Stochastic Models of Epithelial Cancer Initiation and Glioblastoma Recurrence

A DISSERTATION
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Kathleen M. Storey

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Jasmine Foo

June, 2018

Acknowledgements

I am incredibly lucky to have such a wonderful support system, without whom this thesis would not have been possible. First, I would like to thank my advisor, Jasmine Foo, for introducing me to such a fascinating research area and for her guidance and mentorship during this journey. I would also like to thank Kevin Leder for his helpful insight and feedback throughout the last few years. Additionally I am grateful for the encouragement I have received from numerous teachers and mentors at Carleton College, SMP, and the Haverford school district.

I owe many thanks to my family: Mom, Dad, Eileen, and Pat, thank you for your endless loving support and guidance and for always making Havertown feel like home, even when I am 1200 miles away. I am also eternally grateful for the inspiration of my grandparents and entire extended family.

Thank you to my friends in the math department, especially Will, Danika, Sam, Daniel, Nicole, and honorary math department members Blake and Libby; some of my fondest grad school memories are of board games, spades, and Acadia Fridays with all of you. Will, in particular, thank you for leading the way through each graduate school hurdle and for your advice and encouragement as I followed each a year later – I cannot imagine completing this thesis without your partnership and constant support.

Abstract

Cancer development involves the inherently stochastic accumulation of genetic mutations, conferring growth advantages to the cells affected by these mutations. Thus, stochastic modeling provides useful insight when studying the evolutionary processes of cancer initiation and tumor progression. This thesis consists of three projects within the field of stochastic modeling of cancer evolution.

First we explore the temporal dynamics of spatial heterogeneity during the process of carcinogenesis from healthy tissue. We utilize a spatial stochastic model of mutation accumulation and clonal expansion to describe this process. Under a two-step carcinogenesis model, we analyze two new measures of spatial population heterogeneity. In particular, we study the typical length-scale of genetic heterogeneity during carcinogenesis and estimate the size of the clone surrounding a sampled premalignant cell.

Next we study the propagation speed of a premalignant clone during carcinogenesis. We approximate a premalignant clone in epithelial tissue containing w layers of proliferating cells (referred to as a "basal zone") with a biased voter model on a set of w stacked integer lattices. Using the dual process of the biased voter model, we determine the asymptotic propagation speed of the premalignant clone in this setting and compare it to the previously determined speed in epithelial tissue with a single layer of proliferating cells. We then use this speed to investigate clinical implications for primary tumors detected in various types of epithelial tissue.

Finally we develop a multi-type branching process model of the tumor progression and treatment response in glioblastoma multiforme (GBM). GBM recurrence is often attributed to acquired resistance to the standard chemotherapeutic agent temozolomide (TMZ). Promoter methylation of the DNA repair gene MGMT is frequently linked to TMZ sensitivity. We develop and parameterize a model using clinical and experimental data, to investigate the interplay between TMZ and MGMT methylation during GBM treatment. Our model suggests that TMZ may have an inhibitory effect on maintenance methylation of MGMT after cell division. Incorporating this effect, we study the optimal TMZ dosing regimen for GBM patients with high and low levels of MGMT methylation at diagnosis.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Cancer is a genetic evolutionary process involving abnormal cell growth. Typically this process involves the accumulation of multiple genetic alterations, disrupting the carefully regulated cellular behavior within healthy tissue. The genetic alterations can create an imbalance of cell proliferation and cell death, driving tumor growth, triggering the growth of new blood vessels, and eventually invading distant tissue to form metastatic tumors [3]. The mechanisms by which these hallmark events occur vary greatly between cancer types and individual patients, and many of these precise mechanisms remain poorly understood. Due to the complex and variable nature of cancer, mathematical models provide useful insight regarding the evolutionary dynamics of cancer initiation and tumor growth. In this thesis, I describe the use of stochastic models to study various aspects of cancer progression, including the process of tumor initiation from healthy tissue, the emergence of heterogeneity within a tumor, and the evolution of drug resistance.

The body of this thesis is divided into three chapters, each describing a project within the field of evolutionary cancer modeling. Chapters 2 and 3 utilize spatially structured stochastic models to study carcinogenesis in premalignant epithelial tissue. Chapter 4 focuses on the emergence of drug resistance within glioblastoma, a highly aggressive type of brain tumor, using a nonspatial branching process model.

In particular, the project described in Chapter 2 uses a spatial version of a Moran population model, previously analyzed in [4, 5, 6], in which cells are arranged in a d-dimensional lattice. In this spatial Moran model, each cell waits an exponentially distributed amount of time, with respect to its fitness, before dividing, and then randomly chooses one of its nearest neighbors to replace with its progeny. During division, cells can acquire random mutational advances that confer fitness advantages to the affected cells. Further details are

provided in section 2.1. In joint work with Marc Ryser[1], Kevin Leder[2], and Jasmine Foo[3], we use an approximation of the spatial Moran model to analyze the spatial heterogeneity that arises over time during carcinogenesis [7].

Spatial heterogeneity is an important clinical issue because it provides useful information for biopsy procedure. Standard biopsy practice involves taking a single sample from an arbitrary location in a suspected premalignant lesion. This sample is used to determine whether the lesion is benign, precancerous, or cancerous and to determine the treatment strategy with the highest likelihood of success. Due to heterogeneity, this single-biopsy approach can lead to a misdiagnosis; several biopsies across a suspected tissue region would help to ascertain the full extent of the lesion. In Chapter 2, we develop two spatial measures of heterogeneity that can be used to suggest how fine or coarse the spatial sampling should be and to predict the spatial extent of a sampled premalignant clone.

Chapter 3 also focuses on the premalignant phase of cancer development in a spatially structured population. In joint work with Jasmine Foo and Kevin Leder, we analyze the the dependence of the propagation speed of a mutant clone on the underlying structure of the affected tissue. The motivation for this project stems from the fact that some epithelial tissue-types have multiple layers of proliferating cells, which we refer to as a "basal zone." We approximate a basal zone with w layers as a set of w stacked two-dimensional lattices ($\mathbb{Z}^2 \times \mathbb{Z}/w\mathbb{Z}$), with cells occupying each lattice site, and we use a particle system known as the biased voter model (BVM) to model the interactions between normal and mutant cells. Under this framework, the propagation speed of a premalignant clone in a single basal layer ($w = 1$) has been studied previously in [5]. We use a similar approach in order to study the propagation speed in a basal zone with multiple layers of proliferating cells. This approach relies upon a duality relationship between the BVM and a branching coalescing random walk. Further details regarding this duality relationship are provided in section 3.2 and in [8].

After determining the propagation speed as a function of w, we apply the speed within the framework of 'field cancerization.' The term 'field cancerization,' or the 'cancer field effect,' refers to the observation that regions ('fields') surrounding primary tumors frequently have an increased risk for the development of recurrent tumors or multiple distinct primary tumors. Slaughter and colleagues first introduced this terminology in 1953 after repeatedly observing the emergence of multiple oral squamous cell cancers within a single region of

[1]Dept. of Surgery, Dept. of Mathematics, Duke University
[2]Dept. of Industrial and Systems Engineering, University of Minnesota
[3]Dept. of Mathematics, University of Minnesota

tissue [9]. In addition to oral squamous cell cancer, the field effect is commonly observed in Barrett's esophagus, ductal carcinoma in the breast, and prostate cancer, among others [10, 11, 12]. There is evidence that the clonal expansion of mutated cells, possessing fitness advantages over surrounding healthy cells, drives the process of field cancerization [13, 14, 15]. Thus, we can approximate a premalignant field using the group of mutant cells modeled by the BVM. When we look beyond a single premalignant clone and generalize this model to allow for further mutations during cell division, we obtain the spatial Moran model, described previously. Hence in section 3.5, we utilize the clonal propagation speed, as a function of w, to compare properties of premalignant fields in tissue with basal zones of varying thickness, and we discuss the resulting clinical implications.

Chapter 4 shifts the focus to malignant stages of cancer development. With collaborators Kevin Leder, Andrea Hawkins-Daarud[4], Kristin Swanson[5], Atique Ahmed[6], Russ Rockne[7], and Jasmine Foo, we developed a multitype branching process model to describe the evolutionary dynamics driving the progression of a highly aggressive type of brain tumor, known as glioblastoma multiforme (GBM). We incorporate standard treatment components for GBM, detailed in [2], which consist of surgical resection, radiation, and chemotherapy with the alkylating agent temozolomide (TMZ) . Increased expression levels of the DNA repair protein MGMT are associated with resistance to TMZ, and epigenetic silencing of the MGMT gene via promoter methylation is associated with TMZ sensitivity [16, 17, 18]. For this reason, we integrate detailed mechanisms of DNA methylation and demethylation within the model, using a variant of the model dynamics described in [1]. Then we investigate the role of MGMT demethylation in TMZ resistance during the standard treatment regimen for GBM. We aim to gain understanding of observed methylation patterns between GBM diagnosis and recurrence, indicating a frequent downward shift in methylation percentage between tumor detection and recurrence [19, 20]. We also explore optimal TMZ dosing strategies during the adjuvant chemotherapy treatment phase, and we compare the optimal dosing results, contingent upon MGMT methylation status at diagnosis.

[4]Dept. of Neurosurgery, Mayo Clinic Arizona
[5]Dept. of Neurosurgery, Mayo Clinic Arizona
[6]Dept. of Neurological Surgery, Northwestern University
[7]Beckman Research Institute, City of Hope National Medical Center

Chapter 2

Spatial measures of genetic heterogeneity during carcinogenesis

Carcinogenesis, the transformation from healthy tissue to invasive cancer, is a lengthy and complex process driven by a variety of factors including hereditary predisposition [21], exposure to environmental factors [22] and a changing microenvironment in the affected organ [23]. Irrespective of the driving factors, most cancers are characterized by the progressive accumulation of genetic alterations in a small group of founder cells. These alterations are either deleterious or neutral (passenger mutation), and some can confer a fitness advantage to the affected cell (driver mutation) by increasing the reproductive rate or inhibiting cell-regulatory mechanisms [24]. These selective advantages in turn lead to clonal expansion of a mutant cell population, which provides a fertile backdrop for further genetic alterations. Importantly, the underlying tissue architecture strongly influences the spatial growth patterns of the premalignant lesions, leading to complex patterns of spatial heterogeneity caused by competing and overlapping clones of various sizes and genetic ancestries [25].

The extent of spatial heterogeneity arising from this evolutionary process has been shown to correlate with clinical outcome. For example, genetic clonal diversity in premalignant tissue found in cases of Barrett's esophagus has been shown to predict progression to esophageal carcinoma [26]. However, the translation of heterogeneity into patient-specific clinical progression markers remains challenging because multiple point biopsies per patient are needed to reliably ascertain the degree of heterogeneity. Thus, there is a critical need for quantitative tools that (i) inform optimal sampling strategies, (ii) infer the degree of

heterogeneity in premalignant tissue based on sparse sample data, and (iii) predict the evolution of premalignant lesions and time scale of progression.

In this chapter we develop and analyze a cell-based stochastic model that describes the evolutionary process of cancer initiation in a spatially structured tissue. This model is a spatial version of a Moran population model, and has previously been analyzed in [4, 5, 6]. Using a mesoscopic approximation of this model, we analyze two spatial measures of heterogeneity that are relevant for the clinical setting. First, we study the probability that two samples, taken a fixed distance apart from each other, are genetically identical. This corresponds to a spatial analog of Simpsons' Index, a traditionally non-spatial measure of diversity which is defined as the probability that two individuals sampled at random from a population are identical. This measure, taken as a function of the distance between samples, provides an estimate of the length scale of heterogeneity in the premalignant tissue. Heterogeneity measures in premalignant conditions such as Barrett's esophagus have been correlated with likelihood of progression to esophageal cancer [26]. As a second measure of heterogeneity, we study the expected size of a premalignant lesion. This measure may be useful in scenarios where an isolated point biopsy indicates premalignant tissue without further information about the extent of the lesion. For both measures, we determine how they evolve during the transformation from healthy tissue to onset of malignancy, and we characterize their dependence on cancer-specific parameters such as mutation rates and fitness advantages. Due to the general formulation of the model, these results provide a useful tool for studying how heterogeneity and the extent of premalignant lesions vary between different cancer types.

The influence of spatial structure on the diversity of evolving populations has previously been studied in the ecological literature. Within that context, R.H. Whittaker introduced the measures of α-,β- and γ-diversity to denote the average species richness at the single habitat level (α), the diversity between habitats (β), and total species richness (γ) [27]. These measures are useful to quantify large scale organismal diversity in an ecological setting with spatial variation between well-defined habitats. However in this chapter we are interested in developing new measures of diversity to specifically explore the intrinsic length scales of genetic heterogeneity driven by clonal expansion dynamics in a spatially structured tissue population.

There have been other mathematical modeling efforts on the topic of heterogeneity during cancer initiation and expansion. In particular, previous work by Iwasa and Michor explored the Simpson's Index in a Moran process of tumorigenesis [28]. This study focused

on understanding the impact of neutral and advantageous mutations in a non-spatial, homogeneously mixed population setting. The work by Durrett et. al. [29] developed formulas for Simpson's Index and other heterogeneity measures in a multitype branching process model of cancer evolution. More recently, Dhawan and colleagues [30] developed a computational platform for the comparison of alternative spatial heterogeneity measures as potential biomarkers for tumor progression.

Finally, within the broader context of spatial tumor growth, our work adds to a vast body of literature. In [31], Williams and Bjerknes introduced the idea of a spatial model for clone spread in the basal layer of epithelial tissue and characterized the dynamics by means of simulations. In [32], Nowak, Michor, and Iwasa used a spatially explicit linear process to model the mechanisms that organisms have developed to slow down the cellular evolution leading to cancer. Komorova showed in [33] that the rate of cancer initiation is higher in a spatial model than in a well-mixed model when the first event is the inactivation of a TSG. Thalhauser et al. introduced a spatial Moran process that incorporates migration in [34], and Komorova expanded upon the migration model and looked at invasion probability in [35]. In addition, Durrett and Moseley studied how major results in the nonspatial Moran model change in the spatial version and considered how space changes the expected waiting time for a cell to develop two mutations [4].

The outline of this chapter is as follows: In Section 4.2 we introduce a cell-based stochastic evolutionary model of spatial carcinogenesis, as well as a mesoscopic approximation to this model that was analyzed in [5]. In Section 2.2 we first analyze the non-spatial Simpson's Index for this spatially-structured population. Then, in Section 2.3 we formulate and analyze two clinically relevant spatial measures of heterogeneity and study their dependence on cancer-specific parameters. Finally, we summarize and discuss our findings in Section 2.4.

2.1 Model

We introduce a spatial evolutionary model that describes the dynamic transition from physiological homeostasis to the onset of invasive cancer. In between, the tissue undergoes a sequence of genetic changes that manifest themselves at the phenotypic level in the form of increased proliferation rates, and hence a fitness advantage of mutant cells over normal cells. It is important to note that in many cancers, there is a succinct lack of a clearly defined genetic sequence [36]. On the other hand, the morphological changes from normal tissue to dysplasia, carcinoma in situ and invasive cancer is common in carcinomas, which

account for over 80% of all cancers. Therefore, one might prefer to interpret *mutations* as phenotypic transitions rather than genetic aberrations. With this interpretation in mind, we are going to introduce a linear 3-stage model, where type-0 cells represent normal tissue, type-1 cells are pre-malignant (dysplasia/CIS) and type-2 cells are malignant cancer cells. Since disadvantageous mutants die out exponentially fast and hence are very unlikely to produce a type-2 cell, we only model mutations with a net fitness advantage [37]. Advantageous mutants can arise in many different ways, but a common early event in carcinogenesis is loss of TP53 function. In this case, a mutation to the first allele of the gene can reduce function, and a second event, usually loss of the healthy allele, can lead to complete loss of function [38]. For notational simplicity, we assume that the fitness advantage is the same for both mutations. Note that this model can be extended to a setting with more than two mutations, either to represent a more refined phenotypic progression, or to account for select cancer-specific genetic events.

To render this model spatial, we introduce a cell-based stochastic model on the integer lattice $\mathbb{Z}^d \cap [-L/2, L/2]^d$, where $L > 0$, and equip this domain with periodic boundary conditions. On this lattice we have three different types of cells, labeled as type-0, type-1 and type-2. For $i \in \{0, 1, 2\}$ a type i cell reproduces at rate $(1 + s)^i$, and when the cell reproduces it replaces one of its $2d$ neighboring cells at random. In addition, we assume that for $i \in \{0, 1\}$, a type i cell mutates to type $i+1$ at rate u_{i+1}. Initially our entire lattice is occupied by type 0 cells which represent normal cells without any oncogenic mutations. Tumor initiation is defined as the birth of the first type-2 cell that does not go extinct. In the biological application we are interested in (somatic cells in the body), L is generally at least 10^6 while s, u_1 and u_2 are quite small. Therefore we will, unless stated otherwise, restrict our analysis to the regime $L \gg 1$, $u_1 \ll 1$, $u_2 \ll 1$, and $s \ll 1$. Before we can discuss the specific conditions imposed on the model parameters, we need to review the dynamic properties of the model.

In [5] we established that the arrival of type-1 mutants that are successful (i.e. whose progeny does not go extinct) can be described as a Poisson process with rate $u_1 s/(1 + s)$. Here, u_1 is the mutation rate to type-1 and $s/(1+s)$ is the survival probability of each type-1 mutant. We also characterized the radial expansion rate of type-1 families as a function of the selective advantage s in each dimension. In particular, it was established in [39, 40] that each successful type-1 family eventually has a convex, symmetric shape D, whose radius expands linearly in time. Let e_1 be the unit vector in the first axis, and let $c_d(s)$ be the linear expansion rate of the radius of this ball: $D \cap \{ze_1 : z \in \mathbb{R}\} = [-c_d(s), c_d(s)]$. Then,

we established that as $s \to 0$,

$$c_d(s) \sim \begin{cases} s & d = 1 \\ \sqrt{4\pi s/\log(1/s)} & d = 2 \\ \sqrt{4\beta_d s} & d \geq 3, \end{cases}$$

where β_d is the probability that two d dimensional simple random walks started at 0 and $e_1 = (1, 0, \ldots, 0)$ never collide. Through simulation, we studied the convergence of these clones in a previous work and demonstrated that a circular shape is reached fairly quickly (see [5] for details); thus we approximate the growth as circular from the clone's inception. With regards to the biology of specific cancer types, this asymptotic shape is a good approximation for clones in the premalignant stage as long as the microscopic model (the biased voter process) accurately reflects the dynamics of tissue maintenance. As anecdotal evidence of this, in [41], imaging shows premalignant lesions in Barrett's esophagus to be convex and circular or oval-shaped with smooth edges.

Based on this result, we then introduced a mesoscopic approximation to the model. Here, the growth of successful mutant families is deterministic, while the arrival of these families follows a non-homogeneous Poisson process. To ensure that this mesocopic model accurately recapitulates the dynamics of the cell-based model, we will make the following assumptions on the relationships between parameters in the model:

$$(A0) \qquad u_1 \ll 1/\ell(s)^{(d+2)/2} \tag{2.1}$$

$$(A1) \qquad \left(\frac{c_d}{u_2 s}\right)^{d/(d+1)} \ll N$$

$$(A2) \qquad (N u_1 s)^{d+1} (c_d^d u_2 s)^{-1} \to c \in [0, \infty) \tag{2.2}$$

$$(A3) \qquad u_2 \ll 1/\ell(s)$$

These assumptions generally hold for the parameter ranges appropriate for our biological application of carcinogenesis, see [5] for details. In addition, we will focus on dimensions $d = 1$ and $d = 2$, since most epithelial tissues can be viewed as one or two dimensional structures, e.g., the cells lining a mammary duct ($d = 1$), the crypts in the colon ($d = 2$), or the stratified squamous epithelia of the bladder, the cervix and the skin ($d = 2$).

In the simplified mesoscopic model we consider the cells to live on a spatial continuum $D = [-L, L]^d$. The state-space of the system is given by a set-valued function χ_t, which

characterizes the regions of D occupied by type-1 cells at time t. Mutations to type-1 cells occur as a Poisson process at rate $u_1 s$ in the set $\chi_t^c = D \backslash \chi_t$, i.e. in regions where type-0 cells reside. This Poisson process is non-homogeneous because the rate at which the mutations arrive depends on the size of χ_t^c, which depends on t. Each newly created type-1 mutation initiates an expanding ball whose radius grows linearly at rate c_d. Then after k mutations at the space-time points $\{(x_1, t_1), \ldots, (x_k, t_k)\}$, we have

$$\chi_t = \bigcup_{i=1}^{k} B_{x_i, c_d(t-t_i)},$$

where $B_{x,r} = \{y : ||y - x|| \leq r\}$ denotes the Euclidean ball of radius r and centered at x. Thus, the state of the system at any time t is the union of balls occupied by expanding mutant type-1 families. In [5] we proved that under assumption (A2) we can neglect the possibility that a second mutation arises from a type-1 family that dies out eventually. Therefore, we model successful type-2 mutations as Poisson arrivals into the space occupied by type-1 cells, χ_t, with rate $u_2 s$. Recall that in our two-step cancer initiation model, the type-2 mutant represents a malignant cancer cell.

We define the cancer initiation time σ_2 as the time when the first successful type-2 cell is born. Then, σ_2 is a random variable with complimentary cumulative distribution function given by

$$\mathbb{P}(\sigma_2 > t) = \mathbb{E} \exp\left(-u_2 s \int_0^t |\chi_t| dt\right),$$

where $|\chi_t|$ is the area of type-1 cells at time t.

2.2 Simpson's Index

Simpson's Index and Shannon's Index are two common measures of diversity, but we chose to look at Simpson's Index because its spatial analog lends itself well to informing biopsy procedure. Simpson's Index, a traditional non-spatial measure of heterogeneity, is defined as the probability that two individuals, sampled at random from a population, are genetically identical. More precisely, if there are N types of individuals in a population, the Simpson's Index is defined as

$$R = \sum_{i=1}^{N} \left(\frac{Y_i}{Y}\right)^2, \tag{2.3}$$

where Y_i is the number of individuals of the i-th type, and Y is the size of the entire population. Although this measure is usually used to characterize well-mixed populations, we investigate here how it evolves over time within the spatially structured population described by the mesoscopic spatial model from Section 4.2. In the cancer setting, one question of interest is to determine the degree of heterogeneity of the premalignant cell population. In our mesoscopic model, suppose there are N_t type-1 clones present at time $t > 0$. We then extend definition (2.3) as the time-dependent quantity

$$R(t) = \sum_{i=1}^{N_t} \left(\frac{Y_{1,i}(t)}{Y_1(t)} \right)^2, \tag{2.4}$$

where N_t is a Poisson random variable with parameter Nu_1st, $Y_{1,i}$ denotes the volume of the type-1 subclone originating from the ith type 1 mutation, and $Y_1(t)$ is the total volume of all the type-1 families present at time t, i.e. $Y_1(t) = \sum_{k=1}^{N_t} Y_{1,i}(t)$. From conditions (A0)-(A3) and Theorem 4 of [5] we know that overlaps between distinct type-1 clones occur with negligible probability by time σ_2. Define the event $A = \{\text{two clones overlap by } \sigma_2\}$, since $R(t) \in [0, 1]$ we have that

$$E[R(t); A^c] \leq E[R(t)] \leq E[R(t); A^c] + P(A).$$

Thus if $P(A) \ll 1$ we can safely ignore this overlap in the computation of Simpson's Index.

Building on the theory of size-biased permutations, it is possible to characterize the distribution of $R(t)$ as follows.

Proposition 2.2.1. *The conditional expectation of Simpson's Index for the spatial mesoscopic model is*

$$\mathbb{E}[R(t)|N_t = n] = n\mathbb{E}\left[\left(\frac{S_1}{S_n} \right)^2 \right], \tag{2.5}$$

where $S_n := B_1 + \ldots + B_n$ with B_i are i.i.d. $Beta(\frac{1}{d}, 1)$ random variables.

The proof of this result is found in Appendix A.1.2.

Proposition 2.2.2. *The conditional variance of the Simpson's Index is bounded as follows*

$$\left[\mathbb{E}\left(R(t) \mid N_t = n \right) \right]^2 \leq n \int_0^\infty \int_0^\infty \left(\frac{r}{x} \right)^3 \nu_1(r)\, \nu_{n-1}(x - r)\, dx\, dr, \tag{2.6}$$

where ν_k is the probability density function of S_k as defined above.

The derivation of this bound is found in Appendix A.1.3. Finally, the following result establishes the behavior of Simpson's Index for large n.

Proposition 2.2.3. *Conditioned on $N_t = n$, $R(t)$ converges to zero in probability as $n \to \infty$.*

This result tells us that as the number of clones increases, the probability of selecting two cells from the same clone goes to zero.

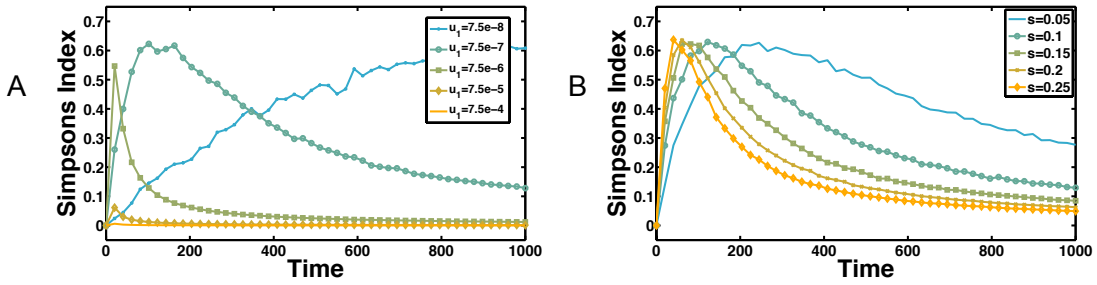


Figure 2.1: **Time-dependence of non-spatial Simpson's Index $R(t)$.** The temporal evolution of the expected value of the non-spatial Simpson's Index is shown **(A)** for varying values of the mutation rate u_1, and **(B)** for varying values of the fitness advantage s of preneoplastic cells over normal cells. In both simulations: $M_1 = M_2 = 500, N = 10^4$, and $u_1 = 7.5 \times 10^{-7}$, $s = 0.1$ unless specified.

