel, editor. *Mathematics of evolution and phylogeny*. Oxford University Press, Oxford, 2007.

[82] Olivier Gascuel and Mike Steel, editors. *Reconstructing evolution*. Oxford University Press, Oxford, 2007. New mathematical and computational advances.

[83] Andrew Gelman and Xiao-Li Meng. Simulating normalizing constants: from importance sampling to bridge sampling to path sampling. *Statist. Sci.*, 13(2):163–185, 1998.

[84] D. Graur and W. H. Li. *Fundamentals of Molecular Evolution*. Sinauer, 2000.

[85] Lawrence F. Gray. A reader's guide to P. Gács's "positive rates" paper: "Reliable cellular automata with self-organization". *J. Statist. Phys.*, 103(1-2):1–44, 2001.

[86] Olle Häggström and Jeffrey E. Steif. Propp-Wilson algorithms and finitary codings for high noise Markov random fields. *Combin. Probab. Comput.*, 9(5):425–439, 2000.

[87] J. B. S. Haldane. *The Causes of Evolution*. Longmans, Green and Co., London, New-York, 1932.

[88] Asger Hobolth. A Markov chain Monte Carlo expectation maximization algorithm for statistical analysis of DNA sequence evolution with neighbor-dependent substitution rates. *J. Comput. Graph. Statist.*, 17(1):138–162, 2008.

[89] M. Holmes and R. Sun. A monotonicity property for random walk in a partially random environment. *Stochastic Process. Appl.*, 122:1369–1396, 2012.

[90] Mark Holmes. Excited against the tide: A random walk with competing drifts. *Ann. Inst. H. Poincaré Probab. Statist.*, 48:745–773, 2012.

[91] D. G. Hwang and P. Green. Bayesian Markov chain Monte Carlo sequence analysis reveals varying neutral substitution patterns in mammalian evolution. *Proc. Natl. Acad. Sci. U.S.A.*, 101(39):13994–14001, 2004.

[92] Dima Ioffe and Yvan Velenik. Ballistic phase of self-interacting random walks. In P. Mörters and al., editors, *Analysis and Stochastics of Growth Processes and Interface Models*. Oxford University Press, 2008.

[93] M. Jara, G. Moreno, and A. F. Ramírez. Front propagation in an exclusion one-dimensional reactive dynamics. *Markov Process. Related Fields*, 14(2):185–206, 2008.

[94] David Jerison, Lionel Levine, and Scott Sheffield. Logarithmic fluctuations for internal DLA. *J. Amer. Math. Soc.*, 25(1):271–301, 2012.

[95] T.H. Jukes and C.R. Cantor. Evolution of protein molecules. In H.N. Munro, editor, *Mammalian protein metabolism*, pages 21–132. Academic Press, New York, 1969.

[96] H. Kesten, M. V. Kozlov, and F. Spitzer. A limit law for random walk in a random environment. *Compositio Math.*, 30:145–168, 1975.

[97] Harry Kesten. A renewal theorem for random walk in a random environment. In *Probability (Proc. Sympos. Pure Math., Vol. XXXI, Univ. Illinois, Urbana, Ill., 1976)*, pages 67–77. Amer. Math. Soc., Providence, R.I., 1977.

[98] Harry Kesten. Branching Brownian motion with absorption. *Stochastic Processes Appl.*, 7(1):9–47, 1978.

[99] Harry Kesten, Alejandro F. Ramírez, and Vladas Sidoravicius. Asymptotic Shape and Propagation of Fronts for Growth Models in Dynamic Random Environment. In *Probability in Complex Physical Systems*, volume 11 of *Springer Proc. Math.*, pages 195–224. Springer, Heidelberg, 2012.

[100] Harry Kesten and Vladas Sidoravicius. The spread of a rumor or infection in a moving population. *Ann. Probab.*, 33(6):2402–2462, 2005.

[101] Harry Kesten and Vladas Sidoravicius. A phase transition in a model for the spread of an infection. *Illinois J. Math.*, 50(1-4):547–634, 2006.

[102] Harry Kesten and Vladas Sidoravicius. A problem in one-dimensional diffusion-limited aggregation (DLA) and positive recurrence of Markov chains. *Ann. Probab.*, 36(5):1838–1879, 2008.

[103] Harry Kesten and Vladas Sidoravicius. A shape theorem for the spread of an infection. *Ann. of Math. (2)*, 167(3):701–766, 2008.

[104] C. Kipnis and S. R. S. Varadhan. Central limit theorem for additive functionals of reversible Markov processes and applications to simple exclusions. *Comm. Math. Phys.*, 104(1):1–19, 1986.