Next, we use Monte Carlo simulations to evaluate (2.5) and study the temporal evolution of Simpson's Index (see Appendix A.1.5 for details on evaluating (2.5)). In Figure 2.1, we observe that the index first increases until it reaches a maximum, and then starts decaying in a monotone fashion. Essentially this result stems from the fact that in the early phase of the model the first few clones are developing and expanding, so the likelihood that samples come from the same clone increases. The population diversifies as more mutations are produced. Then Simpson's Index decreases as it becomes less likely for two cells to share the same family. Note that this is consistent with the result in Proposition 2.2.3. Figure 2.1A illustrates that as the mutation rate u_1 increases, this process of establishing mutant families and diversification occurs faster. In particular, the maximum Simpson's Index decreases with increasing mutation rate due to shrinking time periods during which a single clone exists. Finally, Figure 2.1B shows that as the selective advantage s increases, the growth of mutant families and diversification occurs sooner due to the faster spread of mutant cells. This result may imply that more aggressive tumors will have a higher level of heterogeneity.

2.3 Spatial measures of heterogeneity

In the clinical setting, the spatial heterogeneity of premalignant tissues poses considerable challenges. It is standard practice to take one biopsy sample from an arbitrary location in suspected premalignant tissue. Then clinicians typically use molecular information from this sample to determine whether it is (pre)cancerous as well as its specific cancer sub-type, if applicable. This information is used to help guide the diagnosis, prediction of prognosis, and treatment strategies. Due to the heterogeneity of premalignant tissue, such a single-biopsy approach may lead to incorrect subtype labeling or diagnoses and subsequently, to suboptimal therapeutic measures. For example, the spatial extent of this clone is unknown and thus surgical excision or prognosis prediction may be difficult. In view of these issues, the analysis of several biopsies across the tumor mass upon excision seems necessary. However, this raises another question of the *length-scale of heterogeneity*: how fine or coarse should the spatial sampling be, i.e. how many sections are required for a representative genetic fingerprint of the heterogeneous tissue?

In order to gain insight into these issues, we focus here on two specific clinical questions and introduce corresponding measures of spatial heterogeneity.

- Question 1: Given a region of premalignant tissue, what is the expected length-scale of heterogeneity? (i.e. how far apart should biopsy samples be taken?)
- Question 2: Provided that only a single point biopsy is available, what is the expected size of the clone present at the biopsy?

Before we introduce analytical expressions for these two measures of spatial heterogeneity I_1 and I_2, we introduce notation that will be useful below. Suppose two type-1 mutations occur at space-time points (x_0, t_1) and (y_0, t_2), respectively. Then the two clones will collide at time

$$t_* = \frac{t_1 + t_2}{2} + \frac{||x_0 - y_0||}{2c_d}.$$

Define the vector $v = (y - x)/||y - x||$. Then the first interaction between the two clones occurs at location

$$v_* = x_0 + c_d v(t_* - t_1) = y_0 - c_d v(t_* - t_2).$$

Next define the half-spaces

$$H_+ = \{x \in \mathbb{R}^d : \langle x, v_* \rangle > 0\} \text{ and } H_- = \{x \in \mathbb{R}^d : \langle x, v_* \rangle < 0\},$$

where $\langle \cdot, \cdot \rangle$ is the inner product in $\mathbb{R}^d$. If $x_0 \in H_+$ and $y_0 \in H_-$ then the region of space

influenced by the mutation that occurred at (x_0, t_1) is

$$B_{x_0, c_d(t-t_1)} \cap H_+,$$

and similarly the region influenced by the mutation that occurred at (y_0, t_2) is given by

$$B_{y_0, c_d(t-t_2)} \cap H_-.$$

Note that we still have

$$\chi_t = B_{x_0, c_d(t-t_1)} \cup B_{y_0, c_d(t-t_2)},$$

but we have decomposed χ_t into regions influenced by the two distinct mutations.

2.3.1 Spatial measure I_1: length scale of heterogeneity

A mutation at point (t_i, x_i) generates a ball $B_{x_i, c_d(s)(t-t_i)}$ growing linearly in t. Thus at time $t > t_i$, barring interference, the type-1 family is of size

$$Y_{1,i}(t) = \gamma_d c_d^d(s) \, (t - t_i)^d \,, \tag{2.7}$$

where γ_d is the volume of the d-dimensional unit sphere. To determine the length-scale of spatial heterogeneity, consider a fixed distance $r > 0$ and pick two cells separated by r uniformly at random. We define $I_1(r, t)$ to be the probability that these two cells are genetically identical (from the same mutant clonal expansion) at time t. The functional dependence of $I_1(r, t)$ on r provides an estimate of the length scale of heterogeneity and thus may provide guidance on sampling procedures. For example, a suggested sampling distance $r_{.5} \equiv \{\text{argmin}_{r>0} I_1(r, t) < 0.5\}$ between biopsies would ensure that sampled clones would be genetically different from neighboring samples 50 percent of the time. The measure $I_1(r, t)$ is a spatial analog of the Simpson's Index.

The actual analysis of $I_1(r, t)$ is quite technical so we will leave the details to the Appendix for interested readers. However, here we will provide some intuition for our approach and also provide some graphs demonstrating the dependence of I_1 on parameters and time. We will also provide some comparisons between our analysis (based on the mesoscopic model approximation) and simulations of I_1 in the full cell-based stochastic evolutionary model.

Idea behind calculation. First, the following are the main steps involved in the proofs.

1. Define space-time regions V_a and V_b that can influence the state of the cell samples, i.e. if a mutation occurs in V_a or V_b, then its clone can spread to site a or b, respectively,

by time t

2. Split up the probability that the sampled cells are different by conditioning on the number of mutations that occur in the space-time region $V_a \cup V_b$ (greater than two mutations in $V_a \cup V_b$ will have negligible probability on our time scale)

3. Use the Poisson distribution and the volume of the regions to calculate the probability that one or two mutations occur in the region and the probability that the cells are different, given that one mutation occurs

4. For the two mutation case, calculate the probability that the cells are different by splitting up $V_a \cup V_b$ into subregions and then conditioning on the subregion containing the first mutation

5. Define associated regions based on the location of the first mutation, which represent the section of $V_a \cup V_b$ in which a second mutation will make the sampled cells either different or the same

6. Obtain the probability that the two cells are different by integrating over each subregion, using the associated regions mentioned in the previous step

7. Subtract the probability that the two cells are different from 1 to obtain I_1

Now we will describe these steps in a little more detail. Let a, b be the positions of two cell samples taken at time t_0 and assume that $\|a-b\| = r$. Define D_{ab} as the event that the cells at positions a and b are genetically different at time t. Calculations in sections A.2.1 and A.2.2 of the Appendix demonstrate that $\mathbb{P}(D_{ab})$ only depends on the distance between the samples $\|a-b\|$, as long as $c_d t + r \ll L$. Thus, conditioning on the location of a and b we can conclude that $I_1(r,t) = 1 - \mathbb{P}(D_{ab})$.

Next we discuss the idea behind calculating $\mathbb{P}(D_{ab})$. Recall that if two clones meet, then each continues to spread in all directions away from the interacting clone. We will denote a cell in position x at time t by the coordinate (x,t). $V_a(t_0)$ and $V_b(t_0)$ are the space-time regions in which a mutation can influence the genetic state of the samples at a and b, respectively. The union of these regions $(V_a(t_0) \cup V_b(t_0))$ represents the space-time region in which mutations can influence the genetic state of the samples examined at locations a or b at time t_0. Let $E(A)$ be the number of mutations that occur in a region A, and let E_k be the event $\{E(V_a(t_0) \cup V_b(t_0)) = k\}$. Then, the event D_{ab} can be divided into sub-events

according to how many mutations have occurred in the spacetime region $V_a(t_0) \cup V_b(t_0)$

$$\mathbb{P}(D_{ab}) = \sum_{i=1}^{\infty} \mathbb{P}(D_{ab} \cap E_i)$$

The following simple calculation demonstrates that the probability of more than two type-1 mutations occurring in the region $V_a(t_0) \cup V_b(t_0)$ is small for the carcinogenesis setting. First, we note that the volume $|V_a(t_0) \cup V_b(t_0)|$ is bounded above by $|V_a(t_0)| + |V_b(t_0)|$. In 1D this sum of volumes is $c_1(s)t_0^2$, and in 2D it is $\frac{2\pi}{3} c_2(s)^2 t_o^3$, where $c_1(s)$ and $c_2(s)$ are the spreading speeds of the single mutant clones in dimensions 1 and 2 respectively, provided in Section 4.2. Since type-1 mutations arrive into $V_a(t_0) \cup V_b(t_0)$ as a Poisson process with rate $u_1 s$, the number of mutations in $V_a(t_0) \cup V_b(t_0)$ is stochastically dominated by a Poisson random variable with rate $\lambda = u_1 s^2 t_0^2$ in 1D and

$$\lambda = u_1 s t_0^3 \frac{4\pi s}{\log(1/s)} \frac{2\pi}{3}$$

in 2D. We note that t_0 is a time prior to carcinogenesis when premalignant tissue is sampled, and in our model tumor initiation occurs at time σ_2 when the first successful cell with two mutations arises. In [5] we found that the appropriate time scale of this process is $1/N u_1 s$. Replacing t_0 by this time scale in the Poisson rate λ in each dimension we obtain $\lambda \equiv 2/N^2 u_1$ in 1D and $\lambda \equiv \frac{8\pi^2}{3N^3 u_1^2 s \log(1/s)}$ in 2D.

We assume the point mutation rate in healthy tissue is within the range of 10^{-7} to 10^{-10} per base pair per cell division [42, 43, 44]. Selection advantages are more difficult to ascertain experimentally, but one study has estimated the average advantage s to be approximately 0.004 [24]. Lastly, the cell population sizes of interest in tissues at risk of initiating cancer are in the range of 10^6 and upward. Using these estimates we can easily calculate that the probability of more than two mutations arriving within the region of interest is negligible across all reasonable parameter ranges. Thus we can approximate $\mathbb{P}(D_{ab})$ with the first two terms of the sum above:

$$\mathbb{P}(D_{ab}) \approx \mathbb{P}(D_{ab} \cap E_1) + \mathbb{P}(D_{ab} \cap E_2). \tag{2.8}$$

We then use Bayes' Theorem to obtain

$$\mathbb{P}(D_{ab}) \approx \mathbb{P}(D_{ab}|E_1)\mathbb{P}(E_1) + \mathbb{P}(D_{ab}|E_2)\mathbb{P}(E_2).$$

We use the Poisson distribution with parameter $u_1 s|V_a(t_0) \cup V_b(t_0)|$ to calculate $\mathbb{P}(E_1)$

and $\mathbb{P}(E_2)$. Then we define the following notation

$$D(r,t_0) = V_a(t_0)\Delta V_b(t_0)$$
$$M(r,t_0) = V_a(t_0) \cap V_b(t_0).$$

Then given that there is one mutation, the cells will only be different if that mutation occurs in $D(r,t_0)$, so we have

$$\mathbb{P}(D_{ab}|E_1) = \frac{|D(r,t_0)|}{|V_a(t_0) \cup V_b(t_0)|}.$$

In 1-D, if there are two mutations, then we divide $V_a(t_0) \cup V_b(t_0)$ into seven regions R_i and condition on whether the first mutation occurrs in each of these regions. The most important distinction between the regions is whether they are located in $M(r,t_0)$ or $D(r,t_0)$. Each region R_i has an associated region Z_i. If $R_i \in M(r,t_0)$, then Z_i represents the region in which a second mutation makes the sampled cells different. If $R_i \in D(r,t_0)$, then Z_i represents the region in which a second mutation makes the cells the same.

Then we calculate $\mathbb{P}(D_{ab}|E_2)$ by summing the probability that the first mutation occurs in R_i and the second occurs in Z_i for each $R_i \in M(r,t_0)$ or the probability that the first mutation occurs in R_i and the second does not occur in Z_i for each $R_i \in D(r,t_0)$.

We integrate over the volume of the regions to find the probability that the second mutation occurs in Z_i, given that the first occcurs in R_i, as shown below

$$\mathbb{P}(X_2 \in Z_i | X_1 \in R_i) = \frac{1}{|R_i|} \int_{R_i} \frac{|Z_i(x,t)|}{|V_a \cup V_b \setminus A_i(x,t)|} dxdt,$$

where $A_i(x,t)$ is the region that is affected by the first mutation at (x,t), and thus, the region where a second mutation cannot occur.

In two dimensions we do not split the space into seven regions, but we use a similar process and condition on whether the first mutation occurs in $M(r,t_0)$ or $D(r,t_0)$. The resulting probability calculations are more complicated in the two-dimensional case. We do not compute I_1 in three dimensions, but we expect it to be much more technically demanding because the space-time diagram will be four dimensional. In addition we have found in most epithelial cancers that the early carcinogenesis process is driven by dynamics in the epithelial basal layer, so the 2-dimensional setting is the most relevant.

The exact calculations for $I_1 = 1 - \mathbb{P}(D_{ab})$ are provided in Sections A.2.1 and A.2.2 of the Appendix.

Agreement with microscopic model simulations. To verify our results we simulated the full cell-based stochastic evolutionary model and compared the spatial measure $I_1(r,t)$ with our derivations from the previous section. Table 2.1 shows the results of these comparisons. Since the cell-based model is very computationally intensive, only 100 simulations were performed in each set of parameter values; however the Wald confidence intervals are provided for each set of simulations. In this table we see a close agreement between our theoretical values for I_1 based on the mesoscopic model and the simulations of I_1 based on the microscopic model. For all cases, $N = 10000$ and $r = 10$.

Table 2.1: **Comparison of $I_1(r,t)$ between theory and simulation of the cell-based stochastic model.**

u_1	s	t	Theory $I_1(r,t)$	Simulation $I_1(r,t)$	95% CI
0.001	0.01	30	0.98	0.97	(0.94, 1)
0.001	0.01	40	0.96	0.92	(0.87, 0.97)
0.001	0.01	50	0.94	0.9	(0.84, 0.96)
0.001	0.01	60	0.90	0.9	(0.84, 0.96)
0.001	0.01	70	0.88	0.83	(0.76, 0.90)
0.0001	0.1	30	0.9	0.95	(0.91, 0.99)
0.0001	0.1	40	0.84	0.91	(0.85, 0.97)
0.0001	0.1	50	0.8	0.84	(0.77, 0.91)
0.0001	0.1	60	0.79	0.8	(0.72, 0.88)

In Figures 2.2 and 2.3 we demonstrate how the spatial measure of heterogeneity varies in time for different parameters. In Figure 2.2 we observe that as time increases, $I_1(r,t)$ decreases; this reflects the clonal expansion of existing mutant families leading to an increase in heterogeneity over time. In addition, as the distance r increases the probability of the two samples being genetically the same decreases, as expected. Figure 2.3 shows this result as a function of u_1. As the mutation rate increases, $I_1(r,t)$ decreases since the heterogeneity of the tissue increases as more mutant clones emerge. In both figures we see that sensitivity of I_1 to parameter changes increases as r increases. This is natural since the likelihood that two cells are identical is more likely to change as the distance between the cells increases.

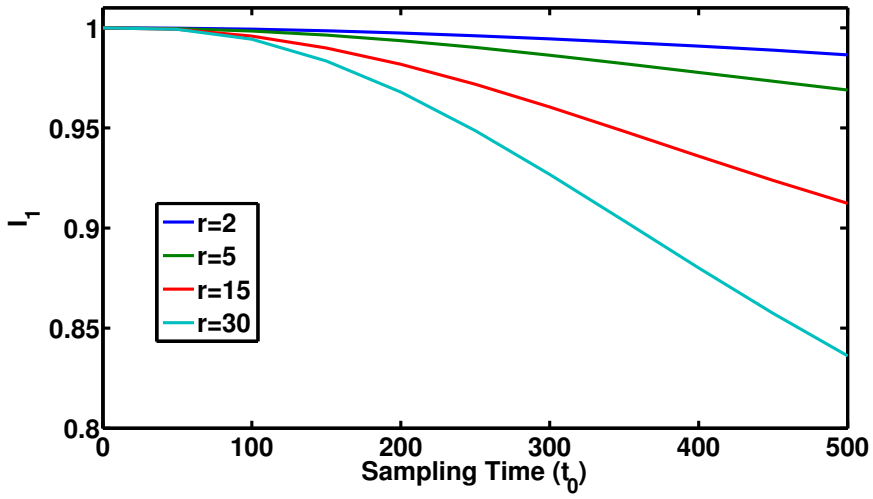


Figure 2.2: I_1 **in 2D as a function of sampling time** t_0**.** We vary the sampling radius r and set $s = 0.01$ and $u_1 = 1e-5$, so the mutation rate is $1e-7$. We also set the mutant growth rate $c_d = 0.25$.

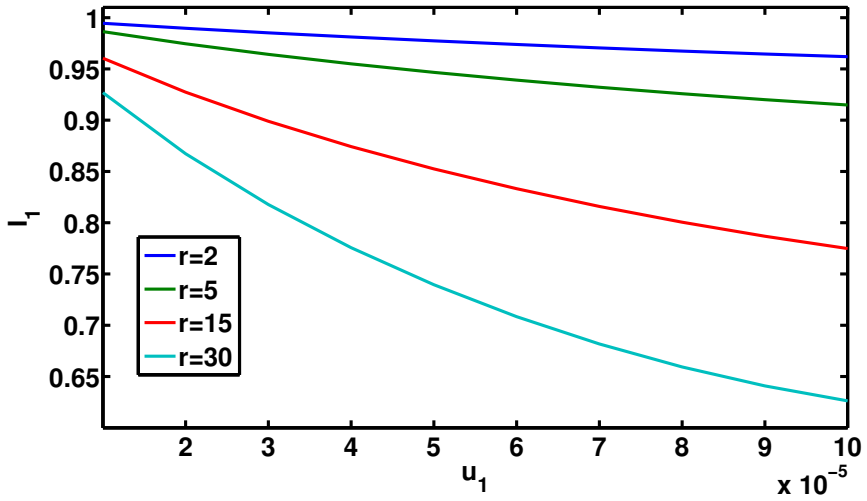


Figure 2.3: I_1 **in 2D as a function of** u_1**, which contributes to the mutation rate.** Mutations arise according to a Poisson process with rate $u_1 s$, and we set $s = 0.01$. We vary the sampling radius r and set the sampling time $t_0 = 300$ and the mutant growth rate $c_d = 0.25$.

2.3.2 Spatial measure I_2: extent of a premalignant lesion

Next, suppose we have obtained a premalignant (type-1) biopsy at a single point in the tissue at time t. We would like to estimate the expected size of the corresponding clone. In particular we define $I_2(r,t)$ to be the probability that an arbitrarily sampled cell at distance r is from the same clone as the original sample.

In order to study I_2 it is necessary to define the concept of 'size-biased pick' from a sequence of random variables $\{X_i\}_{i\geq 1}$.

Definition 1. *A size-biased pick from the sequence (X_i) is a random variable $X_{[1]}$ such that*

$$\mathbb{P}\left(X_{[1]} = X_i \mid X_1, \ldots X_n\right) = \frac{X_i}{X_1 + \ldots + X_n}.$$

Suppose that by time t there have been successful mutations at space time points $\{(x_i,t_i)\}_{i=1}^N$ initiating populations $C_{1,i}$, $1 \leq i \leq N$. In the model description we stated that the assumptions (A0)-(A3) hold throughout the paper. A consequence of these assumptions, proved in Theorem 4 of [5], is that overlaps between distinct type-1 cell is unlikely. Thus we assume here that for $1 \leq i \leq N$, $C_{1,i} = B_{x_i, c_d(t-t_i)}$.

In order to calculate I_2, we first choose a clone $C_{[1]}$ via size-biased pick from the different clones $\{C_{1,i}\}$. The radius of this pick is denoted $R_{[1]}$. For ease of notation we will make the following substitution throughout the rest of the section $C = C_{[1]}$ and $R = R_{[1]}$. We next choose a point, p_1, at random from the chosen clone C. We choose a second point p_2, at random a distance r away from p_1. In other words, the point p_2 is chosen at random from the circle

$$\mathcal{S} = \{x \in \mathbb{R}^2 : |x - p_1| = r\}.$$

To calculate I_2 we are interested in determining the probability that p_2 is contained in C. More specifically, let us denote the center of C by x_o. It is useful to define $X = |p_1 - x_o|$ which is a random variable with state space $[0,R]$. The heterogeneity measure $I_2(r,t)$ is given by:

$$I_2(r,t) \equiv P(p_2 \in C) = E\left[P\left(p_2 \in C | R, X\right)\right]. \tag{2.9}$$

The following two properties are useful in determining I_2:

(i) If $X + r \leq R$, then $\mathcal{S} \subset C$. To see this consider $z \in \mathcal{S}$. Then

$$|z - x_o| = |z - p_1 + p_1 - x_o| \leq |z - p_1| + |p_1 - x_o| \leq r + X \leq R.$$

(ii) If $R + X < r$ then $\mathcal{S} \cap C = \emptyset$. To see this take $z \in C$, which of course implies $|z - x_o| \leq R$ and thus,

$$|z - p_1| = |z - x_o + x_o - p_1| \leq R + X < r.$$

We then have that

$$P(p_2 \in C|R, X) = \begin{cases} 0, & R + X < r \\ 1, & X + r \leq R \\ \phi(X, R), & \text{otherwise.} \end{cases} \tag{2.10}$$

We can use the cosine rule to see that

$$\phi(X, R) = \frac{1}{\pi} \cos^{-1}\left(\frac{X^2 + r^2 - R^2}{2Xr}\right). \tag{2.11}$$

Substituting expressions (2.10) and (2.11) into (2.9) results in a formula for I_2 that can be easily approximated via Monte Carlo simulation. Details of this procedure are provided in Appendix A.3.

The heterogeneity measure I_2 is designed to be an estimate of the extent of a premalignant lesion that has already been detected via one point biopsy. Thus it is of interest to determine the value of I_2 at the time of detection of the premalignant condition (which itself may be random). Premalignant detection is realistic for some types of cancer that are tested early, e.g. the premalignant stages of oral squamous cell carcinomas [45] and esophageal adenocarcinoma in Barrett's esophagus [26, 46]. We hypothesize that detection of the premalignancy may occur at a random time τ, which occurs with a rate proportional to the total man-hours of premalignant lesions. In other words, detection of the condition is driven by the size and duration of premalignant lesion presence. Let us define τ with the following:

$$\mathbb{P}(\tau > t) = \mathbb{E} \exp\left(-\mu \int_0^t |\chi_t| dt\right), \tag{2.12}$$

where we recall that $|\chi_t|$ is the volume of type-1 cells at time t. Display (2.12) tell us that detection occurs at rate μ. Note we assume that (A1-A3) hold with $u_2 s$ replaced by μ.

In Section A.3 of the Appendix we also develop a numerical approach for estimating

I_2 at the random detection time τ. Interestingly enough, it is computationally easier to compute $I_2(r,\tau)$ than $I_2(r,t)$.

Numerical examples. In Figure 2.4, $I_2(r,t)$ is plotted as a function of r for various values of u_1, s, and t. Figure 2.5 shows analogous plots of $I_2(r,\tau)$ at the random detection time. Comparing Figures 2.4 and 2.5 we observe an interesting phenomenon. In particular when looking at I_2 at a fixed time in Figure 2.4 we see that for each r and t, I_2 is an increasing function of both u_1 and s. This makes sense if we consider the system at a fixed time. Then increasing the mutation rate will increase the expected growing time of any clones, i.e., they are more likely to be born earlier; therefore any clone we select is likely to be larger, so it is more likely that the second point selected a distance r away will be in the original clone. Similarly increasing s increases the expected size of clones present at time t, and thus increases $I_2(r,t)$. However, when we look at Figure 2.5 we see that $I_2(r,\tau)$ is decreasing in u_1. Interestingly by observing the process at the random time τ we flip the dependence on the parameters u_1. This phenomenon results from the fact that increasing the mutation rate allows for detection to be caused by multiple clones, which will therefore be smaller at detection than if the detection were driven by a single clone.

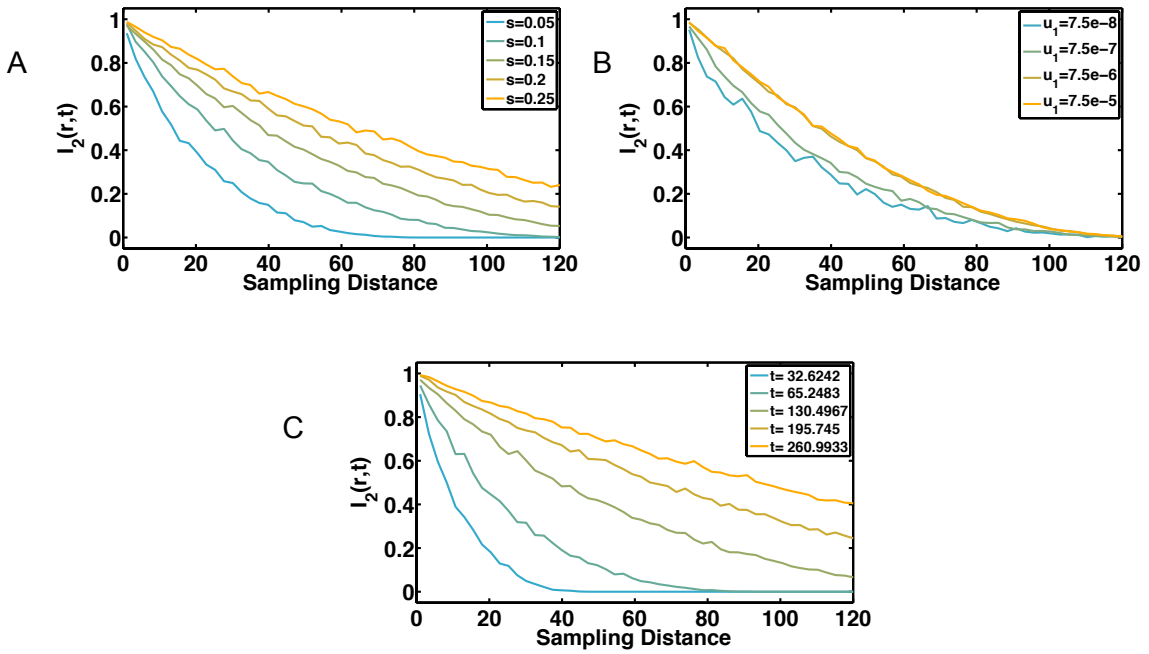


Figure 2.4: $I_2(r,t)$ **in 2D as a function of sampling radius.** Displayed for (A) varying selection strength, s, (B) varying u_1, and (C) varying t. In all panels $N = 2e5$, and $1e4$ Monte Carlo simulations are performed. Unless varied, $s = 0.1$, $u_1 = 7.5e-7$, and t is the median of the detection time τ with $\mu = 2e-6$.

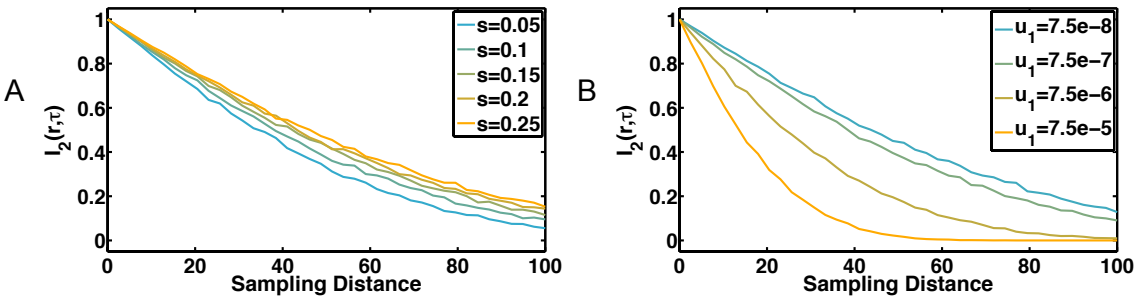


Figure 2.5: $I_2(r,\tau)$ **in 2D as a function of sampling radius.** In panel (A) we vary the selection strength, s, and in panel (B) we vary u_1. In all panels $N = 2e5$, we use $1e4$ Monte Carlo simulations, and for the random detection time τ we use $\mu = 2e-6$. If not mentioned we set $s = 0.1$, $u_1 = 7.5e-7$.

2.4 Discussion

In this chapter we have analyzed and examined several measures of heterogeneity in a spatially structured model of carcinogenesis from healthy tissue. In particular, we first derived estimates of the traditionally nonspatial measure of diversity, Simpson's Index, in the premalignant tissue and studied how the Simpson's Index changes in time and varies with parameters. We observed that as expected, the Simpson's Index decreases over time as more mutants are produced, and that this process occurs faster in settings with higher mutation rates or with larger selective advantages. We also formulated and analyzed two spatially-dependent measures of population heterogeneity, motivated by clinical questions. In particular we analyzed a measure (I_1) that can identify the length-scale of genetic heterogeneity during carcinogenesis, as well as a measure (I_2) that can estimate the extent of a surrounding premalignant clone, given a premalignant point biopsy.

We note that in this work we have confined our analysis to a two-step model of carcinogenesis. The results can be used in a setting in which a larger number of genetic hits is required for full malignant transformation. However our heterogeneity estimates would apply to the population of cells with a single mutation and thus can be used in the setting of early stages of carcinogenesis only. Incorporating further mutations in the model will be the subject of future work.

These analyses facilitate a better understanding of how to interpret discrete (in both time and space) samples from a spatially evolving population during carcinogenesis. For example, the quantity I_2 can be calculated to help determine the expected size of a premalignant lesion, given a point biopsy that is premalignant. In addition the quantity I_1 may be used to generate suggestions for optimal sample spacing in situations where multiple

biopsies or samples are possible. Finally, we note that although it is possible to calculate these diversity indices using computational simulation of similar cell-based or agent-based models, it can be extremely computationally onerous to simulate such models for even small sized lattices (100x100 sites) for a lengthy period of time, such as during the process of carcinogenesis. Therefore the heterogeneity estimates we derive based on our mesoscopic model in many cases provide the only feasible way to estimate spatial diversity in models living on a larger lattice e.g., 1000x1000 sites or larger. Given the large number of epithelial cells in a small area, realistic simulations to determine statistical properties of diversity measures during carcinogenesis may be completely infeasible. Our results here provide analytical or rapidly computable expressions that enable a detailed assessment of how these heterogeneity measures vary depending on time and depending on tissue/genetic parameters such as mutation rate and selective advantage conferred by the genetic alteration. These tools can be utilized to study how tissue heterogeneity in premalignant conditions varies between sites and tissue types, and thus guide sampling or biopsy procedures across various cancer types.

Chapter 3

Mutant clone propagation and field cancerization in epithelial basal zones

3.1 Introduction

Recall that cancer initiation is typically characterized by the accumulation of successive genetic alterations that can confer fitness advantages due to increased proliferation capabilities or avoidance of apoptosis signals. The fitness advantages in the affected cells lead to the expansion of premalignant clones. These clones are more susceptible than normal tissue to subsequent mutations and eventually tumor initiation. However, the clones are difficult to detect clinically, so mathematical models can provide useful insight regarding the temporal dynamics and spread of premalignant tissue.

There have been number of previous papers that use mathematical models to study the evolutionary process of cancer initiation. One model that has been used to study carcinogenesis is known as a Moran process, first proposed by Moran in [47], which is a nonspatial evolutionary model that assumes a well-mixed population. In [48], Nowak et al. studied the waiting time until a cell in the Moran process acquires two mutations, signifying cancer initiation. Durrett, Schmidt, and Schweinsberg generalized this framework to a Moran process, in which cancer initiation occurs after a cell has mutated k times, and they analyzed the time until this event [49, 50]. Additionally, Komarova et al. introduced the concept of stochastic tunneling between two homogeneous states in [51], and Iwasa et al. used the Moran process to analyze the inactivation dynamics of tumor suppressor genes

in [52, 53].

In this chapter, we use a spatially structured model to approximate premalignant clones in epithelial tissue, i.e. the tissue lining the exterior and interior surfaces of the body. Tumors arising from epithelial tissue are known as carcinoma, and they account for more than 80% of human cancer cases [54]. Stratified epithelial tissue is made up of layers of cells, and the deepest layer is known as the basal layer. The basal layer contains the stem cells of the epithelium, so these cells proliferate, and the differentiated cells are pushed upward through the other layers of the epithelial tissue. Hence, the genetic alterations that eventually lead to cancer initiation arise within the basal layer.

In [55], Durrett, Foo, and Leder analyze the propagation speed of a premalignant clone on a two-dimensional lattice. The two-dimensional lattice is a good approximation of a single basal layer. However, some epithelial tissue has more than one layer of proliferating cells, e.g. glabrous skin or hyperproliferative epidermis. To account for these cases, we refer to the set of layers containing proliferating cells as the "basal zone." We provide a more detailed description of the types of epithelial tissue with multi-layered basal zones in Section 3.5. This paper focuses on the clonal propagation speed in epithelial basal zones that are more than one cell thick. We represent such a basal zone as a set of stacked 2-dimensional lattices, and we modify the techniques in [55] to determine the asymptotic speed as a function of the width of the basal zone.

We use the propagation speed to investigate how various temporal and spatial properties of premalignant tissue change when the basal zone is more than one cell thick. Many of these properties were analyzed in [56], and they describe a phenomenon known as 'field cancerization' or the 'cancer field effect.' This field effect describes the observation that some regions surrounding tumors seem to have an increased risk for the development of recurrent tumors. These regions are known as premalignant fields, and recent studies suggest that they are caused by genetic alterations that lead to faster cell growth, making the cells more susceptible to subsequent mutation [13, 14]. Curtius and colleagues review the biological mechanisms driving field cancerization and the resulting clinical implications in [15]. These fields are represented by the premalignant clone regions in our model, so after determining the clonal propagation speed, we use it to compare various properties of premalignant fields in basal zones that are multiple cells thick to those in tissue containing a single basal layer.

We approximate a premalignant clone using a spatial model of clonal expansion known as the biased voter model, introduced by Williams and Bjerknes in [57]. Later, Bramson and Griffeath analyzed this model and proved a shape theorem that describes the asymptotic behavior of the process in [39, 40]. In this paper, we will discuss the biased voter model

and a generalized model, which is a spatial version of the Moran process. This generalized model incorporates mutation, and it has been studied previously in several different papers. Among these, in [58], Komarova analyzed the time until a cell develops two mutations in a one dimensional spatial Moran model. Durrett and Moseley extend this work and analyze the time until a cell has acquired two mutations in d dimensions [59].

In section 3.2 we present the asymptotic propagation speed of a premalignant clone in a basal zone that is w cells thick. We also include a local central limit theorem and random walk return time result in section 3.2, which we use within the proof of our main result. Then in section 3.3 we provide precise descriptions of the relevant particle systems used in the proof of the speed, and we include a convergence proof that allows us to approximate the dual process to the biased voter model with a branching Brownian motion. We present upper and lower bounds for the propagation speed and use these bounds to prove the main result in section 3.4. Next in section 3.5, we describe an application of the propagation speed within the context of field cancerization, in which we compare the size of the premalignant fields in basal zones of differing thickness. Finally, we discuss our findings and future goals in section 3.6.

3.2 Main result

In this section, we present the asymptotic propagation speed of a mutant clone in an epithelial basal layer that is w cells thick. First, we establish the relevant notation and make the notion of propagation speed precise.

Let $\mathbb{Z}_w = \mathbb{Z}/w\mathbb{Z}$. Suppose that each site in $\mathbb{Z}^2 \times \mathbb{Z}_w$ is occupied either by a type-0 cell or type-1 cell. Type-1 cells, representing mutant cells, have a relative fitness advantage $\beta > 0$ over type-0 cells, which represent normal cells. Each type-0 and type-1 cell waits an exponentially distributed amount of time with parameter 1 and $1 + \beta$ (with $\beta > 0$), respectively, before splitting into two cells of its type. Then it randomly chooses one of its nearest neighbors to replace with its daughter cell.

Let ξ_t denote the set of sites in occupied by type-1 cells $\mathbb{Z}^2 \times \mathbb{Z}_w$ at time t. The process $\{\xi_t : t \geq 0\}$ is known as the biased voter model (BVM) on $\mathbb{Z}^2 \times \mathbb{Z}_w$. We will on occasion want to make the fitness advantage of the biased voter model explicit, in these cases we will use the notation $\{\xi_{t,\beta} : t \geq 0\}$. In [39, 40], Bramson and Griffeath showed that ξ_t, conditioned on the event that it does not die out, eventually has a convex, symmetric shape

D, which is a unit ball in an unknown norm. More precisely Theorem 2 of [39] states that for any $\epsilon > 0$,

$$\mathbb{P}\left(\exists t_* < \infty : (1-\epsilon)tD \cap (\mathbb{Z}^2 \times \mathbb{Z}_w) \subset \xi_t \subset (1+\epsilon)tD, \quad t \geq t_* | \tau_\emptyset = \infty\right) = 1,$$

where $\tau_\emptyset = \inf\{t > 0 : \xi_t = \emptyset\}$. Note that [39] proved their result on $\mathbb{Z}^d$ for $d > 1$, but theIr result can be generalized to $\mathbb{Z}^2 \times \mathbb{Z}_w$. We provide a description of this generalization in Section B.5.

Our primary mathematical result is to better understand the shape D, and how this shape depends on the carcinogenic advantage β and the tissue geometry w. In pursuit of this define

$$[-c_w(\beta)e_1,\ c_w(\beta)e_1] = \{x \in \mathbb{R} : (x,0,0) \in D\}, \tag{3.1}$$

where $e_1 = (1,0,0)$. Our main result describes the small β behavior of the asymptotic growth rate $c_w(\beta)$.

Theorem 3.2.1. *Let ξ_t be a biased voter model on $\mathbb{Z}^2 \times \mathbb{Z}_w$ with fitness advantage $\beta > 0$. Let $c_w(\beta)$ be the growth rate of ξ_t, conditioned on the event that it does not die out, as described by 3.1. As $\beta \to 0$,*

$$c_w(\beta) \sim \begin{cases} \frac{4}{5}\sqrt{\frac{\pi w \beta}{\log(1/\beta)}} & w = 2 \\ \\ \frac{2}{3}\sqrt{\frac{\pi w \beta}{\log(1/\beta)}} & w > 2. \end{cases}$$

The proof of Theorem 3.2.1 is provided in section 3.4 and it follows from upper and lower bounds on the speed, whose proofs are included in the Appendix. The proof of the propagation speed relies on the duality between the biased voter model and a system of branching coalescing random walks with migration rate 1 and branching rate β. During a migration event a particle jumps to one of the $2d$ nearest neighbors chosen at random, also during a branching event the resulting daughter particle is placed at random on one of the $2d$ nearest neighboring sites. Anytime two particles simultaneously occupy the same lattice location they coalesce into a single particle. This dual process is constructed using a graphical representation of the BVM, in which the dual reverses time and traces the lineages of the particles in ξ_t. See [60] or [8] for a description of this construction.

Let $\tilde{\zeta}_t \subset \mathbb{Z}^2 \times \mathbb{Z}_w$ denote the dual process of the BVM ξ_t. By construction, $\tilde{\zeta}_t$ and ξ_t

satisfy the following duality relation,

$$\mathbb{P}(\xi_t^A \cap B \neq \emptyset) = \mathbb{P}(\tilde{\zeta}_t^B \cap A \neq \emptyset), \tag{3.2}$$

where the notation $\xi_t^A, \tilde{\zeta}_t^B$ indicates that $\xi_0 = A$, and $\tilde{\zeta}_0 = B$.

Since each particle in the dual process $\tilde{\zeta}_t$ behaves as a random walk, we proved the following local central limit theorem (LCLT) on $\mathbb{Z}^d \times \mathbb{Z}_w$ to describe the asymptotic behavior of each particle.

Theorem 3.2.2 (Local Central Limit Theorem on $\mathbb{Z}^d \times \mathbb{Z}_w$)**.** *Let $(Y_t)_{t\geq 0}$ be a simple symmetric random walk on $\mathbb{Z}^d \times \mathbb{Z}_w$ with jump rate α. Then if $w = 2$,*

$$\lim_{t\to\infty} (\alpha t)^{d/2}\, \mathbb{P}(Y_t = \boldsymbol{x}) = w^{-1} \left(\frac{2d+1}{4\pi}\right)^{d/2}, \quad \forall \boldsymbol{x} \in \mathbb{Z}^d \times \mathbb{Z}_w.$$

and if $w > 2$,

$$\lim_{t\to\infty} (\alpha t)^{d/2}\, \mathbb{P}(Y_t = \boldsymbol{x}) = w^{-1} \left(\frac{d+1}{2\pi}\right)^{d/2}, \quad \forall \boldsymbol{x} \in \mathbb{Z}^d \times \mathbb{Z}_w.$$

We include the proof of this theorem in Appendix section B.1.

In order to prove an upper bound for the speed, we consider a pruned dual process, which does not include the particles that coalesce with their parents shortly after birth. Then we approximate this pruned dual process with a branching random walk, which excludes every particle that quickly coalesces with its parent but does not allow for any other coalescence in the process. Since coalescence slows down the process, the propagation speed of the branching random walk provides an upper bound for the speed of the dual process.

We use the LCLT on $\mathbb{Z}^2 \times \mathbb{Z}_w$ to prove the following result, which describes the timing of a random walk's first return to the origin. This is used to determine the probability of coalescence between a parent and daughter particle before a specific time.

For the random walk $\{Y_t : t \geq 0\}$ starting at the origin define the time of its first jump by J_1 and the time of first return to the origin as

$$T_0 = \inf\{t > J_1 : Y_t = 0\}.$$

Then we have the following result for the asymptotics of the tail probability of T_0.

Theorem 3.2.3. *As $t \to \infty$,*

$$\mathbb{P}(T_0 > t) \sim \begin{cases} \frac{4\pi w}{5\log(\alpha t)} & w = 2 \\ \\ \frac{2\pi w}{3\log(\alpha t)} & w > 2. \end{cases}$$

The proof of this theorem is provided in Appendix section B.2.

We also use the branching random walk resulting from pruning the dual process to provide a lower bound on the speed of the dual process. This proof relies on the weak convergence of a scaled version of the branching random walk to a branching Brownian motion; we provide the proof of this convergence in the next section. We work with branching Brownian motion because it is possible to use a block construction argument to show that it dominates a percolation process, and known percolation results provide then provide the desired lower bound on the propagation speed.

Throughout this paper we will use the following notation. We will denote a vector in $\mathbb{Z}^2 \times \mathbb{Z}_w$ by $z = (z_1, z_2, z_3)$. Also we will denote the ith unit vector by e_i. In addition we will use the following Landau notation for non-negative functions as $x \to \infty$: $f(x) = O(g(x))$ if $\limsup_{x\to\infty} f(x)/g(x) < \infty$, and $f(x) = o(g(x))$ if $\lim_{x\to\infty} f(x)/g(x) = 0$.

3.3 Approximating the dual process

In this section, we define notation to describe the dual process, branching Brownian motion, and a few modified processes that we use to determine the propagation speed of the BVM.

3.3.1 Dual process

Recall that the dual process $\tilde{\zeta}_t$ is a system of branching coalescing random walks. Each particle in $\tilde{\zeta}_t$ has jump rate 1, and each gives birth to a new particle at rate β. The new offspring is placed at a randomly chosen neighboring site.

We represent each particle with a path $(\tilde{Z}_t^i)_{t\in[0,T]}$, in which $\tilde{Z}_t^i \in \mathbb{Z}^2 \times \mathbb{Z}_w$ is the location of the i-th particle at time t. Let $\tilde{Z}$ be the product of these paths,

$$\tilde{Z} = \left(\tilde{Z}^1, \tilde{Z}^2, \tilde{Z}^3, \ldots\right),$$

where

$$\tilde{Z}_t = \left(\tilde{Z}_t^1, \tilde{Z}_t^2, \ldots, \tilde{Z}_t^N, \infty, \ldots\right),$$

and we use the symbol ∞ as a place holder for particles that have not been created yet. For all $i \in \mathbb{Z}^+$,

$$\tilde{Z}^i : [0,T] \to (\mathbb{Z}^2 \times \mathbb{Z}_w) \cup \{\infty\}.$$

Note that

$$\tilde{\zeta}_t = \left\{\tilde{Z}_t^i : \tilde{Z}_t^i < \infty\right\}.$$

For any set $A \in \mathbb{Z}^2 \times \mathbb{Z}_w$, let

$$\tilde{\zeta}_t(A) := \left\{\tilde{Z}_t^i \in A\right\}.$$

For particles i and j define their coalescence time as

$$\sigma_{ij} = \inf\{t > 0 : \tilde{Z}_t^i = \tilde{Z}_t^j, \tilde{Z}_t^i \neq \infty\}.$$

If $i < j$ then for $t > \sigma_{ij}$ we set $\tilde{Z}_t^j = \infty$. In other words, the path of the particle with the higher index changes to ∞ after time t, and the path of the lower index particle continues to follow the system dynamics.

Let K be the number of initial particles in $\tilde{\zeta}_t$, so $K = |\tilde{\zeta}_0|$. For all $i > K$, let $p_t(i)$ be the index of the parent of the particle with path $\{\tilde{Z}_s^i : 0 \leq s \leq t\}$. We will say $p_t(i) = 0$ if $\tilde{Z}_t^i \equiv \infty$, and $p_t(i) = i$ for $i \leq K$ and all $t \geq 0$. In addition define the nth iterate of p by $p^{(n)}$.

Let $\tilde{C}_t(i)$ be the set of children of particle $\tilde{Z}^i$ by time t, so

$$\tilde{C}_t(i) = \{\tilde{Z}^j : p_t(j) = i\}.$$

Let $\tilde{D}_t(i)$ be the set of all descendants of particle $\tilde{Z}^i$ by time t, so

$$\tilde{D}_t(i) = \{\tilde{Z}^j : \exists n \text{ s.t. } p_t^{(n)}(j) = i\}.$$

Define $\tilde{b}_j$ as the branching time of the particle with path $\tilde{Z}^j$. We let $\tilde{b}_i = 0$ for $i \leq K$, since the the first K particles are included in the process at time 0.

3.3.2 Pruned dual process

Next, we introduce a pruned dual process $\hat{\zeta}_t$, which eliminates the particles that coalesce with another particle shortly after birth.

Analogously to $\tilde{Z}_t$, we can represent the pruned dual with a product of paths $\hat{Z}^i = (\hat{Z}^i_t)_{t\in[0,T]}$, such that

$$\hat{Z} = \left(\hat{Z}^1, \hat{Z}^2, \hat{Z}^3, \ldots\right),$$

and

$$\hat{\zeta}_t = \left\{\hat{Z}^i_t : \hat{Z}^i \in \hat{Z}, \hat{Z}^i_t < \infty\right\}.$$

Let

$$\tau(\beta) := \frac{1}{\beta\sqrt{\log(\frac{1}{\beta})}},$$

where $\tau(\beta)$ is the waiting period after each branching event. A newborn particle in the dual process only joins the pruned dual if it does not coalesce with another particle by $\tau(\beta)$ time units after its birth. Also we do not allow particles to branch during the initial $\tau(\beta)$ time units after its creation.

Note that the path of a new particle that coalesces with a non-parent before time $\tau(\beta)$ will not be a part of the process at time $\tau(\beta)$, but to serve as a placeholder, its path $\hat{Z}^i \in \hat{Z}$, and $\hat{Z}^i \equiv \infty$.

After a particle is added to the process ($\tau(\beta)$ time units after its birth), we assume that branching and coalescence occurs exactly as it does in the dual $\tilde{\zeta}_t$.

Note that for $A, B \in \mathbb{Z}^2 \times \mathbb{Z}_w$,

$$\mathbb{P}\left(\tilde{\zeta}^B_t \cap A \neq \emptyset\right) \geq \mathbb{P}\left(\hat{\zeta}^B_t \cap A \neq \emptyset\right).$$

3.3.3 Branching random walk

We also introduce a branching random walk ζ_t with jump rate 1 and branching rate β, which we will use to approximate the pruned dual $\hat{\zeta}_t$.

Similarly to the dual and pruned dual processes, we represent the location of each particle in the BRW with a path $(Z^i_t)_{t\in[0,T]}$, and

$$Z = \left(Z^1, Z^2, Z^3, \ldots\right)$$

$$\zeta_t = \left\{Z^i_t : Z^i \in Z, Z^i_t < \infty\right\}.$$

Most of the coalescence events in $\tilde{\zeta}_t$ occur between parent and daughter particles shortly after the daughter particle is born. Thus in ζ_t, we exclude every particle that coalesces with its parent before time $\tau(\beta)$.

Recall $(\tilde{b}_j)_{j\in\mathbb{Z}^+}$ are the branching times in $\tilde{\zeta}_t$, and let $(\hat{b}_j)_{j\in\mathbb{Z}^+}$ and $(b_j)_{j\in\mathbb{Z}^+}$ be the branching times in $\hat{\zeta}_t$ and ζ_t, respectively. For $i > K$, if

$$\tilde{Z}_t^i = \tilde{Z}_t^{p_i} \text{ for some } t \in \left[\tilde{b}_i, \tilde{b}_i + \tau(\beta)\right),$$

then $\tilde{Z}^i \notin Z$ and $\tilde{b}_i \notin (\hat{b}_j)_{j\in\mathbb{Z}^+} \cup (b_j)_{j\in\mathbb{Z}^+}$. However, if

$$\tilde{Z}_t^i \neq \tilde{Z}_t^{p_i} \text{ for all } t \in \left[\tilde{b}_i, \tilde{b}_i + \tau(\beta)\right),$$

then this is considered a successful branching event in ζ_t and $\hat{\zeta}_t$; therefore,

$$(\tilde{b}_i + \tau(\beta)) \in (\hat{b}_j)_{j\in\mathbb{Z}^+} \cap (b_j)_{j\in\mathbb{Z}^+}.$$

Suppose $\tilde{b}_i + \tau(\beta) = b_n$, i.e. the branching time for the n-th new particle in ζ_t. Then

$$Z_t^n = \infty \text{ for all } t < \tilde{b}_i + \tau(\beta), \text{ and } Z^n_{\tilde{b}_i+\tau(\beta)} = \tilde{Z}^i_{\tilde{b}_i+\tau(\beta)}.$$

Otherwise, since ζ_t is a branching random walk, we assume that particles in ζ_t do not coalesce when they meet.

Similarly to $\hat{\zeta}_t$, we let

$$C_t(i) = \{Z_t^j : p_t(j) = i\}$$

and

$$D_t(i) = \{Z^j : \exists n \text{ s.t. } p_t^{(n)}(j) = i\}.$$

3.3.4 Scaled process

We define a scaling factor $h(\beta)$ in order to obtain weak convergence to branching Brownian motion as $\beta \to 0$. Let

$$h(\beta) = \frac{1}{\beta}\log(1/\beta).$$

Let ζ_t^β be the scaled version of ζ_t, in which we scale space so that the particles live on

$(\mathbb{Z}^2 \times \mathbb{Z}_w)/\sqrt{h(\beta)}$, and we run time at rate $h(\beta)$. Hence for any $A \subset (\mathbb{Z}^2 \times \mathbb{Z}_w)/\sqrt{h(\beta)}$,

$$\begin{aligned}
|\zeta_t^\beta(A)| &= \sum_{Z_t^i \in Z_t} \mathbb{1}_{\left\{\frac{Z_t^i(h(\beta)t)}{\sqrt{h(\beta)}} \in A\right\}} \\
&= \sum_{Z_t^i \in Z_t} \mathbb{1}_{\left\{Z_t^i(h(\beta)t) \in \sqrt{h(\beta)}A\right\}} \\
&= \left|\zeta_{h(\beta)t}\left(\sqrt{h(\beta)}\ A\right)\right|.
\end{aligned}$$

Similarly to the unscaled BRW, we represent ζ_t^β as a product of paths,

$$Z^\beta = \left(Z^{1,\beta}, Z^{2,\beta}, Z^{3,\beta}, \ldots\right).$$

Let $(b_j^\beta)_{j\in\mathbb{Z}^+}$ be the branching times in ζ_t^β, where each $b_j^\beta = h(\beta)b_j$. As before we denote the children and offspring of particle i at time t by $C_t^\beta(i)$ and $D_t^\beta(j)$ respectively.

3.3.5 Branching Brownian motion

Let $\mathbb{R}_w := \mathbb{R} \mod w$. Let χ_t be a branching Brownian motion on $\mathbb{R}^2 \times \mathbb{R}_w$ with branching rate $\frac{4\pi w}{5}$ if $w = 2$ or branching rate $\frac{2\pi w}{3}$ otherwise.