[105] A. Kolmogorov, I. Petrovsky, and N. Piscounov. Etude de l'équation de la diffusion avec croissance de la quantité de matière et son application à un problème biologique. *Bull. Univ. Etat Moscou Sér. Int. Sect. A Math. Mécan.*, 1(6):1–25, 1937.

[106] T. Komorowski, C. Landim, and S. Olla. *Fluctuations in Markov processes*, volume 345 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin, 2012. Time Symmetry and Martingale Approximation.

[107] Elena Kosygina and Thomas Mountford. Limit laws of transient excited random walks on integers. *Ann. Inst. Henri Poincaré Probab. Stat.*, 47(2):575–600, 2011.

[108] Elena Kosygina and Martin P. W. Zerner. Excited random walks: results, methods, open problems. arXiv:math.PR/1204.1895.

[109] Elena Kosygina and Martin P. W. Zerner. Positively and negatively excited random walks on integers, with branching processes. *Electron. J. Probab.*, 13:no. 64, 1952–1979, 2008.

[110] Gady Kozma. Excited random walk in three dimensions has positive speed. arXiv: math/0310305.

[111] Gady Kozma. Excited random walk in two dimensions has linear speed. arXiv: math/0512535.

[112] N. Lartillot and H. Philippe. Computing Bayes factors using thermodynamic integration. *Syst. Biol.*, 55(2):195–207, 2006.

[113] Y. Le Jan and S. Lemaire. Products of Beta matrices and sticky flows. *Probab. Theory Related Fields*, 130(1):109–134, 2004.

[114] Thomas M. Liggett. An improved subadditive ergodic theorem. *Ann. Probab.*, 13(4):1279–1285, 1985.

[115] Thomas M. Liggett. *Interacting particle systems*, volume 276 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, New York, 1985.

[116] Thomas M. Liggett. *Stochastic interacting systems: contact, voter and exclusion processes*, volume 324 of *Grundlehren der Mathematischen Wissenschaften*. Springer-Verlag, Berlin, 1999.

[117] G. Lunter and J. Hein. A nucleotide substitution model with nearest-neighbour interactions. *Bioinformatics*, 20(Suppl. 1):i216–i223, 2004.

[118] D. Mai, I. M. Sokolov, V. N. Kuzovkov, and A. Blumen. Front form and velocity in a one-dimensional autocatalytic $A+B \to 2A$ reaction. *Phys. Rev. E*, 56(4):4130–4134, 1997.

[119] P. Maillard. Mouvement brownien branchant avec sélection. *PhD thesis, Univ. Pierre et Marie Curie*, 2012.

[120] M. Menshikov, S. Popov, A. Ramírez, and M. Vachkovskaia. On a general many-dimensional excited random walk. *Ann. Probab.*, to appear.

[121] Y. Mohylevskyy, C. M. Newman, and K. Ravishankar. Ergodicity and Percolation for Variants of One-dimensional Voter Models. 2011, arXiv:1112.1893.

[122] Thomas Mountford, Leandro P. R. Pimentel, and Glauco Valle. On the speed of the one-dimensional excited random walk in the transient regime. *ALEA Lat. Am. J. Probab. Math. Stat.*, 2:279–296 (electronic), 2006.

[123] C. Mueller, L. Mytnik, and J. Quastel. Small noise asymptotics of traveling waves. *Markov Process. Related Fields*, 14, 2008.

[124] Carl Mueller, Leonid Mytnik, and Jeremy Quastel. Effect of noise on front propagation in reaction-diffusion equations of KPP type. *Invent. Math.*, 184(2):405–453, 2011.

[125] Rasmus Nielsen, editor. *Statistical methods in molecular evolution*. Statistics for Biology and Health. Springer, New York, 2005.

[126] D. Panja. Effects of fluctuations in propagating fronts. *Phys. Rep.*, 393:87–174, 2004.

[127] A.-M. K. Pedersen and J. L. Jensen. A dependent-rates model and a MCMC-based methodology for the maximum-likelihood analysis of sequences with overlapping reading frames. *Mol. Biol. Evol.*, 18:763–776, 2001.

[128] Robin Pemantle. Search cost for a nearly optimal path in a binary tree. *Ann. Appl. Probab.*, 19(4):1273–1291, 2009.

[129] J. Peterson. Large deviations and slowdown asymptotics for one-dimensional excited random walks. arXiv:math.PR/1201.0318.

[130] Agoston Pisztora, Tobias Povel, and Ofer Zeitouni. Precise large deviation estimates for a one-dimensional random walk in a random environment. *Probab. Theory Related Fields*, 113(2):191–219, 1999.