More specifically, we assume that χ_t starts with K initial particles, each of which moves independently as standard Brownian motion. Each particle waits an exponentially distributed amount of time with parameter $\frac{4\pi w}{5}$ if $w = 2$ or parameter $\frac{2\pi w}{3}$ otherwise before dividing to form two identical particles. Both particles continue along independent Brownian paths starting at the same location, with the same branching rate.

We represent χ_t as a product of continuous paths $Y^i = (Y_t^i)_{t\in[0,T]}$, such that

$$Y_t = \left(Y^1, Y^2, Y^2, \ldots\right).$$

Let $(\tau_j)_{j\in\mathbb{Z}^+}$ be the branching times of χ_t.

3.3.6 Modified Skorokhod topology

Now we describe the space in which $\tilde{Z}, \hat{Z}_t, Z_t$, and Y_t live. Let $\mathbf{D}([0,T], \mathbb{R}^2 \times \mathbb{R}_w)$ be the space of Cadlag paths. For simplicity, we will denote this space by $\mathbf{D}_T$.

Let $\hat{\mathbf{D}}_T$ be a modification of $\mathbf{D}_T$ so that paths can take on the value ∞. The ∞ acts as a placeholder for the time before a particle is born. $\hat{Z}^i, Z^i, Y^i \in \hat{\mathbf{D}}_T$ for all $i \in \mathbb{Z}^+$.

We must make some adjustments to the standard Skorokhod metric d_T in order to

account for the fact that paths can take on the value ∞, so we define

$$|(f-g)(t)| = \begin{cases} 0 & \text{if } f(t)=\infty, g(t)=\infty \\ \infty & \text{if } f(t)=\infty, g(t)\neq\infty \\ \infty & \text{if } f(t)\neq\infty, g(t)=\infty \end{cases}$$

If $f(t) \neq \infty$ and $g(t) \neq \infty$, then $|(f-g)(t)|$ is the Euclidean norm, modified for the periodic boundary on $\mathbb{R}_w$. In particular, if $f(t) = (p_1, p_2, p_3)$ and $g(t) = (q_1, q_2, q_3)$, then

$$|(f-g)(t)| = \sqrt{(q_1-p_1)^2 + (q_2-p_2)^2 + (\min\{q_3-p_3, w-(q_3-p_3)\})^2}.$$

Otherwise, d_T is defined as usual, and it induces the Skorokhod topology on the modified space $\hat{\mathbf{D}}_T$ of Cadlag paths.

Let $\mathcal{D}_T$ be the space consisting of infinite products of Cadlag paths in $\hat{\mathbf{D}}_T$, as shown below,

$$\mathcal{D}_T = \{(x_1, x_2, \ldots) : \exists k_0 \ \ s.t. \ \ \forall k \leq k_0, x_k \in \hat{\mathbf{D}}_T, \ \ \forall k > k_0, x_k = \infty\}.$$

Note that $\tilde{Z}, \hat{Z}, Z \in \mathcal{D}_{h(\beta)T}$, while Z^β and Y take values in $\mathcal{D}_T$.

We use the following metric to measure the distance between any $X, Y \in \mathcal{D}_T$:

$$d_{p,T}(X,Y) = \sup_i \{d_T(x_i, y_i)\},$$

where $X = (x_1, x_2, \ldots)$ and $Y = (y_1, y_2, \ldots)$.

3.3.7 Convergence of BRW to BBM

Next, we show that the scaled branching random walk ζ_t^β converges weakly to the branching Brownian motion χ_t. We will use this in the proof of the lower bound for the speed of spread of the BVM.

Lemma 3.3.1. *Let χ_t be a branching Brownian motion with infinitesimal variance $\sigma^2 = \frac{1}{3}$ and branching rate $\frac{4\pi w}{5}$ if $w = 2$ or branching rate $\frac{2\pi w}{3}$ otherwise.*

Assume initial conditions $Y_0 = (x_1, x_2, \ldots, x_K, \infty, \ldots)$ and $Z_0^\beta = (x_1^\beta, x_2^\beta, \ldots, x_K^\beta, \infty, \ldots)$, where $x_1, x_2, \ldots, x_K \in \mathbb{R}^2 \times \mathbb{R}_w$ and $x_i^\beta = \dfrac{1}{\sqrt{h(\beta)}} \left[\sqrt{h(\beta)}\, x_i\right]$. Then as $\beta \to 0$,

$$Z^\beta \Longrightarrow Y,$$

where the weak convergence is with respect to the metric $d_{p,T}$.

Proof. We first identify the limit of the branching rate of ζ_t^β. We will show that as $\beta \to 0$, $\mu_\beta \to \frac{4\pi w}{5}$ when $w = 2$ and $\mu_\beta \to \frac{2\pi w}{3}$ when $w > 2$.

The branching rate of the scaled dual process $\hat{\zeta}_t^\beta$ is $\beta h(\beta)$. In our branching random walk, we only count branching events where particles do not collide with their parent for their first $\tau(\beta)$ time units. We can use Theorem 3.2.3 to find the rate of branching events in ζ_t^β. In our particular case, let X^1 and X^2 be independent simple random walks on $\mathbb{Z}^d \times \mathbb{Z}_w$ with jump rate 1, representing a parent and daughter particle. The initial conditions are as follows: $X_0^1 = 0$, and X_0^2 is one of the nearest neighbors of 0, each with probability $\frac{1}{6}$. Define their difference as

$$\bar{X}_t = X_t^1 - X_t^2,$$

which is a simple random walk with jump rate 2. Let

$$T_0 = \inf\{t > 0 : X_t^1 = X_t^2\} = \inf\{t > 0 : \bar{X}_t = 0\}.$$

A successful branching event occurs when $T_0 > \tau(\beta)$. Then by Theorem 3.2.3, if $w > 2$,

$$\begin{aligned}
\mu_\beta &= \beta h(\beta) \mathbb{P}\left(T_0 > \tau(\beta)\right) \\
&\sim \beta h(\beta) \cdot \frac{2\pi w}{3 \log\left(2\tau(\beta)\right)} \\
&= \log(1/\beta) \cdot \frac{2\pi w}{3 \log\left(\dfrac{2}{\beta\sqrt{\log(1/\beta)}}\right)} \\
&= \frac{2\pi w \log(1/\beta)}{3\log(1/\beta) + 3\log 2 - \frac{3}{2}\log(\log(1/\beta))} \\
&\sim \frac{2\pi w}{3} \quad \text{as } \beta \to 0.
\end{aligned} \tag{3.3}$$

Analogously, if $w = 2$, then $\mu_\beta \sim \frac{4\pi w}{5}$ as $\beta \to 0$.

Based on this result we know that $|\zeta^\beta|$ is dominated by branching process with birth rate $2\pi w/3$ (in the case in which $w > 2$), and in particular

$$E\left[|\zeta_T^\beta|\right] \leq K e^{2\pi w T/3}.$$

Therefore given $\varepsilon > 0$ we can find an $M_\varepsilon < \infty$ such that $P\left(|\zeta_T^\beta| > M_\varepsilon\right) < \varepsilon$. We can thus restrict our analysis to the set where $|\zeta_T^\beta| \leq M_\varepsilon$. Therefore to establish weak convergence we only need to show the following three things: i) the rate of branching converges to the

desired limit, ii) the displacement of offspring from parent converges to zero, and iii) the paths of particles converge to Brownian motion. We have already established item (i) in (3.3), so we will next establish item (ii).

We will show that as $\beta \to 0$, the displacement between any parent and daughter particle $Z^i, Z^j \in \zeta_t$ at the time of branching converges to 0 in probability. Define

$$U = \beta^{-1/2}(\log(1/\beta))^{1/3}.$$

In the unscaled BRW ζ_t, (3.2.3) implies that as $\beta \to 0$, if $w > 2$,

$$\begin{aligned}
\mathbb{P}\left(T_0 > \tau(\beta)\right) &\sim \frac{2\pi w}{3\log\left(2\tau(\beta)\right)} \\
&= \frac{2\pi w}{3\log\left(\dfrac{2}{\beta\sqrt{\log(1/\beta)}}\right)} \\
&\sim \frac{2\pi w}{3\log(1/\beta)} \quad \text{as } \beta \to 0. \qquad (3.4)
\end{aligned}$$

Then by (3.4), as $\beta \to 0$,

$$\begin{aligned}
\mathbb{P}\left(|X^1_{\tau(\beta)} - X^2_{\tau(\beta)}| \geq U \;\middle|\; T_0 > \tau(\beta)\right) &\leq \frac{1}{\mathbb{P}(T_0 > \tau(\beta))}\mathbb{P}\left(|X^1_{\tau(\beta)} - X^2_{\tau(\beta)}| \geq U\right) \\
&\leq \frac{3\log(1/\beta)}{2\pi w}\left(\beta^{-1/2}(\log(1/\beta))^{1/3}\right)^{-2}\tau(\beta) \\
&= \frac{3\log(1/\beta)}{2\pi w} \cdot \frac{\beta}{\log(1/\beta))^{2/3}} \cdot \frac{1}{\beta\sqrt{\log(\frac{1}{\beta})}} \\
&= \frac{3}{2\pi w \log(1/\beta))^{1/6}}, \qquad (3.5)
\end{aligned}$$

where the second inequality follows from Chebyshev's inequality.

Note that $\frac{|X^1_{\tau(\beta)} - X^2_{\tau(\beta)}|}{\sqrt{h(\beta)}}$ is the corresponding distance on $(\mathbb{Z}^2 \times \mathbb{Z}_w)/\sqrt{h(\beta)}$. Therefore,

$$\begin{aligned}
\lim_{\beta\to 0}\mathbb{P}\left(|X^1_{\tau(\beta)} - X^2_{\tau(\beta)}| \geq U \;\middle|\; T_0 > \tau(\beta)\right) &= \lim_{\beta\to 0}\mathbb{P}\left(\frac{|X^1_{\tau(\beta)} - X^2_{\tau(\beta)}|}{\sqrt{h(\beta)}} \geq \frac{U}{\sqrt{h(\beta)}}\middle| T_0 > \tau(\beta)\right) \\
&= 0,
\end{aligned}$$

and

$$\lim_{\beta\to 0}\frac{U}{\sqrt{h(\beta)}} = \lim_{\beta\to 0}\frac{(\log(1/\beta))^{1/3}}{\sqrt{\beta}}\cdot\sqrt{\frac{\beta}{\log(1/\beta)}}$$
$$= \lim_{\beta\to 0}\frac{1}{(\log(1/\beta))^{1/6}} = 0.$$

Similarly, if $w = 2$, then $\mathbb{P}\left(T_0 > \tau(\beta)\right) \sim \dfrac{4\pi w}{5\log(1/\beta)}$, and the previous argument holds in this case as well.

Hence, for all w and for parent and daughter particles $Z^{1,\beta}, Z^{2,\beta} \in Z^\beta$ that split at time b_i,

$$|Z^{i,\beta}_{b_i} - Z^{j,\beta}_{b_i}| \Longrightarrow 0. \tag{3.6}$$

We next establish that the paths of Z^β converge to Brownian motions as $\beta \to 0$. Let $R_i = [b_i, T)$. Then let $Z^{i,\beta}|_{R_i}$ be the restriction of $Z^{i,\beta}$ to the domain R_i. In other words,

$$Z^{i,\beta}|_{R_i} : R_i \to (\mathbb{Z}^2 \times \mathbb{Z}_w)/\sqrt{h(\beta)},$$

and for $t \in R_i$,

$$Z^{i,\beta}|_{R_i}(t) = Z^{i,\beta}_t.$$

For $i \le K$, $Z^{i,\beta}|_{R_i} \equiv Z^{i,\beta}$. For each i, $Z^{i,\beta}|_{R_i}$ is a random walk, who steps have mean 0 and covariance matrix $(1/3)I_3$, where I_3 is the 3×3 identity matrix. Therefore, by Donsker's Functional CLT on $\mathbf{D}$ (Theorem 14.1 in [61])

$$Z^{i,\beta}|_{R_i} \Longrightarrow W|_{R_i} \ \ \forall i \text{ as } \beta \to 0, \tag{3.7}$$

where W is 3 dimensional Brownian motion with zero drift and covariance matrix $(1/3)I_3$.

One remaining technicality is that the path of a particle is not a simple random walk because we condition the particle on not hitting its parent. However, this conditioning is only carried out for the first $\tau(\beta)$ time units which is negligible on the time scale $h(\beta)$.

Combining (3.3), (3.6), and (3.7), we have shown that as $\beta \to 0$,

$$Z^\beta \Longrightarrow Y.$$

■

In Appendix B.3, we show that $\lim_{\beta\to 0}\mathbb{P}\left(d_p(Z^\beta,\hat{Z}^\beta)>\delta\right)=0$. Therefore, by Slutsky's Theorem, Lemmas 3.3.1 and B.3.2 imply that

$$\hat{Z}^\beta \Longrightarrow Y \text{ as } \beta\to 0.$$

3.4 Proof of the main result

In this section, we state a lower bound on the speed, whose proof we provide in the Appendix, and we claim an upper bound on the speed, whose proof is not yet complete. Then assuming this upper bound, we prove the main result.

Our goal will be to show that in the small β regime the speed of the BVM matches with the following expression

$$a(w,\beta)=\begin{cases}\frac{4}{5}\sqrt{\frac{\pi w\beta}{\log(1/\beta)}} & w=2\\ \frac{2}{3}\sqrt{\frac{\pi w\beta}{\log(1/\beta)}} & w>2.\end{cases} \tag{3.8}$$

In section 3.3.7, we proved that the BRW ζ_t^β converges weakly to the branching Brownian motion χ_t as $\beta\to 0$. We use this result in the proof of the lower bound for the propagation speed, which is stated below. We will show that for any $\varepsilon>0$ the BVM conditioned on survival will hit the point

$$x_\varepsilon(t;\beta)=\lfloor a(w,\beta)te_1(1-\varepsilon)\rfloor$$

at time t in the small β and large t limit.

Lemma 3.4.1. *Let $\{\xi_{t,\beta}:t\geq 0\}$ be a biased voter model on $\mathbb{Z}^2\times\mathbb{Z}_w$ with fitness advantage $\beta>0$ and define $\tau_\emptyset=\min\{t:\xi_t^0=\emptyset\}$, Then for any $\epsilon>0$*

$$\lim_{\beta\to 0}\lim_{t\to\infty}\mathbb{P}\left(\xi_{t,\beta}\cap\{x_\varepsilon(t;\beta)\}\neq\emptyset|\tau_\emptyset=\infty\right)=1.$$

The proof is provided in Appendix section B.4. In the proof, we use a block construction to renormalize the paths of the branching Brownian motion χ_t and show that the χ_t dominates a percolation process. Then we use the continuous mapping theorem and the convergence result 3.3.1 to extend the argument to the pruned dual process. Using the Brownian motion paths' normal distribution, we show that the BBM can fill up the first block in a given amount of time, and then our lower bound follows from a percolation result in [62].

The following claims that $a(w,\beta)$ is an upper bound on the speed of the BVM for all β. We will suppress the w,β and use the notation $a \equiv a(w,\beta)$. The proof of the upper bound is currently unfinished but involves the analysis of a process that dominates the dual process. In this dominating process, coalescence can only occur between a parent and daughter shortly after birth, provided that the daughter particle has not already produced any offspring.

Claim 3.4.2. *Let $\{\xi_t : t \geq 0\}$ be a biased voter model on $\mathbb{Z}^2 \times \mathbb{Z}_w$ with fitness advantage $\beta > 0$, and for $\varepsilon > 0$ and $t > 0$ define*

$$\mathcal{B}_\varepsilon(t) = \{z \in \mathbb{Z}^2 \times \mathbb{Z}_w : |z_1| > at(1+\varepsilon)\}.$$

Then

$$\lim_{t\to\infty} \mathbb{P}\left(\xi_t \cap \mathcal{B}_\varepsilon(t) \neq \emptyset\right) = 0.$$

Under the assumption that the upper bound holds, we combine upper and lower bounds from Claim 3.4.2 and Lemma B.4.1 to prove Theorem 3.2.1 on the speed of propagation.

Proof. We first fix $\varepsilon > 0$. The shape theorem by Bramson and Griffeath ([39, 40]) shows that there exists $c_w(\beta) < \infty$ such that

$$[-c_w(\beta)e_1,\ c_w(\beta)e_1] = \{x \in \mathbb{R} : (x,0,0) \in D\},$$

where D is the limiting shape of the biased voter model. This implies that for $z(t) = (z_1(t),0,0)$

$$\begin{aligned} &\text{if } |z_1(t)| \leq c_w(\beta)t(1-\varepsilon), \text{ then } \lim_{t\to\infty} \mathbb{P}(z(t) \in \xi_t | \tau_\emptyset = \infty) = 1 \\ &\text{if } |z_1(t)| \geq c_w(\beta)t(1+\varepsilon), \text{ then } \lim_{t\to\infty} \mathbb{P}(z(t) \in \xi_t | \tau_\emptyset = \infty) = 0. \end{aligned} \tag{3.9}$$

Assume there exists a positive sequence $\{\beta_n\}_{n\geq 1}$ converging to 0 such that

$$c_w(\beta_n)(1+\varepsilon) \leq a(w,\beta_n)(1-\varepsilon)$$

for all n. Then by (3.9) we get that

$$\lim_{t\to\infty} \mathbb{P}\left(x_\varepsilon(t;\beta_n) \in \xi_{t,\beta_n} | \tau_\emptyset = \infty\right) = 0 \quad \text{for all } n,$$

which clearly contradicts Lemma B.4.1. We thus conclude that

$$\liminf_{\beta\to 0} \frac{c_w(\beta)}{a(w,\beta)} \geq \frac{1-\varepsilon}{1+\varepsilon}.$$

Since $c_w(\beta)(1-\varepsilon) \geq a(w,\beta)(1+\varepsilon)$ leads to a contradiction between Lemma 3.4.2 and (3.9) we can conclude that for all $\beta > 0$

$$\frac{c_w(\beta)}{a(w,\beta)} \leq \frac{1+\varepsilon}{1-\varepsilon},$$

and since $\varepsilon > 0$ is arbitrary the result follows. ■

3.5 Application to cancer initiation in epithelial tissue

Epithelial basal zone thickness. In this section we discuss the biological application of the mutant clone propagation speed that we have analyzed in this paper. We use $\mathbb{Z}^2 \times \mathbb{Z}_w$ to approximate the layers of progenitor cells in stratified epithelial tissue, which we refer to as the "basal zone," as designated in [63]. In general, stratified epithelium consists of two or more layers of cells, and the layers provide protection for underlying tissue. The most common type of stratified epithelial tissue is known as stratified squamous epithelium, in which the cells toward the surface of the tissue are very flat. Stratified squamous epithelium can be classified further into two types: keratinized and nonkeratinized. The outer layer of keratinized stratified squamous epithelium is made up of dead cells that contain the protein keratin, which makes the surface waterproof. On the other hand, the outer layer of nonkeratinized stratified squamous epithelium contains moist, living cells. All outer layers of human skin are keratinized, and nonkeratinized stratified squamous epithelium is found in the esophagus, oral cavity, vagina, anus, cornea, and part of the pharynx.

The proliferating cells in epithelial tissue, composing the basal zone, typically express keratins K5 and K14, so this keratin expression can be used to determine the types of epithelial tissue whose basal zone is made up of multiple layers. Often K19 is also expressed in nonkeratinized basal cells [64]. Many of the tissue types with multi-layered basal zones fall within the branch of nonkeratinized stratified squamous epithelium. In [63], Geboes describes the basal zone in esophageal tissue, consisting of the basal layer and a few layers of cells above it, known as the suprabasal cell layers. The esophageal basal zone occupies at most 15-20% of the total epithelial thickness. Cells in both the deepest basal layer and the suprabasal cell layers express K14, K19, and an epidermal growth factor receptor (EGFR),

and the cells above this zone do not express the receptor and express keratins K1 and K10, rather than K14 and K19.

In addition to esophageal tissue, other thick epithelia like cheek and palate tissue have a basal zone 2-3 layers thick, expressing the keratins K5 and K14. In contrast, thin epithelia, such as the tissue on the floor of the mouth has one layer of proliferating cells [64]. The outer root sheath around a hair follicle can also have two layers of cells that express K5 and K14 and that proliferate to populate the rest of the epithelium around the hair follicle [65]. Typically the basal zone in keratinized epithelia, consists of a single layer of cells, but skin tissue that has been affected by hyperproliferative skin diseases, including psoriasis, papillomas, and benign epidermal tumors, have 2-3 layers of proliferating cells. With treatment, the epithelia that has been affected by these diseases can return to its baseline state, with one layer of progenitor cells [66].

Boundary condition comparison. Since the periodic boundary condition that we apply in the third dimension of our model is not biologically realistic, we use simulation to compare the speeds when using a periodic boundary condition versus a reflecting boundary condition. When the third dimension is equipped with a reflecting boundary condition, the cells on the top sheet can only place their progeny in one of the four neighboring sites on the same sheet or on the sheet directly below. Analogously cells on the lowest sheet can only place their progeny in the five nearest neighboring sites, resulting in a more biologically realistic growth process. Let $I_w = [0, w-1] \cap \mathbb{Z}$. In the reflecting case, $\mathbb{Z}^2 \times I_w$ forms the domain for the biased voter model. We simulated the BVM on $\mathbb{Z}^2 \times \mathbb{Z}_w$, i.e. with a periodic boundary condition in the third dimension, and on $\mathbb{Z}^2 \times I_w$, i.e. with a reflecting boundary condition in the third dimension. In order to compare the speeds in these cases, we simulated the process for width values $w = 2, 3, 4, 5$ and with fitness advantage $\beta = 0.01, 0.05, 0.1$. We ran at least 30 simulations for each set of parameters, and we recorded the propagation speed when the process reached $(100, 0, 0)$ or $(-100, 0, 0)$. We used this data to determine an average speed and 95% confidence interval for each set of parameters. Plots 3.1 and 3.2 display the mean speed and 95% CI under both boundary conditions, as functions of β and w, respectively.

Note that the periodic and reflecting boundary conditions are equivalent when $w = 2$ because each cell has exactly five nearest neighbors in this case. Thus with either boundary condition, a cell at site (x, y, z) can place its progeny at $(x \pm 1, y, z)$, $(x, y \pm 1, z)$, or $(x, y, (z + 1) \mod 2)$. We observe that when $w > 2$, equipping the model with a reflecting boundary condition in the third dimension results in a slightly smaller propagation speed than in the periodic case. Recall that we measure this speed using the time it takes to travel a certain

distance in the x-dimension on the first sheet. In the reflecting case, there are fewer cells placing their progeny on the outer sheets, since only one neighboring sheet can contribute daughter cells to each outer sheet. Hence, we expect a slightly slower propagation speed when measured on an outer sheet than on an inner sheet. Consequently, since we have tracked the simulation speed on the first sheet, this provides an explanation for the slower speed in the reflecting boundary case versus the periodic boundary case, in which two neighboring sheets can contribute daughter cells to the first sheet.

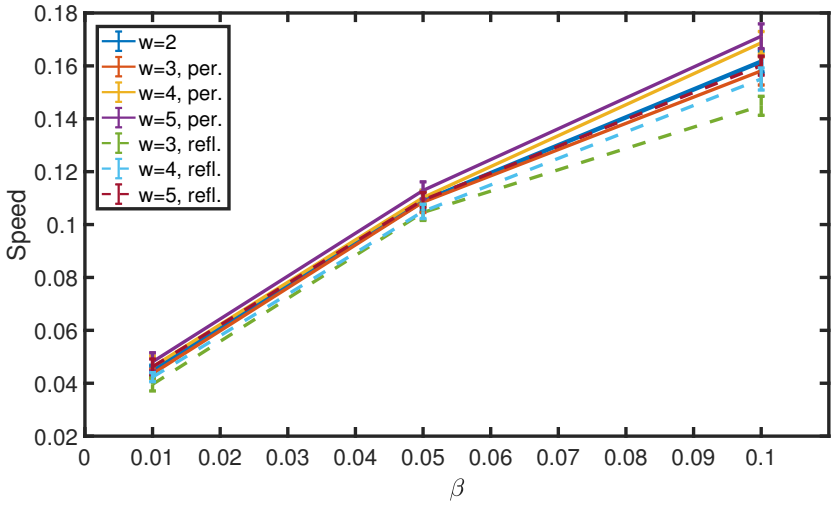


Figure 3.1: **BVM propagation speed, as a function of the fitness advantage β.** The speed is obtained using Monte Carlo simulation on w sheets, for $2 \leq w \leq 5$. This figure shows the average speeds when simulated with both periodic boundary conditions and reflecting boundary conditions, and the error bars indicate the 95% confidence interval for each speed.

As mentioned, the reflecting boundary condition induces different cell behavior on the inner and outer sheets, resulting in distinct propagation speeds on these two types of sheets. Note this is not the case in the periodic case because all cells have six nearest neighbors in that setting, whereas in the reflecting case, cells on the outer sheets have five nearest neighbors. Our simulations of the speed with a reflecting boundary condition measured on sheets 1,2, and 3, as shown in Figure 3.3, support the hypothesis that the propagation speed is larger on inner sheets than on outer sheets. We observe that, excluding the case when $w = 2$, the speed on sheet 2 is greater than the speed on sheet 1 for all β, since sheet 2 is an inner sheet. When $w = 2$, both sheet 1 and 2 are outer sheets, so the propagation speeds are about the same. Analogously, the speed on sheet 3 is similar to the speed on

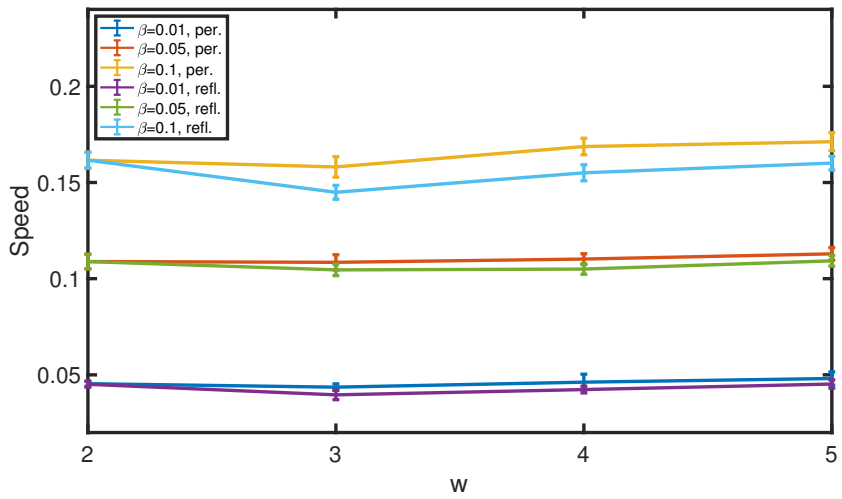
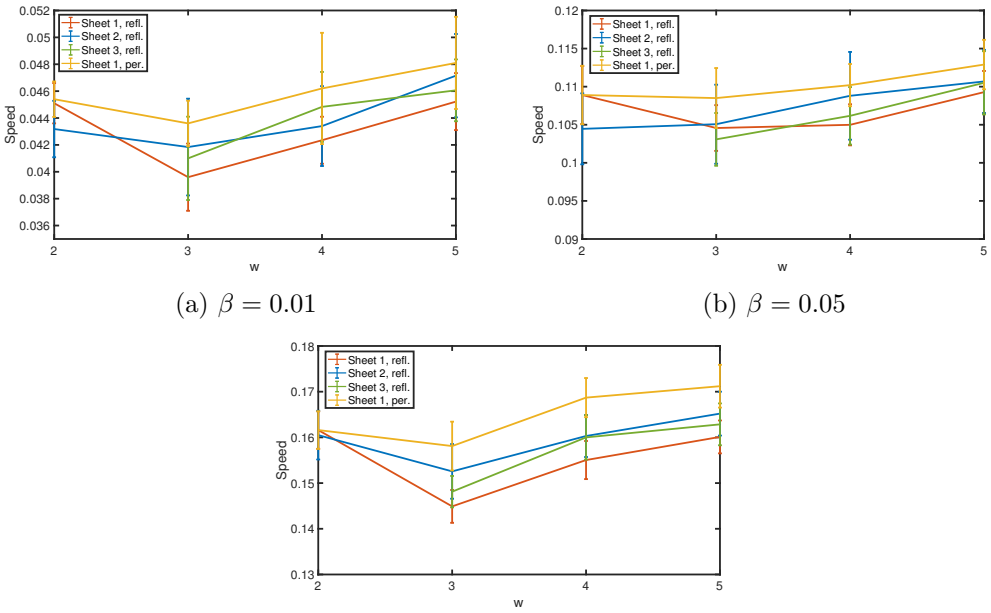


Figure 3.2: **BVM propagation speed, as a function of the number of sheets, w.** This figure shows the average speeds for the cases in which $\beta = 0.01, 0.05, 0.1$, with both periodic boundary conditions and reflecting boundary conditions. The error bars indicate the 95% confidence interval for each speed.

sheet 2 and larger than the speed on sheet 1, except when $w = 3$. In that case, sheet 3 is an outer sheet, so the speed on sheet 3 is similar to the speed on sheet 1 and smaller than the speed on sheet 2.

Notice that the speeds measured on all sheets in the reflecting case are slightly smaller than the speed in the periodic case. An intuitive explanation for the overall difference in simulation speed in the time frame that we have examined is that in the reflecting case, it takes longer for the type-1 population to fill in the gaps and converge to the asymptotic shape that propagates outward on each sheet. This behavior stems from the fact that the type-1 population has reduced ability to spread to new sites in the reflecting case, since cells on the lowest sheet cannot place cells on the highest sheet, and vice versa. In the periodic case, the type-1 population can fill in the gaps occupied by type-0 cells on all of the sheets more quickly, allowing fronts to emerge and spread outward slightly more quickly than in the reflecting case.

In general, the propagation speeds in the periodic and reflecting cases are similar, particularly for small β. These simulations suggest that the speed that we have determined using a periodic boundary condition provides a slight overestimate, but certainly a reasonable approximation, of the speed at which a mutant clone would realistically spread outward in epithelial tissue.



(a) $\beta = 0.01$

(b) $\beta = 0.05$

(c) $\beta = 0.1$

Figure 3.3: **BVM propagation speed comparison on various sheets.** The propagation speed is displayed as a function of w, in the periodic boundary case and on sheets 1,2, and 3 in the reflecting boundary case. The average speeds obtained from simulation are shown when (a) $\beta = 0.01$, (b) $\beta = 0.05$, and (c) $\beta = 0.1$. The error bars indicate the 95% confidence interval for each speed.

Field cancerization application. Next we consider the mutant propagation speed within the broader context of cancer initiation. In order to do this, we extend the biased voter model on $\mathbb{Z}^2 \times \mathbb{Z}_w$ to a generalized spatial Moran process η_t, which allows cells to acquire mutations and includes k cell types. In particular,

$$\eta_t : \mathbb{Z}^2 \times \mathbb{Z}_w \to \{1, 2, \ldots, k\},$$

where $\eta_t(x) = i$ indicates that site x is occupied by a type-i cell at time t.

This is known as a k-step model of cancer initiation, in which type-k cells represent malignant cancer cells. In this model, cancer initiation is defined to be the time of the birth of the first type-k cell that does not go extinct. In this work, we use a two-step model, in which type-0 cells are considered normal cells, type-1 cells represent premalignant mutant cells, and type-2 cells represent cancer cells [67]. In the model, type-i cells mutate to type-$(i+1)$ cells at rate u_{i+1}, and type-i cells reproduce at rate $(1+\beta)^i$.

We can approximate the spatial Moran model with a mesoscopic model, as described in [56], due to Bramson and Griffeath's shape theorem on $\mathbb{Z}^d$ and the extension of the shape theorem to $\mathbb{Z}^2 \times \mathbb{Z}_w$. In the mesoscopic approximation, mutant clones live on a spatial continuum and grow deterministically with rate $c_w(\beta)$. Mutations arise according to a Poisson process with rate $u_1 N$, where N is the total number of cells in the tissue. We consider a mutation successful if it does not eventually die out. Maruyama showed that mutations are successful with probability $\frac{\beta}{1+\beta}$ [68]. In [56], they use this mesoscopic model to analyze the field cancerization process. In this work, we use the speed derived in Section 3.4 within the context of this mesoscopic model. Hence, we approximate a successful mutant clone with a ball whose radius grows linearly with rate

$$c_w(\beta) \sim \begin{cases} \sqrt{\frac{\pi\beta}{\log(1/\beta)}} & w = 1 \\ \frac{4}{5}\sqrt{\frac{\pi w\beta}{\log(1/\beta)}} & w = 2 \\ \frac{2}{3}\sqrt{\frac{\pi w\beta}{\log(1/\beta)}} & w > 2 \end{cases},$$

where the case $w = 1$ corresponds to clonal expansion on $\mathbb{Z}^2$ and was analyzed in [55]. Note that the volume of w stacked discs of radius 1 in $\mathbb{Z}^2 \times \mathbb{Z}_w$ is $w\pi$.

We use this mesoscopic model to characterize properties of the premalignant field during

cancer initiation. These properties allow us to determine how the process of field cancerization differs in single-layered basal zone from a basal zone that is $w > 1$ cells thick. An important property is the size of the local field, which is the region of premalignant cells that gives rise to the first malignant cell.

The values of parameters β, u_1, u_2, and N depend on the cancer and tissue type. Durrett, Foo, and Leder show that the cancer initiation behavior can be classified using three different regimes. See [55] if you are interested in the explicit definition of these regimes in terms of the parameter values, but for the scope of this paper, we will simply provide a description of the mutant behavior within each of the regimes.

In Regime 1 (R1), the first successful type-2 mutation occurs within the clone of the first successful type-1 mutation. In Regime 2 (R2), there are several successful type-1 clones by the time the first successful type-2 mutation occurs within one of these clones. Finally, in Regime 3 (R3), many successful type-1 clones arise before the first cancer initiation, and the first successful type-2 can emerge within one of these successful clones or within an unsuccessful type-1 clone. Note that there may be borderline cases that do not fit into one regime, so we denote these as R1/R2 or R2/R3.

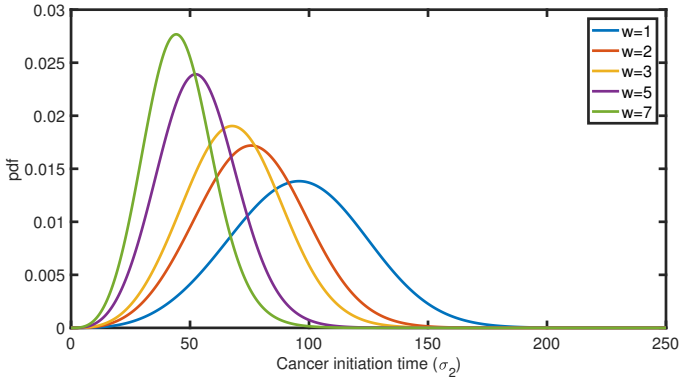


Figure 3.4: **Cancer initiation time σ_2.** Cancer initiation occurs at the time when the first successful type-2 cell arises, which we denote by σ_2. The pdf of σ_2 for various w in parameter regime 2 is shown. The parameters used in this plot are $u_2 = 2 \cdot 10^{-5}, N = 2 \cdot 10^5, u_1 = 7.5 \cdot 10^{-6}$, and $\beta = 0.1$.

Let σ_2 denote the time when the first malignant cancer cell (type-2 cell) arises. Durrett, Foo, and Leder calculated the distribution of σ_2 in [55], and we modified this calculation using the updated speed and shape of a unit ball on $\mathbb{Z}^2 \times \mathbb{Z}_w$. Figure 3.4 shows the updated density of σ_2 in parameter regime 2 for various w. For larger values of w, we expect cancer initiation to occur earlier, due to the increased propagation speed of the premalignant clone.

In Theorem 3.2 of [56], they derive the distribution of the area of the local field, X_l, at time σ_2, conditioned on the event $\{\sigma_2 \in dt\}$. We compare the size-distribution on $\mathbb{Z}^2$ and on $\mathbb{Z}^2 \times \mathbb{Z}_w$ for several different w by varying the type-2 mutation rate u_2 in Figure 3.5 and varying the fitness advantage β in Figure 3.6. In both plots, we condition on cancer initiation occurring at the expected time $\mathbb{E}[\sigma_2]$. In all cases, we see that the larger local fields at the time of cancer initiation are associated with larger w. Due to the conditioning on $\mathbb{E}[\sigma_2]$, the early initiation time for larger w should contribute to a smaller local field, but the density functions imply that the increased propagation speed has a stronger effect on the field size than the faster initiation time, resulting in a larger local field.

In Figure 3.5, we also observe that for all w, the likelihood of a larger local field increases as u_2 decreases. As u_2 decreases, the process moves toward regime 3, in which the premalignant region is made up of many independent clones, increasing the likelihood that a larger clone will give rise to the first type-2 cell. On the other hand, as β decreases, we see in Figure 3.6 that the support of the distribution decreases for all w. A smaller β implies a less aggressive mutation, so we expect the clone to expand more slowly, leading to a smaller local field at the time of initiation. Notice that varying u_2 seems to have a stronger effect on the size-distribution than varying β and that the local field size distribution for smaller w is more sensitive to variation of either parameter u_2 or β than the distribution of a local field with larger w.

Thus, the geometry of the basal zone affects the process of field cancerization during tumorigenesis. The greater the thickness of the basal zone, the larger we expect the local field to be at the time of cancer initiation, implying that there will be a larger region surrounding the tumor with high risk for tumor recurrence. The information provided about basal zone thickness at the beginning of this section suggests that if cancer is detected in thicker nonkeratinized epithelial tissue, such as esophageal, cheek, or palate tissue; in the epithelium surrounding a hair follicle; or in tissue affected by a hyperproliferative skin disease, then a larger region of tissue should be monitored closely for tumor recurrence.

3.6 Conclusions

In this chapter, we have studied the biased voter model on $\mathbb{Z}^2 \times \mathbb{Z}_w$ and have shown how the spread of the model in this setting differs from the spread of the biased voter model on $\mathbb{Z}^d$. The biased voter model on $\mathbb{Z}^2 \times \mathbb{Z}_w$ approximates a basal zone that is w cells thick. We used this model to determine an asymptotic formula for the propagation speed of a premalignant clone in such a basal zone.

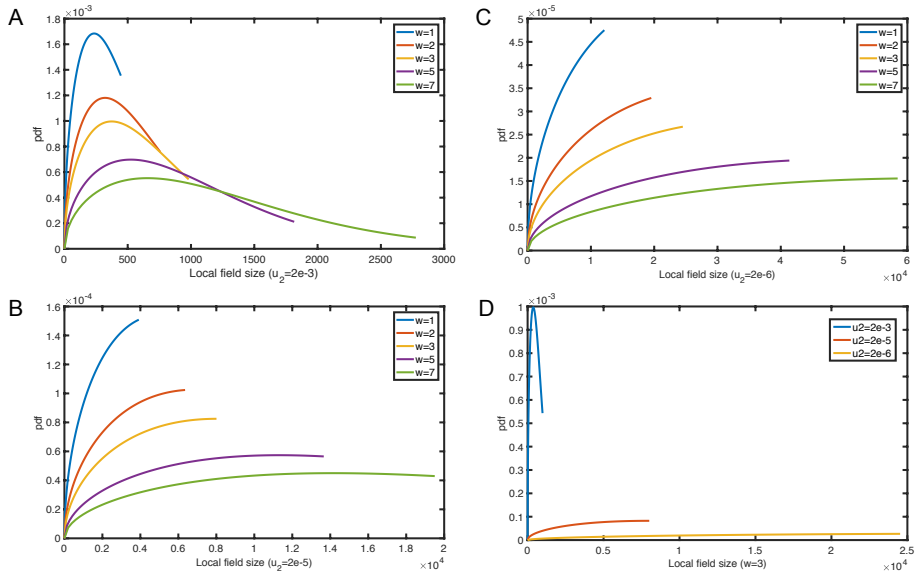


Figure 3.5: **Size-distribution of the local field (varying u_2).** The pdf of the local field size on $\mathbb{Z}^2$ and $\mathbb{Z}^2 \times \mathbb{Z}_w$ for various w is shown. Each plot corresponds to a different type-2 mutation rate and parameter regime. In **(A)** $u_2 = 2 \cdot 10^{-3}$ (R1), in **(B)** $u_2 = 2 \cdot 10^{-5}$ (R2), and in **(C)** $u_2 = 2 \cdot 10^{-6}$ (R2/R3). The other parameters are held constant at $N = 2 \cdot 10^5, u_1 = 7.5 \cdot 10^{-6}, \beta = 0.1$. For the purpose of comparing the scales in the first three plots, in **(D)** we show the graph for all three u_2 values in the case in which $w = 3$.

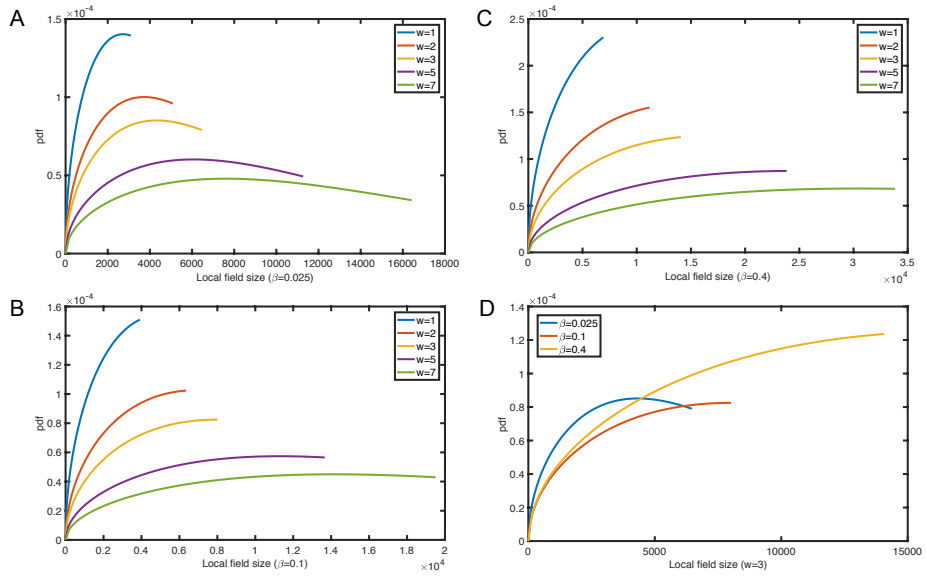


Figure 3.6: **Size-distribution of the local field (varying β).** The pdf of the local field size on $\mathbb{Z}^2$ and $\mathbb{Z}^2 \times \mathbb{Z}_w$ for various w is shown. Each plot corresponds to a different fitness advantage β and parameter regime. In **(A)** $\beta = 0.025$ (R1/R2), in **(B)** $\beta = 0.1$ (R2), and in **(C)** $\beta = 0.4$ (R2/R3). The other parameters are held constant at $N = 10^4, u_1 = 7.5 \cdot 10^{-6}, u_2 = 2 \cdot 10^{-5}$. In **(D)** we show the graph for all three β values in the $w = 3$ case.

In order to determine this speed, we began by proving a local central limit theorem on $\mathbb{Z}^d \times \mathbb{Z}_w$. We used this result when analyzing the dual process of the biased voter model. This analysis allowed us to determine an upper and lower bound for the propagation speed. Theorem 3.2.1 gives the asymptotic formula for the rate of spread of a premalignant clone on such a basal layer, as a function of the selective advantage β. This result, in comparison with the result in [55] on $\mathbb{Z}^d$ shows that the geometry of the basal zone influences the spread of mutant clones in epithelial tissue. We showed that for small β, premalignant clones eventually spread out more quickly when the basal zone is more than one cell thick.

The rate of spread of premalignant clones has an important clinical implication during the time period after a tumor has been detected. The local fields surrounding primary tumors are composed of cells that have mutated and are dividing more quickly than normal cells. Thus, they face a higher risk than normal cells of further mutation and eventually the initation of another tumor. We utilized our speed result and the distributions of cancer initiation time and local field size from [55] and [56], respectively, to investigate the process of field cancerization in tissue with basal zones of varying width. We determined that cancer initiation tends to occur sooner as the width of the basal zone increases. Additionally we showed that the local field at the time of cancer initiation tends to be smallest in tissue with a single basal layer and that the expected field size increases as w increases.

It is difficult for a clinician to detect the local field, since it appears phenotypically normal, but it would be very useful to know the size of this region, so that it could inform the location and frequency with which the clinician should monitor a cancer patient for signs of recurrence. For example, if a tumor is discovered in esophageal tissue, known to have multiple layers of proliferating cells, then a larger region surrounding the tumor should be monitored than if the tumor had been discovered in skin tissue. Thus, our speed result, combined with the local field size-distribution and knowledge about the structure of the basal zone of the patient's affected tissue, can provide important information for predicting recurrence location and timing.

A limitation of our work is that we are confined to a uniform fitness advantage β conferred by all mutations and to a specific sequence of genetic events. In reality, different mutations could lead to various pathways with divergent mutation rates, fitness advantages, and number of events required to initiate cancer. However, our approach provides a useful construction for studying the general differences in field cancerization between tissues with different basal layer thickness. It also equips us to differentiate between various genetic pathways in molecular subtypes of carcinoma and to predict their impact on the spread of local fields. Another limitation is the periodic boundary condition on the third dimension,

which was used because the proof of the local central limit theorem and other parts of the analysis become much more complicated if we remove this condition. However, we showed via simulation that the propagation speed of a BVM with a periodic boundary condition appears to be quite similar to the speed of a BVM with a more biologically realistic reflecting boundary condition in the third dimension. Generalizing the proof of the propagation speed to a BVM with a reflecting boundary condition is a topic of future work.

We also plan to analyze the mutant clone propagation speed on $\mathbb{Z} \times \mathbb{Z}_w$ in the future. Our motivation for studying the speed in this setting arises from epithelial tissue that is structured in thin cylindrical tubes, e.g. mammary ducts of the breast and renal tubules of the kidney. We plan to use our local central limit theorem on $\mathbb{Z} \times \mathbb{Z}_w$ to aid in the analysis, and we hope to determine how the propagation speed and the process of field cancerization change in this geometric setting.

Chapter 4

Glioblastoma recurrence and the role of MGMT promoter methylation

Glioblastoma, also known as glioblastoma multiforme (GBM), is an extremely fast-growing and lethal form of brain cancer. Despite aggressive treatment strategies, the clinical prognosis for glioblastoma is grim. Typically, GBM patients are treated with surgical resection followed by adjuvant radiation therapy and chemotherapy with the oral alkylating agent temozolomide (TMZ). This standard regimen results in a median survival of only 15 months and a two-year survival rate of 30% [2].

The effectiveness of TMZ is impacted by the methylation status of the promoter for DNA repair protein O-6-Methylguanine-DNA Methyltransferase (MGMT). In particular, the poor prognosis in GBM patients is largely due to the near-universal recurrence of tumors after treatment, which is often driven by resistance to the chemotherapy component TMZ [69, 70]. In responsive tumors, TMZ typically induces cellular apoptosis via DNA strand breaks. Resistance to TMZ in GBM has been associated with increased expression levels of MGMT [16, 17, 18]. Indeed, clinical studies have shown that epigenetic silencing of the MGMT gene via promoter methylation is associated with greater sensitivity to TMZ and improved patient prognosis [16, 71]. For example, in [16] it was shown that MGMT-methylated GBM patients treated with TMZ and radiotherapy had a survival benefit, as compared to those treated with only radiotherapy; this survival benefit disappeared in the absence of MGMT promoter methylation.

It has been proposed that TMZ may actually impact the methylation status of the

MGMT promoter during treatment [72]. However, the details and mechanism of this proposed interaction are uncertain. There have been a small number of studies comparing the MGMT promoter methylation in newly diagnosed tumors vs matched recurrence samples, following treatment with TMZ [19, 20, 73, 74]. Many of these studies demonstrate that a majority of patients with MGMT methylated tumors at diagnosis actually present with unmethylated recurrent tumors following standard treatment; thus there is a downward shift in the overall methylation percentage of these tumors during the course of treatment. However, it is unclear whether this transition from methylated to unmethylated recurrent tumors is due to TMZ actively influencing the methylation status of MGMT, simply a result of evolutionary selection for a more drug-tolerant phenotype, or some combination of both processes. In this work we aim to utilize evolutionary modeling to help shed light on this question.

Our work adds to a large body of literature modeling the response of glioblastoma to various treatment strategies. In [75] the authors model chemotherapeutic delivery to brain tumors and the extent of the breakdown in the blood-brain barrier using a two-compartment catenary model. In [76], a spatio-temporal model of glioblastoma response to temozolomide that allows for patient-specific chemotherapeutic optimization is developed. The model in [77] explores the interactions between rapidly proliferating cells and a dormant cell population within GBM, as well as the effect of various treatment schedules on the overall composition of the tumor. There are a number of existing papers studying the effect of fractionated radiation dosing on malignant tumors using the linear-quadratic model [78, 79]. In [80], the linear-quadratic model is used to describe the tumor control probability under radiotherapy, and [81] investigates the optimal radiating dosing strategies specifically in glioblastoma. Powathil and colleagues consider a spatio-temporal brain tumor model that includes effects from both radiotherapy and chemotherapy in [82]. Patient-specific models of glioblastoma are developed in [83] and [84] to predict patient response to radiotherapy and to determine optimal radiation dosing strategies. Many mathematical modeling efforts focusing on glioblastoma tumor growth and therapy response are reviewed in [85].

Mathematical models have also been developed to describe the process of DNA methylation changes in cells. Yatabe and collaborators developed a methylation-based model to trace stem cell dynamics in human colon crypts [86]. Otto and Walbot introduced the first model that described methylation in terms of both maintenance and de novo methylation [87]; a similar model in a continuous-time framework was developed in [88]. Genereux and collaborators built upon this model by allowing for the possibility that de novo methylation occurs on one daughter strand and not the other [89]. In [1], Sontag and colleagues present

a discrete-time Markov chain version of the methylation model introduced by Otto and Walbot.

Here, we develop and parameterize a stochastic model of the evolutionary dynamics driving GBM response to standard treatment with surgery, radiation and TMZ. We focus on investigating the role of MGMT promoter methylation on TMZ resistance and tumor recurrence after treatment. Thus, our model combines a mechanistic description of DNA methylation dynamics at a characteristic site within the MGMT promoter region with an evolutionary model of GBM treatment response and resistance dynamics. We aim to investigate potential mechanisms underlying observed methylation patterns at diagnosis and recurrence, and also explore the impact of these epigenetic processes on tumor response to therapy.

The outline of this chapter is as follows: in Section 4.1 we provide the biological background that motivates the model, and in Section 4.2 we describe the framework and components of the mathematical model. Section 4.3 describes the clinical and experimental data we collected, used to calibrate the parameters within the model. In Section 4.4 we present our findings regarding methylation changes in tumors during therapy and optimal adjuvant TMZ dosing strategies, contingent upon tumor methylation status at diagnosis. We summarize these results and discuss future directions in Section 4.5.

4.1 Background

4.1.1 A review of DNA methylation

DNA methylation involves the addition of a methyl group to DNA, most often at the 5-carbon of a cytosine ring. Typically, methylation in somatic cells of mammals occurs at CpG dinucleotides in the DNA sequence. Gene promoter regions often contain clusters of of CpG sites called CpG islands, and methylation of these CpG islands within promoter regions effectively silences transcription of the the gene [90, 91].

The process of DNA methylation is carried out by three major DNA methyltransferases (DNMT1, DNMT3a, and DNMT3b), during and immediately following cell replication. DNMT1 is responsible for 'maintenance methylation,' in which patterns of methylation in the original parental DNA are preserved in the replicated DNA. DNMT3a and DNMT3b are responsible for *de novo* methylation, in which unmethylated sites in the parental DNA become methylated in the replicated DNA [92, 93, 94]. Figure 4.1 shows an illustration of the roles of the three methyltransferases during cell replication.