[131] James Gary Propp and David Bruce Wilson. Exact sampling with coupled Markov chains and applications to statistical mechanics. In *Proceedings of the Seventh International Conference on Random Structures and Algorithms (Atlanta, GA, 1995)*, volume 9, pages 223–252, 1996.

[132] Olivier Raimond and Bruno Schapira. Excited Brownian motions as limits of excited random walks. *Probab. Theory Related Fields*, to appear.

[133] Olivier Raimond and Bruno Schapira. Excited Brownian motions. *ALEA Lat. Am. J. Probab. Math. Stat.*, 8:19–41, 2011.

[134] A. F. Ramírez and V. Sidoravicius. Asymptotic behavior of a stochastic combustion growth process. *J. Eur. Math. Soc. (JEMS)*, 6(3):293–334, 2004.

[135] Firas Rassoul-Agha. Large deviations for random walks in a mixing random environment and other (non-Markov) random walks. *Comm. Pure Appl. Math.*, 57(9):1178–1196, 2004.

[136] F. Redig and F. Völlering. Random Walks in Dynamic Random Environments: A transference principle. *Ann. Probab.*, to appear.

[137] Christian P. Robert and George Casella. *Monte Carlo statistical methods*. Springer Texts in Statistics. Springer-Verlag, New York, second edition, 2004.

[138] Leonardo T. Rolla and Vladas Sidoravicius. Absorbing-state phase transition for driven-dissipative stochastic dynamics on $\mathbb{Z}$. *Invent. Math.*, 188(1):127–150, 2012.

[139] A. Rzhetsky and M. Nei. Tests of applicability of several substitution models for DNA sequence data. *Mol. Biol. Evol.*, 12(1):131–151, 1995.

[140] P. Schuster. Mathematical modeling of evolution. Solved and open problems. *Theory Biosci.*, 130:71–89, 2011.

[141] A. Siepel and D. Haussler. Phylogenetic estimation of context-dependent substitution rates by maximum likelihood. *Mol. Biol. Evol.*, 21(3):468–488, 2004.

[142] Damien Simon and Bernard Derrida. Quasi-stationary regime of a branching random walk in presence of an absorbing wall. *J. Stat. Phys.*, 131(2):203–233, 2008.

[143] Alain-Sol Sznitman. Slowdown estimates and central limit theorem for random walks in random environment. *J. Eur. Math. Soc. (JEMS)*, 2(2):93–143, 2000.

[144] Alain-Sol Sznitman and Martin Zerner. A law of large numbers for random walks in random environment. *Ann. Probab.*, 27(4):1851–1869, 1999.

[145] J. van den Berg and J. E. Steif. On the existence and nonexistence of finitary codings for a class of random fields. *Ann. Probab.*, 27(3):1501–1522, 1999.

[146] Remco van der Hofstad and Mark Holmes. Monotonicity for excited random walk in high dimensions. *Probab. Theory Related Fields*, 147(1-2):333–348, 2010.

[147] Remco van der Hofstad and Mark Holmes. An expansion for self-interacting random walks. *Braz. J. Probab. Stat.*, 26(1):1–55, 2012.

[148] A. W. van der Vaart. *Asymptotic statistics*, volume 3 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 1998.

[149] S. R. S. Varadhan. Large deviations for random walks in a random environment. *Comm. Pure Appl. Math.*, 56(8):1222–1245, 2003. Dedicated to the memory of Jürgen K. Moser.

[150] Cristiano Varin. On composite marginal likelihoods. *AStA Adv. Stat. Anal.*, 92(1):1–28, 2008.

[151] Cristiano Varin, Nancy Reid, and David Firth. An overview of composite likelihood methods. *Statist. Sinica*, 21(1):5–42, 2011.

[152] Stanislav Volkov. Excited random walk on trees. *Electron. J. Probab.*, 8:no. 23, 15, 2003.

[153] David Bruce Wilson. Perfectly Random Sampling with Markov Chains. `http://dimacs.rutgers.edu/~dbwilson/exact/`.

[154] S. Wright. Evolution in Mendelian populations. *Genetics*, 16:97–159, 1931.

[155] Ofer Zeitouni. Random walks in random environment. In *Lectures on probability theory and statistics*, volume 1837 of *Lecture Notes in Math.*, pages 189–312. Springer, Berlin, 2004.

[156] Martin P. W. Zerner. Multi-excited random walks on integers. *Probab. Theory Related Fields*, 133(1):98–122, 2005.

[157] Martin P. W. Zerner. Recurrence and transience of excited random walks on $\mathbb{Z}^d$ and strips. *Electron. Comm. Probab.*, 11:118–128 (electronic), 2006.