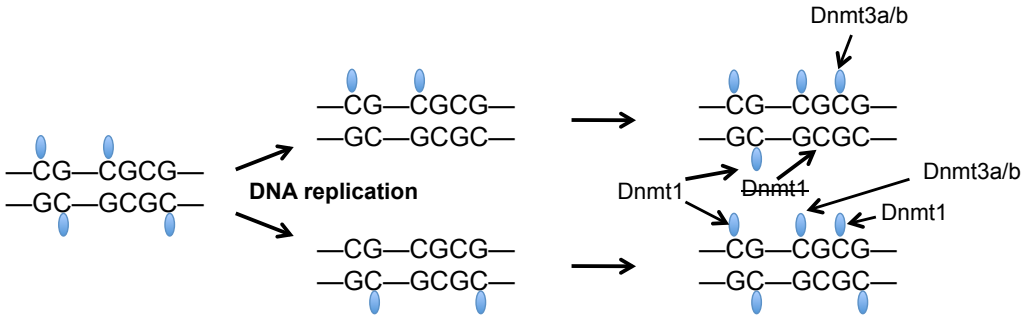


Figure 4.1: **The role of DNA methyltransferases DNMT1 and DNMT3a/b during DNA replication.** This figure illustrates a portion of a DNA molecule splitting during replication. Notice that the DNMT1 methylate the sites in the new strand that were methylated in the parental strand. As this process is not perfect, some sites can be missed. Dnmt3a/b methylates new sites that were not previously methylated in the top strands of the upper and lower molecules. Similar to a figure in [1].

Demethylation, which refers to the loss of methyl groups from DNA, can either be passive, active, or some combination of both. Passive demethylation refers to the failure of DNMT1 to preserve all of the methylated sites in the parental DNA during replication. Active demethylation refers to the removal of methyl groups via the modification of cytosine bases that have been converted by TET enzyme-mediated oxidation [95]. In this work we will primarily be concerned with passive demethylation via DNMT1.

4.1.2 MGMT Methylation and TMZ resistance

TMZ is an alkylating agent that attaches a methyl group to purine bases of DNA (O6-guanine; N7-guanine and N3-adenine); methylation of O6-guanine damages cellular DNA and triggers cell death. The MGMT gene encodes for the DNA-repair protein O6-alkylguanine DNA alkyltransferase (AGT) which removes alkylating groups from the O6-guanine in DNA, thus repairing DNA damage caused by TMZ [17]. Hence, promoter methylation of MGMT results in decreased DNA-repair ability and increased sensitivity to TMZ. On the other hand, unmethylated MGMT promoter regions result in increased DNA-repair ability and decreased sensitivity to TMZ. Indeed, MGMT is methylated in approximately 40-50% of newly diagnosed glioblastomas, and studies have shown that the presence of MGMT promoter methylation is a strong predictor of responsiveness to TMZ [96, 16, 97].

The typical binary stratification of tumors into 'MGMT-methylated' or 'MGMT-unmethylated' requires some clarification, since methylation status can vary between cells of a tumor as

well as between various CpG sites in the same genic promoter region. Thus, the actual description of MGMT methylation within tumors is more nuanced and continuously varying than the terminology suggests. There are 97 CpG sites in the MGMT promoter region, of which 36 have been shown to correlate with MGMT expression in the study [98]. To determine methylation status of a tumor, typically a small subset (< 5) of these sites is examined for the presence of methylation in a sample of tumor cells. The methylation percentage (i.e. percentage of cells methylated) is calculated at each site and then averaged across the sites. Then a threshold mean methylation percentage, which may vary between studies, is used to stratify tumors into a 'methylated' or 'unmethylated' status. Since there is so much variation between studies in how tumors are placed into these categories, here we opt to study quantitative changes in methylation percentage on a representative CpG site, rather than setting a threshold mean methylation percentage for stratification.

There have been a limited number of clinical studies comparing the methylation status of tumors before and after treatment with TMZ. In [19], 8 of 13 patients transitioned from an MGMT methylated primary tumor to an unmethylated recurrent tumor after treatment, and in [20], it was reported that 10 of the 13 patients switched from a methylated primary tumor to an unmethylated recurrent tumor. In this study, two of the primary tumors began unmethylated and remained unmethylated at recurrence, and the other primary tumor's methylation status was not detectable. Hence, all characterizable recurrent tumors in the study were unmethylated. Additionally, in [74], authors observed that 39.1% of pretreatment GB and 5.3% of recurrences were promoter methylated, in addition to an observed increase of MGMT activity in recurrences. Lastly, in [73] 15 of 18 recurrence samples displayed higher MGMT expression as compared to matched primary samples.

Note that one may find some studies that appear to show the opposite result, concluding that the majority of recurrent tumors are methylated. However, these studies do not compare the MGMT promoter methylation status of matched samples of individual tumors at detection and recurrence; instead, they separately compare the total proportion of methylated tumors at detection and recurrence. Due to the unresponsive nature of MGMT-unmethylated tumors, many of these patients do not survive until the clinical definition of tumor recurrence – thus introducing a selection bias. It is unclear from the studies in [19, 20, 74] alone whether the transition to unmethylated recurrent tumors is a result of selection or whether the TMZ treatment increases the rate of demethylation, as some have hypothesized [18, 19].

4.1.3 Standard treatment regimen for GBM

The standard treatment for GBM involves surgical resection of the tumor, followed by concurrent radiotherapy and chemotherapy with TMZ (called CRT), and then adjuvant chemotherapy until tumor recurrence. More specifically, after surgical resection the patient first recovers for three weeks before starting the CRT phase of treatment. During the 6-week CRT phase, radiotherapy is administered in daily fractions of 2 Gy, given five days per week. In total, 60 Gy of radiotherapy is administered during the six-week period. In addition, the tumor is treated every day during the six-week period with 75 mg of TMZ per square meter of body-surface area.

Next, following a three-week recovery period, adjuvant chemotherapy is administered to the patient. One 28-day cycle consists of five daily doses of 150-200 mg/m^2 of body-surface area, followed by a 23-day treatment break; this cycle repeats until the tumor recurs. A schematic of this standard treatment schedule is depicted in Figure 4.2. Further details of the standard regimen can be found in [2].

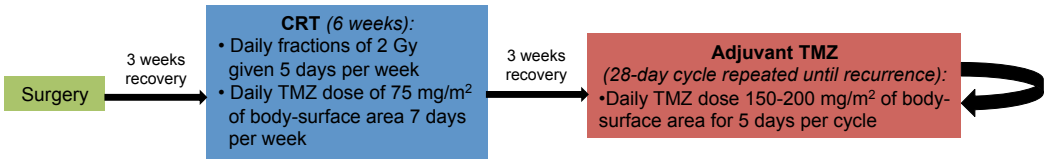


Figure 4.2: **The standard GBM treatment schedule.** The schedule consists of surgery, concurrent radiotherapy and TMZ, and adjuvant TMZ treatment [2].

4.2 Mathematical model

We develop a stochastic model describing the evolutionary dynamics of GBM response to standard treatment. In particular, we utilize a multi-type, continuous-time branching process model (see, e.g. [99]), in which each cell waits an exponential amount of time before division or death, governed by its birth and death rates. The model consists of three cellular subtypes: Type-1, referring to GBM cells with fully methylated MGMT promoters, Type-2, GBM cells with hemimethylated MGMT promoters, i.e. promoters with methylation on one strand and no methylation on the other, and Type-3, GBM cells with unmethylated MGMT promoters. The type-1 cells are TMZ-sensitive, and type-2 and type-3 cells are both considered TMZ-resistant, since they both can repair the lesion created by TMZ.

Let $X_1(t), X_2(t)$, and $X_3(t)$ denote the number of type-1, type-2, and type-3 cells, respectively, at time t. All three populations are birth-death processes, and the TMZ-resistant cells, $X_2(t)$ and $X_3(t)$, are assumed to have the same birth and death rates, while the TMZ-sensitive cells have a distinct birth and death rate. The rates governing these birth-death processes vary during treatment with TMZ and radiation. Conversions also occur between these cell types, driven by methylation and demethylation events during and immediately after cell division. Each of the three cell types has a distinct offspring distribution, driving the immigration events between methylated, hemimethylated, and unmethylated cells.

The model describes three distinct phases of tumor development and treatment, structured to the standard regimen and parameterized using clinical and experimental data.

- Phase 1 (P1): Tumor growth before detection, surgery, and three-week recovery
- Phase 2 (P2): Concurrent radiotherapy and chemotherapy (referred to as CRT), and three-week recovery
- Phase 3 (P3): Adjuvant chemotherapy (administered in 28-day cycles) until tumor recurrence

A schematic of the three model phases is provided in Figure 4.3. Below we describe how the dynamics of the branching process model are adapted to model each specific phase of the treatment.

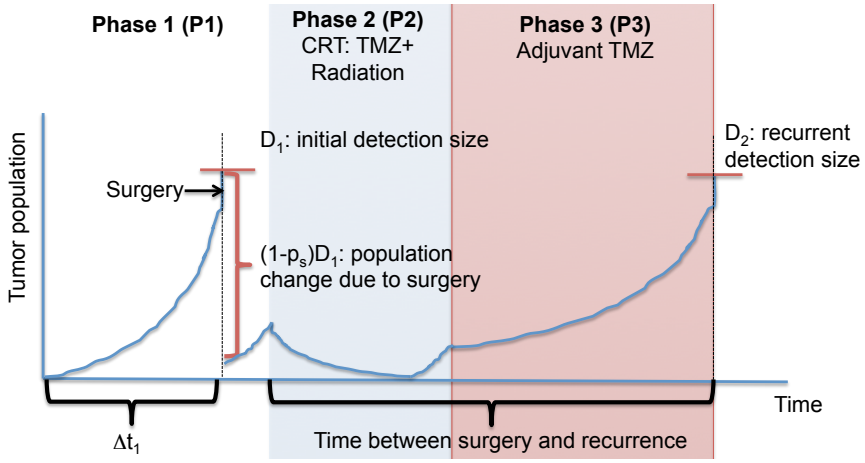


Figure 4.3: **Schematic of the three model phases.** P1 consists of the tumor growth prior to detection and surgery, P2 denotes the concurrent radiation and chemotherapy (CRT) phase of treatment, and P3 refers to the adjuvant chemotherapy following CRT.

Methylation and demethylation processes. We begin by considering in detail the

processes of methylation and demethylation on the MGMT promoter; these processes are ongoing throughout all treatment and pre-treatment phases of the model. We utilize a variant of the model presented in [1], in which a discrete-time Markov chain is used to describe maintenance methylation, and *de novo* methylation at a CpG site. This underlying model of the detailed dynamics of methylation then feeds into the population-level branching process model via the rates of conversion between the cellular subtypes.

Let ρ be the probability of maintaining methylation for any given CpG site after replication, i.e the probability that DNMT1 methylates a CpG dyad after replication, conditioned on the event that the site was methylated before replication. Let ν be the probability of *de novo* methylation, i.e. the probability that DNMT3a or DNMT3b methylates any CpG site that is unmethylated immediately following DNA replication. Inspired by [1], we derive the following offspring distributions for each of the three cell types, conditioned on cell division. In these distributions, $p_i(x, y, z)$ refers to the probability that a type-i cell will produce x type-1 cells, y type-2 cells, and z type-3 cells after replication. For ease of notation, let $A := (1 - \rho)(1 - \nu)$.

$$
\begin{aligned}
p_1((2,0,0)) &= (1-A)^2 \\
p_1((1,1,0)) &= 2A(1-A) \\
p_1((0,2,0)) &= A^2
\end{aligned}
\tag{4.1}
$$

$$
\begin{aligned}
p_2((2,0,0)) &= \nu^2(1-A) \\
p_2((1,1,0)) &= 2\nu(1-\nu)(1-A) + \nu^2 A \\
p_2((1,0,1)) &= (1-\nu)^2(1-A) \\
p_2((0,2,0)) &= 2\nu(1-\nu)A \\
p_2((0,1,1)) &= A(1-\nu)^2
\end{aligned}
\tag{4.2}
$$

$$
\begin{aligned}
p_3((2,0,0)) &= \nu^4 \\
p_3((1,1,0)) &= 4(\nu^3 - \nu^4) \\
p_3((1,0,1)) &= 2\nu^2(1-\nu)^2 \\
p_3((0,2,0)) &= 4\nu^2(1-\nu)^2 \\
p_3((0,1,1)) &= 4\nu(1-\nu)^3 \\
p_3((0,0,2)) &= (1-\nu)^4
\end{aligned}
\tag{4.3}
$$

To derive these offspring distributions, we made use of the fact that a methylated dyad produces two hemimethylated dyads when the DNA strands split during replication, and those sites remain hemimethylated if the site without methylation is not methylated by DNMT1 or DNMT3a/b immediately following replication. Hence, the probability that each dyad remains hemimethylated is $A = (1-\rho)(1-\nu)$, and consequently the probability of producing two hemimethylated dyads, i.e. two type-2 cells, is A^2. Conversely, the probability that one of those hemimethylated sites becomes fully methylated is $1-A$, so the probability of producing two fully methylated cells, i.e. two type-1 cells, is $(1-A)^2$, and the probability of producing one type-2 cell and one type-1 cell is $A(1-A)$.

The offspring distributions for type-2 and type-3 cell replication can be verified similarly upon inspection, using the idea that an unmethylated dyad produces two unmethylated dyads during replication, and each of the CpG sites making up these dyads can only be methylated with DNMT3a/b, i.e via *de novo* methylation.

Note that we will investigate the impact of TMZ on the development of resistance by allowing the methylation probabilities to change in presence of TMZ. In this case ν_z, ρ_z will be used to denote the *de novo* and maintenance probabilities, respectively, in the presence of the drug.

Phase I: Pretreatment, surgery, recovery. In the absence of treatment, the intrinsic birth rates of untreated fully methylated (type-1) and hemi/unmethylated (type-2 and type-3) cells are b_1 and b_2 per day, respectively, and their death rates are c_1 and c_2, respectively. The parameters of the model are determined using experimental and clinical data (see appendix section C.1); however, note that the total tumor population size $X_1(t) + X_2(t) + X_3(t)$ is a supercritical branching process during any untreated period.

Once the tumor population reaches a detection size threshold D_1, we model surgical resection of the tumor by removing p_s percent of the total cells, chosen proportionally for each subtype. After surgery, the patient recovers for three weeks before starting treatment, so in the model, the initial birth and death rates drive the regrowth of the tumor.

Phase II: TMZ, radiation for six weeks and recovery. In P2, the tumor undergoes concurrent radiotherapy and chemotherapy (called CRT) for 6 weeks. The standard schedule for radiotherapy is a daily fraction of 2 Gy, given five days per week, on Monday through Friday. In addition, the tumor is treated every day during the six-week period with 75 mg of TMZ per square meter of body-surface area.

Since TMZ is a cytotoxic treatment, we model its impact as primarily affecting the death rates of the tumor cells, c_1 and c_2. Let $g_1(t), g_2(t)$ be the additional death rate due to TMZ treatment for type-1 and type-2/type-3 cells, respectively. Note that these components vary with time because they depend on the current TMZ concentration level. See appendix section C.1 for details on how $g_1(t), g_2(t)$ are determined from experimental data.

The cytotoxic effect of radiotherapy is modeled using the standard linear-quadratic (L-Q) model [80]. In this model, radiosensitivity parameters α, β are used to account for toxic lesions to DNA and misrepair of repairable damage to DNA, respectively [100]. Under the L-Q model, the probability of cell survival at time t under the L-Q model is dependent on the radiation dose at time t, $D(t)$, in the following manner:

$$S(t) = e^{-\alpha D(t) - \beta D(t)^2}.$$

In our model simulations, at the time t of each radiation dose, we instantaneously remove $(1 - S(t))X_1(t)$ type-1 cells, $(1 - S(t))X_2(t)$ type-2 cells, and $(1 - S(t))X_3(t)$ type-3 cells. During the three-week recovery period, the cellular birth and death rates revert to those

used in the pretreatment growth phase. Given data constraints, we ignore differences in radiosensitivity between Type-1 and Type-2/3 cells.

Phase III: Adjuvant TMZ. During P3, adjuvant chemotherapy is administered to the tumor. In this phase, the additional death rates $g_1(t)$ and $g_2(t)$ due to chemotherapy reflect five daily doses of 150-200 mg/m^2 of body-surface area, followed by 23 days off. This 28-day cycle repeats until the tumor recurs. Tumor recurrence occurs when the tumor population size reaches the threshold D_2, obtained from clinical data.

4.3 Experimental and clinical data

Experimental setup. We performed experiments on PDX cell lines to investigate the differential impact of TMZ on the growth kinetics of MGMT-methylated and unmethylated GBM cells. In these *in vitro* experiments, plates of GBM6 cells were treated at eight concentrations of TMZ (including DMSO) in triplicate, and live and dead cell counts were collected via MTT and trypan blue assays after 8 days of exposure. The average number of live cells after 8 days for each TMZ concentration is displayed in Figure 4.5a. Figure 4.5b plots the average proportion of live cells from the sum of live and dead cells after 8 days of exposure, as a function of TMZ dose.

In addition, the frequency of cells expressing MGMT was assessed in each group after eight days. The average MGMT$^+$ frequency for each concentration of TMZ is displayed in Figure 4.4.

Clinical data. Clinical data was also collected from a group of 20 adult patients that received the standard protocol described in [2]. Information about tumor radius size was collected from each patient at the time of initial detection and at the time of recurrence. This data is summarized in Figure 4.6a.

The growth of the tumor in the absence of treatment was also tracked, resulting in net growth estimates for each patient. Using the patient data and a reaction-diffusion model described in [101], we obtained an average net growth rate estimate before treatment of $\lambda = 0.0897$/cell/day. Individual patient growth rates are summarized in Figure 4.6b. Patient data describing the tumor radius size after surgery is displayed in Figure 4.6c.

4.4 Results

Selection alone does not explain methylation shift in recurrent tumors. We first used the parameter settings obtained from the processes described in Section C.1 to examine

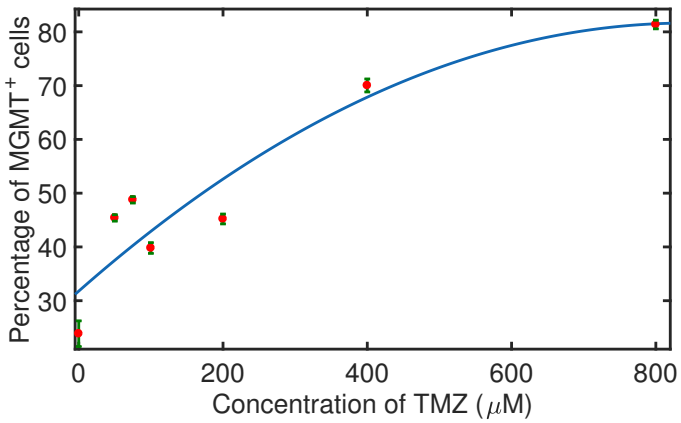
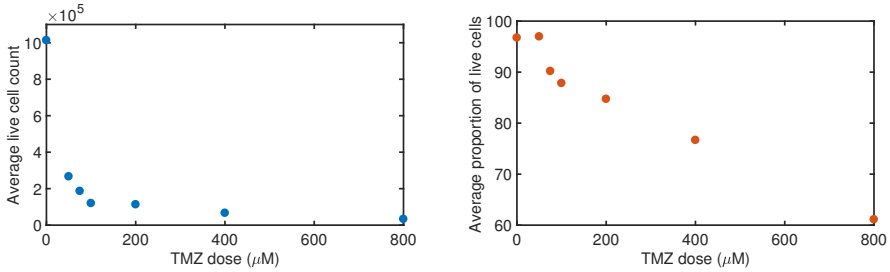
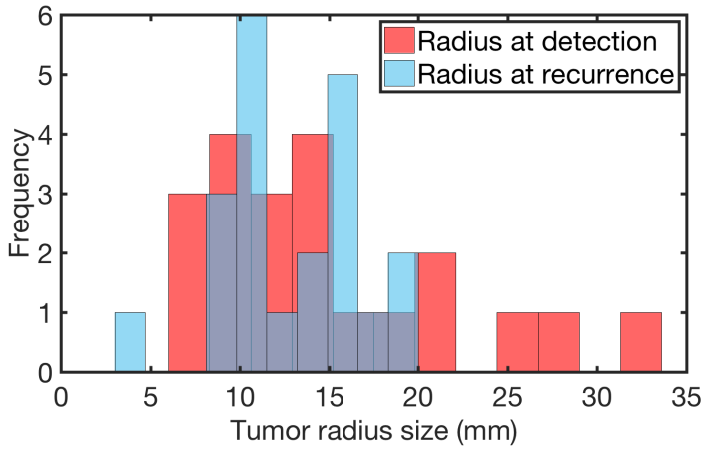


Figure 4.4: **Percentage of cells expressing MGMT, as a function of TMZ dose.** Data was collected after 8 days of exposure to various concentrations of temozolomide (in νM), assessed using PDX experiments. The red dots in the plot denote average percentages of MGMT$^+$ cells, and the error bars indicate the standard deviation.

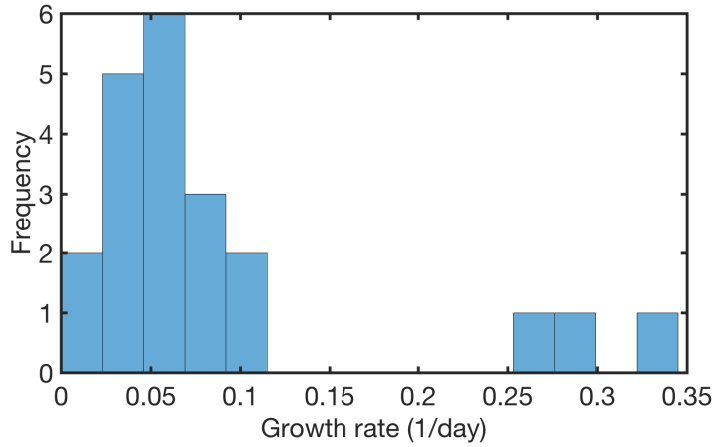


(a) Live cell counts (b) Mean proportion of live cells

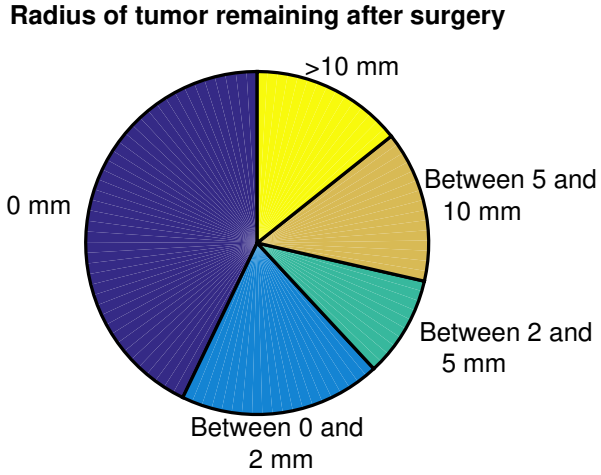
Figure 4.5: **Cell count data, as a function of TMZ dose.** Plots of (a) the average live cell counts and (b) the mean proportion of live cells, out of the total cells (live and dead cells), collected after 8 days of exposure to TMZ, as a function of the concentration of TMZ exposure (in μM).



(a) Tumor radius



(b) Net growth rate



(c) Tumor radius after surgery

Figure 4.6: **Clinical data from GBM patients undergoing standard regimen ($n = 21$).** Histograms of (a) tumor radius at detection and recurrent tumor radius (mm), (b) the overall net growth rate (1/day) before treatment, and (c) a pie chart depicting the radius of tumor remaining after surgery (mm).

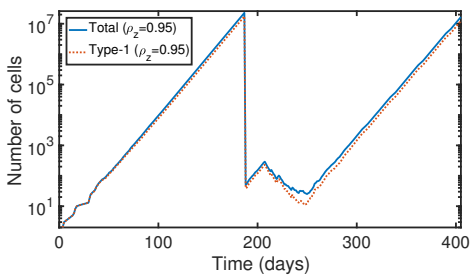
the relative methylation percentages at diagnosis and recurrence, i.e. times at which the tumor population hits D_1 and D_2 as described in the Model section. Figure 4.7a shows the type-1 (methylated proportion) and total population sizes in the tumor during single sample path simulation of the model. We observe that the tumor recurs approximately 214 days after surgery is performed. Figure 4.7b shows the distribution of recurrence times from a computational experiment with 100 Monte Carlo simulations; the median recurrence time is 213.8 days. This is roughly consistent with clinical data reported in [102], where the median recurrence time is 191 days.

Figures 4.7c and 4.7d, show the distribution of methylation percentages found at these times. We observe that the average proportion of methylated (type-1) cells to total cells at the time of recurrence is roughly the same as at the time of diagnosis. The distribution of the change in methylation percentage between detection and recurrence is depicted in Figure 4.7e. The slight reduction in overall methylation percentage suggests that selection alone cannot account for the significant reduction in methylation observed in clinical studies, described in section 4.1.2.

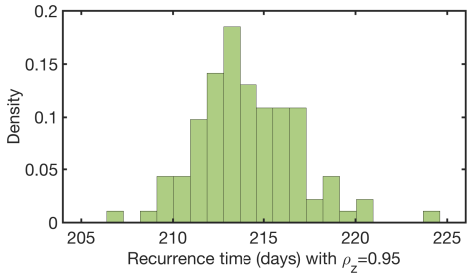
TMZ inhibition of maintenance methylation causes downward methylation shift. We next investigated the hypothesis that an active role of TMZ on the cellular methylation processes may be able to explain the methylation downshift in recurrent tumors. In particular, we investigated the possibility that TMZ may decrease the amount of time spent in the type-1 (methylated) state and increase time spent in type-2/3 states. This may result from a decrease in either the *de novo* methylation probability ν or the maintenance methylation probability ρ. Note that for this investigation, the parameters ν, ρ will deviate from their baseline values only during TMZ treatment periods; thus we denote the parameters during TMZ treatment as ν_z, ρ_z.

To investigate the effects of changing the *de novo* methylation probability, ν_z, in the presence of TMZ, we first note that the lowest possible value of ν_z is 0, representing no *de novo* methylation events. If we let $\nu_z = 0$, then we observe a modest decrease in the expected methylation percentage between detection and recurrence, changing by less than 7%. Figures 4.8a and 4.8b show the distribution of methylation percentages at the time of recurrence and the distribution of change in methylation percentage between detection and recurrence, respectively, when $\nu_z = 0$. Thus, a significant drop in methylation percentage after TMZ treatment cannot be attributed to an inhibitory impact on *de novo* methylation.

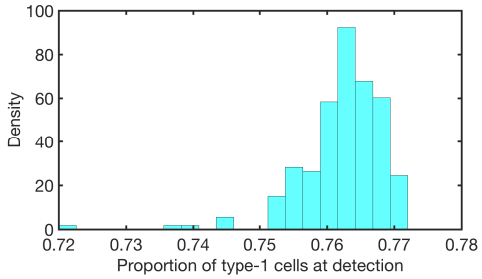
We next investigated the impact of decreasing the probability of maintenance methylation, ρ_z, during chemotherapy. Figure 4.8c displays the type-1 frequency at the time of recurrence when $\rho_z = 0.5$, reduced from the baseline value of $\rho = 0.95$, and figure 4.8d shows



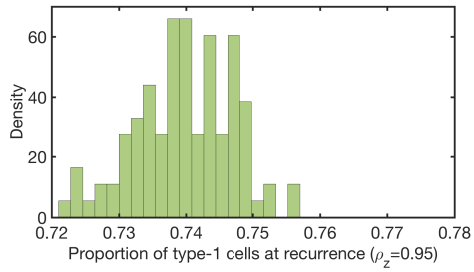
(a) Sample path simulation of model.



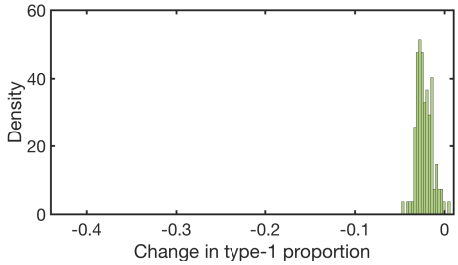
(b) Distribution of recurrence times.



(c) Type-1 frequency at detection.



(d) Type-1 frequency at recurrence.



(e) Change in Type-1 frequency.

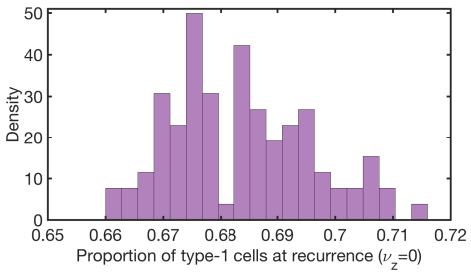
Figure 4.7: **Simulation results – no TMZ impact on methylation rates.** Plots of (a) one sample path simulation of the model, (b) the distribution of recurrence times in a computational experiment with 100 samples, (c) the distribution of methylation percentage at the time of detection, (d) the distribution of methylation percentage at the time of recurrence, and (e) the distribution of change in methylation percentage between detection and recurrence. All parameters are set as described in section C.1.

the change in methylation percentage between detection and recurrence. We observe there is a much more significant decrease in methylation in this case, as we see in clinical observations, than in the case when TMZ has no impact on methylation rates (compare Figures 4.8d and 4.7e). Figure 4.8e displays the expected proportion of type-1 cells at recurrence, as a function of ρ_z. For smaller values of ρ_z, the proportion of type-1 cells after treatment decreases substantially from a mean methylation percentage of 0.762 at detection. Thus, a role of TMZ in inhibiting maintenance methylation, but not *de novo* methylation, can explain the downward shift in methylation that has been observed clinically. In Section C.2 we show that this claim is robust to variability in the model parameters.

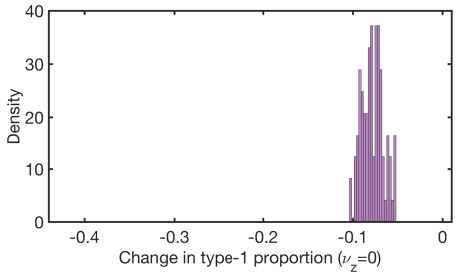
Optimization of adjuvant TMZ schedule to minimize expected tumor size. We next used the model to investigate the optimal number of TMZ doses during Phase III, the adjuvant chemotherapy phase, to minimize the expected tumor size after 4 cycles of treatment. In the standard treatment schedule, five TMZ doses of 150-200 mg/m^2 are administered daily at the beginning of each 28-day cycle. Let n denote the number of TMZ doses in a single 28-day cycle. We vary n in order to determine the number of doses and dose level that minimizes the number of total cells remaining after 4 adjuvant cycles. Let $Z(n)$ denote the TMZ concentration level per dose, in mg/m^2, when n doses are administered per cycle. Each dose concentration is set at $Z(n) = 1000/n$ for varying values of n, where $0 < n \leq 28$, so that the cumulative TMZ dosage during one cycle does not exceed 1000 mg/m^2. Based on our previous investigations, the maintenance methylation probability in the presence of TMZ, ρ_z, is assumed to be 0.5.

Mean calculations for each cell-type, provided in Section C.3, are used to determine the n that minimizes the expected tumor size after 4 cycles. Figures 4.9a and 4.9b show the mean tumor size, number of fully methylated cells (type-1), and cells that are not fully methylated (type-2 and type-3) when $\rho_z = 0.5$. In this case, the optimal number of doses per cycle, i.e. the number that results in the smallest mean tumor population after 4 cycles, is $n = 6$, with $Z(6) = 166.67$ mg/m^2. This is close to the standard administered dose during adjuvant chemotherapy, and we see a small difference in the expected tumor size when 5 vs 6 doses are administered. Hence, our model suggests that the standard dosing schedule is a reasonable, though not optimal, protocol for highly methylated tumors at diagnosis.

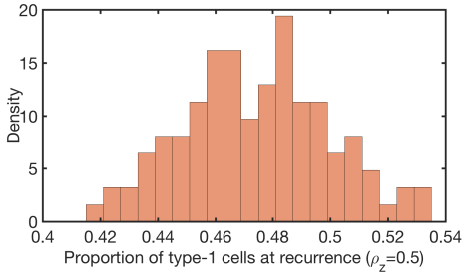
We also used the model to investigate the optimal adjuvant TMZ schedule for tumors with lower methylation percentages at diagnosis. To this end, we first identified the combination of birth rates ($b_1 = .0569$ day^{-1} and $b_2 = 0.1276$ day^{-1}) that satisfied the net growth rate constraint and led to 30% methylation at detection. Figures 4.9c and 4.9d display plots of the mean number of total, fully methylated (type-1), and non-methylated (type-2 and



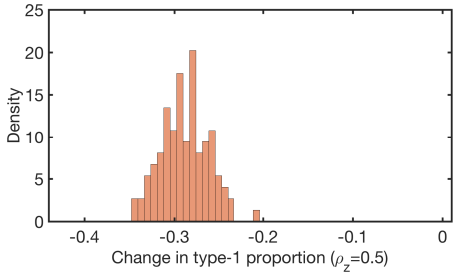
(a) Type-1 frequency at recurrence.



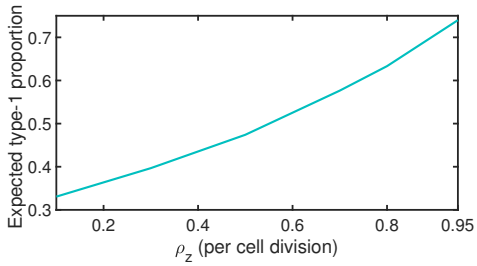
(b) Change in Type-1 frequency.



(c) Type-1 frequency at recurrence.

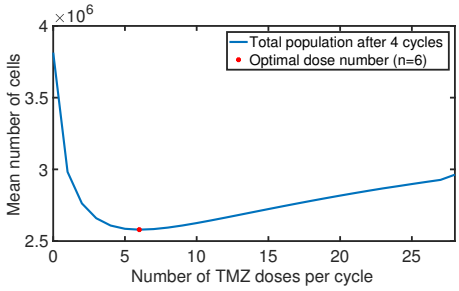


(d) Change in Type-1 frequency.

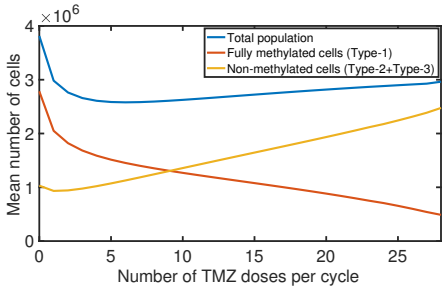


(e) Expected type-1 proportion at recurrence.

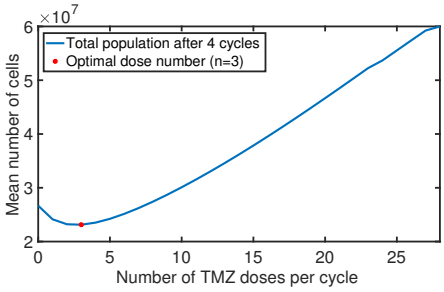
Figure 4.8: **Simulation results – TMZ impacts methylation rates.** Plots of (a) the distribution of methylation percentage at the time of recurrence when $\nu_z = 0$, (b) the distribution of change in methylation percentage between detection and recurrence when $\nu_z = 0$, (c) the distribution of methylation percentage at the time of recurrence when $\rho_z = 0.5$, (d) the distribution of change in methylation percentage between detection and recurrence when $\rho_z = 0.5$, and (e) expected proportion of type-1 cells at recurrence under the standard treatment schedule, as a function of the maintenance methylation probability, ρ_z. Non-varying parameters are set to the baseline values described in Section C.1.



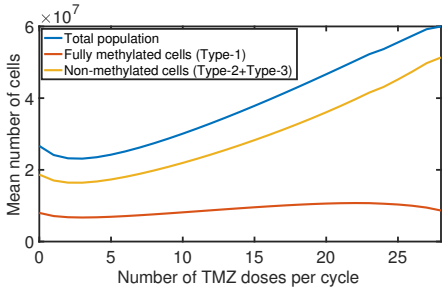
(a) Expected tumor size ($\rho_z = 0.5$).



(b) Cell type sizes ($\rho_z = 0.5$).



(c) Expected tumor size (Initial methyl 30%).



(d) Cell type sizes (Initial methyl 30%).

Figure 4.9: **Adjuvant TMZ optimization results.** Plots of (a) the mean tumor population size and (b) the mean total, type-1, and type-2/3 cell population size when $\rho_z = 0.5$, and (c) the mean tumor population size and (b) the mean total, type-1, and type-2/3 cell population size, when the expected methylation proportion at diagnosis is 0.3. The mean cell populations are calculated after 4 adjuvant chemotherapy cycles, as a function of the number n of doses in one cycle during P3. We use the standard set of parameters. In (a) and (c), we also plot the optimal number of TMZ doses ($n = 6$ and $n = 3$, respectively) and the corresponding tumor size in red.

type-3) cells, as functions of the number of doses per cycle. We observe that the tumor is dominated by non-methylated cells for all n, and the large population of TMZ-resistant cells makes a large number of TMZ doses less effective. Additionally while $n = 3$ is the optimal dose number in this case, it is not significantly more beneficial than no adjuvant TMZ treatment. Such behavior is consistent with clinical observations; the study in [16] found that unmethylated tumors treated with radiotherapy and the standard TMZ regimen had a median overall survival of 12.7 months, versus a median overall survival of 11.8 months for those receiving only radiotherapy. Thus, our model suggests that tumors with low levels of methylation at diagnosis may be better served by alternative therapies, such as O^6-benzylguanine discussed in [103], that can be used in combination with TMZ, to counter TMZ's impact on the methylation process.

Note that a few studies have suggested that there may be a phenomenon of MGMT depletion after cells are exposed to TMZ for an extended period of time, in an attempt to explain observed differences in dose-dense TMZ treatment and the standard TMZ regimen [104, 105]. If this phenomenon occurs, it could increase the benefit of a larger number of small doses of TMZ or generally make TMZ more effective in tumors with low methylation levels. However, other studies have not found any conclusive differences in dose-dense TMZ regimens and the standard TMZ regimen for either MGMT methylated or unmethylated tumors [106, 107]. Thus, due to inconclusive evidence of MGMT depletion after prolonged exposure to TMZ, we have not incorporated such a mechanism in our model.

4.5 Discussion

In this chapter we investigated the interplay between MGMT promoter methylation and TMZ during the treatment of glioblastoma. We developed a mathematical model integrating a mechanistic description of MGMT promoter methylation/demethylation with the evolutionary dynamics of GBM treatment response, contributing a unique perspective to the existing body of literature modeling GBM treatment. In particular, we considered a stochastic branching process model, consisting of distinct cellular subtypes with a methylated, hemimethylated, and unmethylated characteristic CpG site within the MGMT promoter. The model, parameterized using clinical and experimental data, incorporates standard GBM treatment methods. DNA methylation dynamics, inspired by the model in [1], drive the conversion rates between cellular subtypes.

Several clinical studies tracking the methylation status of GBM at initial detection and at recurrence, following standard treatment, have found that the majority of patients with

methylated primary tumors present unmethylated recurrent tumors. Our model results indicate that this clinically observed drop in methylation between diagnosis and recurrence cannot be explained simply by evolutionary selection. Decreasing the rate of maintenance methylation in the presence of TMZ results in a sizable reduction in expected methylation percentage at recurrence, consistent with clinical results. This suggests a hypothesis that TMZ may actively inhibit maintenance methylation, driving the downward shift in methylation between tumor diagnosis and recurrence.

The precise mechanism by which TMZ may contribute to MGMT demethylation is unclear, but experimental studies suggest this may involve the activation of the protein kinase C (PKC) signaling pathway. In [108] it was demonstrated that alkylating drugs similar to TMZ led to an increase in MGMT expression and in PKC activity. In [109], the authors discovered that a number of PKC isoforms induce the attachment of a phosphoryl group to DNMT1. Further testing on the specific isoform PKCζ showed that cells with a high expression of both PKCζ and DNMT1 exhibited a significant reduction in methylation; this was not the case in cells with a high expression of PKCζ or DNMT1 alone. This suggested that the methylation reduction results from the phosphorylation of DNMT1, driven by PKCζ; another study in [110] confirms that the phosphorylation of DNMT1 is associated with hypomethylation of gene promoters. Hence, experimental studies suggest that TMZ may contribute to MGMT demethylation by activating the PKC signaling pathway in GBM cells, leading to the phosphorylation of DNMT1, thereby inhibiting maintenance methylation within the affected cells, as our model suggests.

After incorporating a TMZ-mediated inhibition in maintenance methylation, we also used the model to find the optimal number of TMZ doses administered during adjuvant chemotherapy. We varied the number of daily TMZ doses administered during each 28-day cycle while maintaining the same cumulative dosage per cycle, to determine the dose number that minimizes the mean tumor population after 4 adjuvant cycles. Using our baseline parameter set, we determined an optimal TMZ dosing schedule of 6 daily doses of 166.67 mg/m^2, followed by 22 days off. The standard schedule of 5 daily doses/cycle is nearly optimal, resulting in a slightly larger mean tumor size after 4 cycles.

Due to inter-patient variability in methylation percentage at diagnosis, we also investigated the optimal adjuvant chemotherapy schedule for a tumor with a low methylation percentage at diagnosis. Receiving three larger doses of TMZ is optimal in this case, but it does not provide a significant benefit over the absence of any adjuvant TMZ treatment. This observation is consistent with clinical results comparing the benefit of both radiotherapy and chemotherapy versus radiotherapy alone for unmethylated primary tumors. Therefore,

when such tumors are detected, our model suggests that it may be more beneficial to administer, in combination with TMZ, a therapy that can counteract TMZ's inhibition of maintenance methylation, by stimulating MGMT methylation within the tumor.

A limitation of our model is that we do not incorporate the distant spread of GBM cells or the diffuse nature of GBM tumors. We also assume that all hemimethylated and unmethylated GBM cells behave with the same intrinsic growth rates and that MGMT methylation status does not affect radiosensitivity. Despite these assumptions, the model still provides useful insight regarding the relationship between MGMT methylation, TMZ, and sensitivity to chemotherapy. In the future, we are interested in exploring the role of the IDH1 mutation, an oncogenic mutation that changes the function of enzymes used in a cell's mitochondria, and investigating the hypothesis that the IDH1 mutation drives increased methylation in gliomas, leading to TMZ sensitivity [111]. Additionally, we hope to investigate the role of the stem cell marker $CD133^+$ and its impact on the evolution of TMZ resistance in GBM.

References

[1] L.B. Sontag, M.C. Lorincz, and E.G.Luebeck. Dynamics, stability and inheritance of somatic DNA methylation imprints. *Journal of Theoretical Biology*, 242(4):890 – 899, 2006.

[2] R. Stupp, W. P. Mason, M. J. van den Bent, et al. Radiotherapy plus concomitant and adjuvant temozolomide for glioblastoma. *New England Journal of Medicine*, 352(10):987–996, 2005.

[3] D Hanahan and R Weinberg. The hallmarks of cancer. *Cell*, 100:57–70, 2000.

[4] R. Durrett and S. Moseley. Spatial Moran models I. stochastic tunneling in the neutral case. *The Annals of Applied Probability*, 25(1):104–115, 2015.

[5] R. Durrett, J. Foo, and K. Leder. Spatial Moran models II. Cancer initiation in spatially structured tissue. *Journal of Mathematical Biology*, 72(5):1369–1400, 2016.

[6] J. Foo, K. Leder, and M.D. Ryser. Multifocality and recurrence risk: a quantitative model of field cancerization. *Journal of theoretical biology*, 355:170–184, 2014.

[7] K. Storey, M.D. Ryser, K. Leder, and J. Foo. Spatial measures of genetic heterogeneity during carcinogenesis. *Bulletin of mathematical biology*, 79(2):237–276, 2017.

[8] R. Durrett. Ten lectures on particle systems. In *St. Flour lecture notes*, volume 1608, pages 97–201. Springer-Verlag, New York, 1995.

[9] Danely P Slaughter, Harry W Southwick, and Walter Smejkal. Field cancerization in oral stratified squamous epithelium; clinical implications of multicentric origin. *Cancer*, 6(5):963–968, 1953.

[10] AM Kaz, WM Grady, MD Stachler, and AJ Bass. Genetic and epigenetic alterations in barrett's esophagus and esophageal adenocarcinoma. *Clin. North Am.*, 44(2):473–489, 2015.

[11] BA Virnig, TM Tuttle, T Shamliyan, and RL Kane. Ductal carcinoma in situ of the breast: a systematic review of incidence, treatment, and outcomes. *J. Natl Cancer Inst*, 102(3):170–178, 2010.

[12] DG Bostwick and L Cheng. Precursors of prostate cancer. *Histopathology*, 60:4–27, 2012.

[13] B. Braakhuis, M. Tabor, J. Kummer, C. Leemans, and R. Brakenhoff. A genetic explanation of Slaughter's concept of field cancerization evidence and clinical implications. *Cancer Research*, 63(8):1727–1730, 2003.

[14] H. Chai and R. Brown. Field effect in cancer–an update. *Annals of Clinical & Laboratory Science*, 39(4):331–337, 2009.

[15] K Curtius, NA Wright, and TA Graham. An evolutionary perspective on field cancerization. *Nat Rev Cancer*, 18(1):19–32, 2018.

[16] ME Hegi, AC Diserens, T Gorlia, et al. MGMT gene silencing and benefit from temozolomide in glioblastoma. *N Engl J Med*, 352(10):997–1003, 2005.

[17] J. Zhang, M. Stevens, and T. Bradshaw. Temozolomide: Mechanisms of action, repair and resistance. *Current Molecular Pharmacology*, 5(1):102–114, 2012.

[18] G.J. Kitange, B.L. Carlson, M.A. Schroeder, et al. Induction of MGMT expression is associated with temozolomide resistance in glioblastoma xenografts. *Neuro-Oncology*, 11(3):281–291, 2009.

[19] AA Brandes, E Franceschi, A. Tosoni, et al. O6-methylguanine DNA-methyltransferase methylation status can change between first surgery for newly diagnosed glioblastoma and second surgery for recurrence: clinical implications. *Neuro-Oncology*, 12(3):283–288, 2010.

[20] T. Suzuki, M. Nakada, Y. Yoshida, E. Nambu, N. Furuyama, D. Kita, Y. Hayashi, Y. Hayashi, and J. Hamada. The correlation between promoter methylation status and the expression level of the O6-methylguanine-DNA methyltransferase in recurrent glioma. *Jpn J Clin Oncol*, 41(2):190–196, 2011.

[21] N. Rahman. Realizing the promise of cancer predisposition genes. *Nature*, 505(7483):302–308, 2014.

[22] C.P. Wild, A. Scalbert, and Z. Herceg. Measuring the exposome: a powerful basis for evaluating environmental exposures and cancer risk. *Environmental and molecular mutagenesis*, 54(7):480–499, 2013.

[23] R.J. Gillies and R.A. Gatenby. Metabolism and its sequelae in cancer evolution and therapy. *The Cancer Journal*, 21(2):88–96, 2015.

[24] I. Bozic, T. Antal, H. Ohtsuki, and et. al. Accumulation of driver and passenger mutations during tumor progression. *Proceedings of the National Academy of Sciences*, 107(43):18545–18550, 2010.

[25] N. McGranahan and C. Swanton. Biological and therapeutic impact of intratumor heterogeneity in cancer evolution. *Cancer cell*, 27(1):15–26, 2015.

[26] C.C. Maley, P.C. Galipeau, J.C. Finley, and et. al. Genetic clonal diversity predicts progression to esophageal adenocarcinoma. *Nature genetics*, 38(4):468–473, 2006.

[27] R.H Whittaker. Evolution and measurement of species diversity. *Taxon*, pages 213–251, 1972.

[28] Y. Iwasa and F. Michor. Evolutionary dynamics of intratumor heterogeneity. *PLoS One*, 6:e17866, 2011.

[29] R. Durrett, J. Foo, K. Leder, J. Mayberry, and F. Michor. Intratumor heterogeneity in evolutionary models of tumor progression. *Genetics*, 188:461–477, 2011.

[30] A. Dhawan, T.A. Graham, and A.G. Fletcher. A computational modelling approach for deriving biomarkers to predict cancer risk in premalignant disease. *bioRxiv*, page 020222, 2015.

[31] T. Williams and R. Bjerknes. Stochastic model for abnormal clone spread through epithelial basal layer. *Nature*, 236:19–21, 1972.

[32] M. Nowak, Y. Michor, and Y. Iwasa. The linear process of somatic evolution. *PNAS*, 100:14966–14969, 2003.

[33] N. Komarova. Spatial stochastic models for cancer initiation and progression. *Bull. Math. Biol.*, 68:1573–1599, 2006.

[34] C. Thalhauser, J. Lowengrub, D. Stupack, and N. Komarova. Selection in spatial stochastic models of cancer: Migration as a key modulator of fitness. *Biology Direct*, 5:21, 2010.

[35] N. Komarova. Spatial stochastic models of cancer: Fitness, migration, invasion. *Mathematical Biosciences and Engineering*, 10:761–775, 2013.

[36] K. Sprouffske, J.W. Pepper, and C.C. Maley. Accurate reconstruction of the temporal order of mutations in neoplastic progression. *Cancer Prevention Research*, 4(7):1135–1144, 2011.

[37] R Durrett. *Lecture Notes on Particle Systems and Percolation*. Wadsworth and Brooks/Cole Advanced Books and Software, Pacific Grove, Calif, 1988.

[38] Annemieke de Vries, Elsa R. Flores, Barbara Miranda, Harn-Mei Hsieh, Conny Th. M. van Oostrom, Julien Sage, and Tyler Jacks. Targeted point mutations of p53 lead to dominant-negative inhibition of wild-type p53 function. *Proceedings of the National Academy of Sciences of the United States of America*, 99(5):2948–2953, 2002.

[39] M. Bramson and D. Griffeath. On the Williams-Bjerknes tumor growth model: II. *Mathematical Proceedings of the Cambridge Philosophical Society*, 88:339–357, 1980.

[40] M. Bramson and D. Griffeath. On the Williams-Bjerknes tumour growth model: I. *Annals of Probability*, 9:173–185, 1981.

[41] MA Kara, FP Peters, FJ ten Kate, van Deventer SJ, P Fockens, and JJ Bergman. Endoscopic video autofluorescence imaging may improve the detection of early neoplasia in patients with barretts esophagus. *Gastrointest Endosc*, 61, 2005.

[42] Nachmann M and Crowell S. Estimate of the mutation rate per nucleotide in humans. *Genetics*, 156(1):287–304, 2000.

[43] S Kumar and S Subramanian. Mutation rates in mammalian genomes. *PNAS*, 99(2):803–808, 2002.

[44] CF Baer, MM Miyamoto, and DR Denver. Mutation rate variation in multicellular eukaryotes: causes and consequences. *Nature Rev. Genet.*, 8:619?631, 2007.

[45] Brouwer AF, Eisenberg MC, and Meza R. Age effects and temporal trends in hpv-related and hpv-unrelated oral cancer in the united states: A multistage carcinogenesis modeling analysis. *PLoS ONE*, 11(3), 2016.

[46] K Curtius, C-J Wong, WD Hazelton, AM Kaz, Chak A, and et al. Willis, JE. A molecular clock infers heterogeneous tissue age among patients with barretts esophagus. *PLoS Comput Biol*, 12(5), 2016.

[47] P.A.P. Moran. Random processes in genetics. *Proc. Cambridge Philos. Soc.*, 54:60–71, 1958.

[48] M. A. Nowak, N. L. Komarova, A. Sengupta, P. V. Jallepalli, I. M. Shih, B. Vogelstein, and C. Lengauer. The role of chromosomal instability in tumor initiation. *Proc. Natl. Acad. Sci.*, 99(25):16226–16231, 2002.

[49] R. Durrett, D. Schmidt, and J. Schweinsberg. A waiting time problem arising from the study of multi-stage carcinogenesis. *The Annals of Applied Probability*, 19(2):676–718, 2009.

[50] J. Schweinsberg. Waiting for n mutations. *Electronic Journal of Probability*, 13:1442–1478, 2008.

[51] N.L. Komarova, A. Sengupta, and M.A. Nowak. Mutation–selection networks of cancer initiation: tumor suppressor genes and chromosomal instability. *Journal of Theoretical Biology*, 223(4):433–450, 2003.

[52] Y. Iwasa, F. Michor, and M.A. Nowak. Stochastic tunnels in evolutionary dynamics. *Genetics*, 166:1571–1579, 2004.

[53] Y. Iwasa, F. Michor, N. Komarova, and M. Nowak. Population genetics of tumor suppressor genes. *Journal of Theoretical Biology*, 233:15–23, 2005.

[54] NIH National Cancer Institute. SEER training cancer classification. https://training.seer.cancer.gov/disease/categories/classification.html.

[55] R. Durrett, J. Foo, and K. Leder. Spatial moran models, II: Cancer initiation in spatially structured tissue. *Journal of Mathematical Biology*, 72(5):1369–1400, 2016.

[56] J. Foo, K. Leder, and M.D. Ryser. Multifocality and recurrence risk: a quantitative model of field cancerization. *Journal of theoretical biology*, 355:170–184, 2014.

[57] T. Williams and R. Bjerknes. Stochastic model for abnormal clone spread through epithelial basal layer. *Nature*, 236:19–21, 1972.

[58] N. Komarova. Spatial stochastic models for cancer initiation and progression. *Bulletin of Mathematical Biology*, 68:15731599, 2006.

[59] S. Durrett, R. Moseley. Spatial Moran models I. stochastic tunneling in the neutral case. *Annals of applied probability*, 25(1):104–115, 2015.

[60] D. Griffeath. Additive and cancellative interacting particle systems. In *Lecture notes in mathematics*, volume 724. Springer, New York, 1978.

[61] P. Billingsley. *Convergence of probability measures.* Wiley, New York, 1968.

[62] R. Durrett. Oriented percolation in two dimensions. *Annals of Probability*, 12(4):999–1040, 1984.

[63] K. Geboes. Squamous mucosa and reflux. `http://www.hon.ch/OESO/books/Vol_3_Eso_Mucosa/Articles/ART004.HTML`, 1994.

[64] CA Squier and MJ Kremer. Biology of oral mucosa and esophagus. *JNCI Monographs*, 2001(29):7–15, 2001.

[65] PA Coulombe, R Kopan, and E Fuchs. Expression of keratin k14 in the epidermis and hair follicle: Insights into complex programs of differentiation. *The Journal of Cell Biology*, 109:2295–2312, 1989.

[66] LG Komuves, K Hanley, M Man, PM Elias, ML Williams, and KR Feingold. Keratinocyte differentiation in hyperproliferative epidermis: Topical application of ppar activators restores tissue homeostasis. *Journal of Investigative Dermatology*, 115(3):361–367, 2000.

[67] A. Knudson. Two genetic hits (more or less) to cancer. *Nature Reviews Cancer*, 1:157–161, 2001.

[68] T. Maruyama. A simple proof that certain quantities are independent of the geographical structure of population. *Theor Pop Biol*, 5:148–154, 1974.

[69] JK Park, T Hodges, L Arko, et al. Scale to predict survival after surgery for recurrent glioblastoma multiforme. *Journal of clinical oncology*, 28(24):3838–3843, 2010.

[70] O Gallego. Nonsurgical treatment of recurrent glioblastoma. *Current oncology*, 22(4):e273, 2015.

[71] AB Håvik, P Brandal, H Honne, HS Dahlback, D Scheie, M Hektoen, TR Meling, E Helseth, S Heim, RA Lothe, and GE Lind. MGMT promoter methylation in gliomas-assessment by pyrosequencing and quantitative methylation-specific PCR. *Journal of Translational Medicine*, 10(36), 2012.

[72] R Brown, E Curry, L Magnani, CS Wilhelm-Benartzi, and J Borley. Poised epigenetic states and acquired drug resistance in cancer. *Nature Reviews Cancer*, 14(11):747, 2014.

[73] T. Jung, S. Jung, K. Moon, et al. Changes of the o6-methylguanine-dna methyltransferase promoter methylation and MGMT protein expression after adjuvant treatment in glioblastoma. *Oncology Reports*, 23:1269–1276, 2010.

[74] M. Christmann, G. Nagel, S. Horn, et al. MGMT activity, promoter methylation and immunohistochemistry of pretreatment and recurrent malignant gliomas: a comparative study on astrocytoma and glioblastoma. *Int. J. Cancer*, 127:2106–2118, 2010.

[75] V.A. Levin, C.S. Patlak, and H.D. Landahl. Heuristic modeling of drug delivery to malignant brain tumors. *J. Pharmacokinet. Biopharm.*, 8(3):257–296, 1980.

[76] G.S. Stamatakos, V.P. Antipas, and N.K. Uzunoglu. A spatiotemporal, patient individualized simulation model of solid tumor response to chemotherapy in vivo: the paradigm of glioblastoma multiforme treated by temozolomide. *IEEE Transactions on Biomedical Engineering*, 53(8):1467–1477, 2006.

[77] MA Böttcher, J Held-Feindt, M Synowitz, R Lucius, A Traulsen, and K Hattermann. Modeling treatment-dependent glioma growth including a dormant tumor cell subpopulation. *BMC Cancer*, 18:376, 2018.

[78] J. Fowler. The linear-quadratic formula and progress in fractionated radiotherapy. *British Journal of Radiology*, 62(740):679–694, 1989.

[79] D. Brenner. The linear-quadratic model is an appropriate methodology for determining isoeffective doses at large doses per fraction. *Seminars in Radiation Oncology*, 18:234–239, 2008.

[80] M. Zaider and G.N. Minerbo. Tumour control probability: a formulation applicable to any temporal protocol of dose delivery. *Physics in Medicine and Biology*, 45(2):279–293, 2000.

[81] H. Badri, K. Pitter, E. Holland, F. Michor, and K. Leder. Optimization of radiation dosing schedules for proneural glioblastoma. *Journal of Mathematical Biology*, 72(5):1–36, 2016.

[82] G Powathil, M Kohandel, S Sivaloganathan, A Oza, and M Milosevic. Mathematical modeling of brain tumors: effects of radiotherapy and chemotherapy. *Physics in Medicine and Biology*, 52(11):3291–3306, 2007.

[83] R Rockne, JK Rockhill, M Mrugala, AM Spence, I Kalet K Hendrickson, A Lai, T Cloughesy, EC Alvord Jr, and KR Swanson. Predicting the efficacy of radiotherapy in individual glioblastoma patients in vivo: a mathematical modeling approach. *Physics in Medicine and Biology*, 55(12):3271, 2010.

[84] D Corwin, C Holdsworth, RC Rockne, AD Trister, MM Mrugala, JK Rockhill, RD Stewart, M Phillips, and KR Swanson. Toward patient-specific, biologically optimized radiation therapy plans for the treatment of glioblastoma. *PLoS ONE*, 8(11):1–9, 2013.

[85] H Hatzikirou, A Deutsch, C Schaller, M Md, K Swanson, N Bellomo, and P Maini. Mathematical modelling of glioblastoma tumour development: a review. *Mathematical Models and Methods in Applied Sciences*, 24:1779–1794, 11 2005.

[86] Y. Yatabe, S. Tavare, and D. Shibata. Investigating stem cells in human colon by using methylation patterns. *Proc. Natl. Acad. Sci.*, 98(19):10839–10844, 2001.

[87] S.P. Otto and V. Walbot. DNA methylation in eukaryotes: kinetics of demethylation and de novo methylation during the life cycle. *Genetics*, 124(2):429–437, 1990.

[88] G.P. Pfeifer, S.D. Steigerwald, R.S. Hansen, S.M. Gartler, and A.D. Riggs. Polymerase chain reaction-aided genomic sequencing of an x chromosome-linked CpG island: methylation patterns suggest clonal inheritance, CpG site autonomy, and an explanation of activity state stability. *Proc. Natl Acad. Sci. USA*, 87(21):8252–8256, 1990.

[89] D.P. Genereux, B.E. Miner, C.T. Bergstrom, and C.D. Laird. A population-epigenetic model to infer site-specific methylation rates from double-stranded DNA methylation pattern. *Proc. Natl Acad. Sci. USA*, 102(16):5802–5807, 2005.

[90] S. Saxonov, P. Berg, and D.L. Brutlag. A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. *Proc. Natl Acad. Sci. USA*, 103(5):14121417, 2006.

[91] B Jin, Y Li, and KD Robertson. DNA methylation: Superior or subordinate in the epigenetic hierarchy? *Genes and Cancer*, 2(6):607–617, 2011.

[92] G.D. Kim, J. Ni, N. Kelesoglu, R.J. Roberts, and S Pradhan. Co-operation and communication between the human maintenance and de novo DNA (cytosine-5) methyltransferase. *EMBO J*, 21(15):4183–4195, 2002.

[93] T. Chen, Y. Ueda, J.E. Dodge, Z. Wang, and E. Li. Establishment and maintenance of genomic methylation patterns in mouse embryonic stem cells by Dnmt3a and Dnmt3b. *Mol. Cell. Biol*, 23(16):5594–5605, 2003.

[94] G. Vilkaitis, I. Suetake, S. Klimasauskas, and S. Tajima. Processive methylation of hemimethylated CpG sites by mouse Dnmt1 DNAmethyltransferase. *J. Biol. Chem.*, 280(1):64–72, 2005.

[95] M Tahiliani, KP Koh, Y Shen, et al. Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. *Science*, 324(5929):930–935, 2009.

[96] M Esteller, J Garcia-Foncillas, E Andion, et al. Inactivation of the DNA-repair gene MGMT and the clinical response of gliomas to alkylating agents. *New England Journal of Medicine*, 343(19):1350–1354, 2000.

[97] K. Zhang, X. Wang, B. Zhou, et al. The prognostic value of MGMT promoter methylation in glioblastoma multiforme: a meta-analysis. *Familial Cancer*, 12(3):449–458, 2013.

[98] N Shah, B Lin, Z Sibenaller, et al. Comprehensive analysis of MGMT promoter methylation: Correlation with MGMT expression and clinical response in GBM. *PLoS ONE*, 6(1):e16146, 2011.

[99] K. Athreya and P. Ney. *Branching Processes*. Springer-Verlag, New York, 1972.

[100] NA Franken, S Hovingh, H Rodermond, Stalpers L, Barendsen GW, and Crezee J. Radiosensitization with chemotherapeutic agents and hyperthermia: Effects on linear-quadratic parameters of radiation cell survival curves. *J Cancer Sci Ther*, S5(002), 2011.

[101] D Corwin, C Holdsworth, RC Rockne, AD Trister, MM Mrugala, et al. Toward patient-specific, biologically optimized radiation therapy plans for the treatment of glioblastoma. *PLOS ONE*, 8(11):e79115, 2013.

[102] K Ogura, T Mizowaki, Y Arakawa, M Ogura, K Sakanaka, S Miyamoto, and M Hiraoka. Initial and cumulative recurrence patterns of glioblastoma after temozolomide-based chemoradiotherapy and salvage treatment: a retrospective cohort study in a single institution. *Radiation Oncology*, 8(1):97, 2013.

[103] J.E. Adair, S.K. Johnston, M.M. Mrugala, B.C. Beard, L.A. Guyman, A.L. Baldock, et al. Gene therapy enhances chemotherapy tolerance and efficacy in glioblastoma patients. *Journal of Clinical Investigation*, 124(9):4082–4092, 2014.

[104] AA Brandes, A Tosoni, G Cavallo, et al. Temozolomide 3 weeks on and 1 week off as first-line therapy for recurrent glioblastoma: phase II study from gruppo italiano cooperativo di neuro-oncologia (GICNO). *British Journal of Cancer*, 95:1155–1160, 2006.

[105] W. Wick, M. Platten, and M. Weller. New (alternative) temozolomide regimens for the treatment of glioma. *Neuro-Oncology*, 11(1):69–79, 2009.

[106] M.P. Mehta, M. Wang, K. Aldape, R. Stupp, K.A. Jaeckle, D. Blumenthal, P. Brown, S. Erridge, W. Curran, and M. Gilbert. Rtog 0525: Exploratory subset analysis from a randomized phase III trial comparing standard adjuvant temozolomide with a dose-dense schedule for glioblastoma. *International Journal of Radiation Oncology*Biology*Physics*, 81(2):S128–S129, 2011.

[107] MR Gilbert, M Wang, KD Aldape, et al. Dose-dense temozolomide for newly diagnosed glioblastoma: A randomized phase III clinical trial. *Journal of Clinical Oncology*, 31(32):4085–4091, 2013.

[108] I Boldogh, CV Ramana, Z Chen, et al. Regulation of expression of the DNA repair gene o6-methylguanine-DNA methyltransferase via protein kinase C-mediated signaling. *Cancer Research*, 58:3950–3956, 1998.

[109] G Lavoie, P-O Estve, NB Laulan, S Pradhan, and Y St-Pierre. PKC isoforms interact with and phosphorylate DNMT1. *BMC Biology*, 9(31):doi:10.1186/1741–7007–9–31, 2011.

[110] E Hervouet, L Lalier, E Debien, et al. Disruption of Dnmt1/PCNA/UHRF1 interactions promotes tumorigenesis from human and mice glial cells. *PLoS One*, 5(6):e11333, 2010.

[111] R.J. Molenaar, D. Verbaan, S. Lamba, C. Zanon, J.W.M. Jeuken, S.H.E. Boots-Sprenger, P. Wesseling, T.J.M. Hulsebos, D. Troost, A.A. van Tilborg, S. Leenstra, W.P. Vandertop, A. Bardelli, C.J.F. van Noorden, and F. E. Bleeker. The combination of IDH1 mutations and MGMT methylation status predicts survival in glioblastoma better than either IDH1 or MGMT alone. *Neuro-Oncology*, 2014.

[112] J. Pitman and N.M. Tran. Size biased permutations of a finite sequence with independent and identically distributed terms. *Bernoulli*, 21:2484–2512, 2012.

[113] M.P. Fewell. Area of common overlap of three circles. Technical Report DSTO-TN-0722, Australian Government Defence Science and Technology Organization, 2006.

[114] S.P. Lalley. Convergence rates of markov chains. 2013. URL: `https://galton.uchicago.edu/~lalley/Courses/313/ConvergenceRates.pdf`.

[115] J. Gravner. Lecture notes for introductory probability. chapter 15. University of California, Davis, 2010.

[116] G. Lawler and V. Limic. *Random walk: a modern introduction*. Cambridge University Press, New York, 2010.

[117] W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):3–13, 1963.

[118] C. Cannone. A short note on Poisson tail bounds. 2017. URL: `http://www.cs.columbia.edu/~ccanonne/files/misc/2017-poissonconcentration.pdf`.

[119] A. Dvoretzky and P. Erdos. Some problems on random walk in space. In *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability, 1950*, pages 353–367. University of California Press, Berkeley and Los Angeles, 1951.

[120] R. Durrett and I. Zahle. On the width of hybrid zones. *Stochastic Processes and their Applications*, 117(12):1751–1763, 2007.

[121] Y. Fukai and K. Uchiyama. Potential kernel for two-dimensional random walk. *Ann. Probab.*, 24(4):1979–1992, 1996.

[122] M. Bramson and R. Durrett. A simple proof of the stability criterion of Gray and Griffeath. *Probab. Th. Rel. Fields*, 80:293–298, 1988.

[123] D Richardson. Random growth in a tesselation. *Proc. Cambridge Philos. Soc.*, 74:515–528, 1973.

[124] R Durrett and D Griffeath. Contact processes in several dimensions. *Z. Wahrscheinlichkeitstheorie verw. Gebiete*, 59:535–552, 1982.

[125] K. James, E. Eisenhauer, M. Christian, M. Terenziani, D. Vena, A. Muldal, and P. Therasse. Measuring response in solid tumors: Unidimensional versus bidimensional measurement. *Journal of the National Cancer Institute*, 91(6):523–528, 1999.

[126] YJ Hong, P Marjoram, D Shibata, and KD Siegmund. Using DNA methylation patterns to infer tumor ancestry. *PLOS ONE*, 5(8):1–9, 2010.

[127] GP Pfeifer, SD Steigerwald, RS Hansen, SM Gartler, and AD Riggs. Polymerase chain reaction-aided genomic sequencing of an X chromosome-linked CpG island: Methylation patterns suggest clonal inheritance, CpG site autonomy, and an explanation of activity state stability. *Proc. Nati. Acad. Sci.*, 87:8252–8256, 1990.

[128] C.D. Laird, N.D. Pleasant, A.D. Clark, et al. Hairpin-bisulfite PCR: assessing epigenetic methylation patterns on complementary strands of individual DNA molecules. *Proc. Natl Acad. Sci. USA*, 101(1):204–209, 2004.

[129] H.D. Thames, S.M. Bentzen, I. Turesson, M. Overgaard, and W. Van den Bogaert. Time-dose factors in radiotherapy: a review of the human data. *Radiotherapy and Oncology*, 19(3):219–235, 1990.

[130] B Jones and P. Sanghera. Estimation of radiobiologic parameters and equivalent radiation dose of cytotoxic chemotherapy in malignant glioma. *International Journal of Radiation Oncology, Biology, Physics*, 68(2):441–448, 2007.

[131] L. Hammond, J. Eckardt, S. Baker, et al. Phase I and pharmacokinetic study of temozolomide on a daily-for-5-days schedule in patients with advanced solid malignancies. *J Clin Oncol*, 17:2604–2613, 1999.

[132] M. Rudek, R. Donehower, P. Statkevich, V. Batra, D. Cutler, and S. Baker. Temozolomide in patients with advanced cancer: phase i and pharmacokinetic study. *Pharmacotherapy*, 24(1):16–25, 2004.

[133] J. Portnow, B. Badie, M. Chen, A. Liu, S. Blanchard, and T.W. Synold. The neuropharmacokinetics of temozolomide in patients with resectable brain tumors: Potential implications for the current approach to chemoradiation. *Clinical cancer research:*

an official journal of the American Association for Cancer Research, 15(22):7092–7098, 2009.

[134] C.S. Brock, E.S. Newlands, S.R. Wedge, et al. Phase I trial of temozolomide using an extended continuous oral schedule. *Cancer Research*, 58(19):4363–4367, 1998.

[135] C.D. Britten, E.K. Rowinsky, S.D. Baker, et al. A phase I and pharmacokinetic study of temozolomide and cisplatin in patients with advanced solid malignancies. *Clinical Cancer Research*, 5(7):1629–1637, 1999.

Appendix A

Chapter 2 Appendix

A.1 Details for non-spatial Simpson's Index

A.1.1 Preliminary definitions and results

To characterize the distributions of the Simpson's Index, we introduce two definitions. Let $L_1, L_2, \ldots, L_n$ be independent, identically distributed random variables with distribution F. Recall the definition of a sized-biased pick from Definition 1 in Section 2.3.2. Then we define a *size-biased permutation* as follows.

Definition 2. *We call* $\left(L_{[1]}, \ldots L_{[n]}\right)$ *a size-biased permutation (s.b.p) of the sequence* (L_i) *if* $L_{[1]}$ *is a size-biased pick of the sequence, and for* $2 \leq k \leq n$,

$$\mathbb{P}\left(L_{[k]} = L_j \mid L_{[1]}, \ldots, L_{[k-1]}; L_1, \ldots L_n\right) = \frac{L_j \, \mathbb{1}_{(L_j \neq L_{[i]}, \forall 1 \leq i < k)}}{L_1 + \ldots + L_n - \left(L_{[1]} + \ldots + L_{[k-1]}\right)}.$$

The following results will be useful.

Proposition A.1.1. *(Proposition 2 in [112]) For* $1 \leq k \leq n$, *let* ν_k *be the density of* S_k, *the sum of* k *i.i.d random variables with distribution* F. *Then*

$$\begin{aligned}\mathbb{P}\left(X_{[1]} \in dx_1, \ldots, X_{[k]} \in dx_k\right) =& \frac{n!}{(n-k)!} \left(\prod_{j=1}^{k} x_j \nu_1(x_j) dx_j\right) \ldots \\ & \ldots \int_0^\infty \nu_{n-k}(s) \prod_{j=1}^{k} \left(x_j + \ldots + x_k + s\right)^{-1} ds. \qquad \text{(A.1)}\end{aligned}$$

□

Corollary A.1.2. *(Corollary 3 in [112]) Let $T_{n-k} = X_{[k+1]} + \ldots + X_{[n]}$ denote the sum of the last $n-k$ terms in an i.i.d s.b.p of length n. Then for $k = 1, \ldots, n-1$,*

$$\mathbb{P}\left(T_{n-k} \in ds \mid T_{n-k+1} = t\right) = (n-k+1)\,\frac{t-s}{t}\nu_1(t-s)\frac{\nu_{n-k}(s)}{\nu_{n-k+1}(t)}ds. \tag{A.2}$$

□

A.1.2 Conditional expectation

Recalling that type 1 clones have a linear radial growth rate, Simpson's index (2.4) can be rewritten explicitly

$$R(t) = \sum_{i=1}^{N_t}\left[\frac{(1-t_i/t)^d}{\sum_{j=1}^{N_t}(1-t_j/t)^d}\right]^2, \tag{A.3}$$

where $\{t_i\}_{i=1}^{N_t}$ are the points of a Poisson process with constant intensity Nu_1st. In particular, we note that conditioned on N_t, the t_i are i.i.d and

$$(t_1/t|N_t) \sim U(0,1).$$

We now define $X_i := (1-t_i/t)^{d+1}$ and let

$$T := \sum_{i=1}^{N_t} X_i, \tag{A.4}$$

which allows us to rewrite (A.3) as

$$R(t) = \sum_{i=1}^{N_t}\left(\frac{X_i}{T}\right)^2. \tag{A.5}$$

Note that conditioned on N_t, the X_i are i.i.d with

$$(X_1|N_t) \sim Beta\left(\frac{1}{d}, 1\right).$$

To see this, note first that by symmetry $X_i \sim (t_i/t)^d$; using characteristic functions, it is then easy to verify that for $X \sim Beta\,(\alpha, 1)$ and $Y \sim U(0,1)$, we have $X \sim Y^n$ if and only if $\alpha = 1/n$. Using the above notation and recalling the notion of a size-biased pick in

Definition 1, we condition (A.5) on N_t to find

$$\begin{aligned}(R(t) \mid N_t = n) &= \sum_{i=1}^{n} \frac{X_i}{T} \mathbb{P}\left(X_{[1]} = X_i | X_1, \ldots, X_n\right) \\ &= \mathbb{E}\left(\frac{X_{[1]}}{T} \;\middle|\; X_1, \ldots, X_n\right). \end{aligned} \tag{A.6}$$

To compute the conditional expectation of Simpson's Index $R(t)$, we take the expectation of (A.6) to find

$$\mathbb{E}\left(R(t) \mid N_t = n\right) = \mathbb{E}\left(\frac{X_{[1]}}{T}\right) = \int_0^\infty \int_0^\infty \left(\frac{r}{x}\right) \mathbb{P}\left(X_{[1]} \in dr, T \in dx\right). \tag{A.7}$$

Setting $k = 1$, it follows now from Corollary A.1.2

$$\begin{aligned}\mathbb{E}\left(R(t) \mid N_t = n\right) &= \int_0^\infty \int_0^\infty \left(\frac{r}{x}\right) \mathbb{P}\left(T_{n-1} \in d(x-r), T \in dx\right) \\ &= \int_0^\infty \int_0^\infty \left(\frac{r}{x}\right) \mathbb{P}(T \in dx) \mathbb{P}\left(T_{n-1} \in d(x-r) \mid T = x\right) \\ &= n \int_0^\infty \int_0^\infty \left(\frac{r}{x}\right)^2 \nu_1(r)\, \nu_{n-1}(x-r)\, dx dr. \end{aligned} \tag{A.8}$$

Note that the support of ν_1 is over $[0, 1]$, and the support of ν_{n-1} is over $[0, n]$. Now, by definition, ν_1 is the pdf of $Beta\left(\frac{1}{d}, 1\right)$, i.e.

$$\nu_1(x) = \frac{1}{d} x^{\frac{1}{d}-1}.$$

On the other hand, ν_{n-1} is the density of the sum of $n-1$ i.i.d. $Beta(\frac{1}{d}, 1)$ random variables, i.e.

$$\nu_{n-1}(x) = \left(\nu_1^{*(n-1)}\right)(x).$$

For positive integer n let $S_n = B_1 + \ldots + B_n$ where B_i are independent $Beta(\frac{1}{d}, 1)$ random variables. Finally, from (A.8) we find

$$\mathbb{E}[R(t) | N_t = n] = n\mathbb{E}\left[\left(\frac{S_1}{S_n}\right)^2\right], \tag{A.9}$$

where $S_n := B_1 + \ldots + B_n$, and B_i are independent $Beta(\frac{1}{d}, 1)$ random variables.

A.1.3 Upper bound for variance

We derive an upper bound for the variance of the conditional Simpson's Index as follows:

$$\begin{aligned}
\left[\mathbb{E}\left(R(t) \mid N_t = n\right)\right]^2 \leq & \mathbb{E}\left(R^2(t) \mid N_t = n\right) = \mathbb{E}\left(\left[\mathbb{E}\left(\frac{X_{[1]}}{T} \;\middle|\; X_1, \ldots, X_n\right)\right]^2 \;\middle|\; N_t = n\right) \\
\leq & \mathbb{E}\left(\mathbb{E}\left(\left[\frac{X_{[1]}}{T}\right]^2 \;\middle|\; X_1, \ldots, X_n\right) \;\middle|\; N_t = n\right) \\
= & \mathbb{E}\left(\left[\frac{X_{[1]}}{T}\right]^2 \;\middle|\; N_t = n\right) \\
= & n \int_0^\infty \int_0^\infty \left(\frac{r}{x}\right)^3 \nu_1(r)\, \nu_{n-1}(x-r)\, dx\, dr, \qquad \text{(A.10)}
\end{aligned}$$

where the second to last equality follows from the fact that the sub-sigma algebra $\sigma(N_t = n)$ is coarser than $\sigma(X_1, \ldots, X_n)$.

A.1.4 Proof of Proposition 2.2.3

Proof. Let $Y_n = (R(t)|N(t) = n)$, and note that by definition $Y_n \geq 0$. Thus it suffices to show that $\mathbb{E}[Y_n] \to 0$ as $n \to \infty$. Note that

$$\mathbb{E}[Y_n] = \frac{1}{n}\mathbb{E}\left[\left(\frac{S_1}{S_n/n}\right)^2\right],$$

and by the law of large numbers $S_1/(S_n/n) \to S_1/\mathbb{E}[B_1]$ as $n \to \infty$. Thus if we establish that

$$\sup_{n<\infty} \mathbb{E}\left[\left(\frac{S_1}{S_n/n}\right)^3\right] < \infty, \qquad \text{(A.11)}$$

then by uniform integrability we will have $\mathbb{E}[(S_1/(S_n/n))^2] \to 1$, and thus $\mathbb{E}[Y_n] \to 0$.

In order to establish (A.11), we define $S_{2,n} = B_2 + \ldots + B_n$ and for $\varepsilon > 0$ the event

$$A_n = \{S_{2,n} > (1-\varepsilon)(n-1)\mathbb{E}[B_1]\}.$$

We then have that

$$\begin{aligned}\mathbb{E}\left[\left(\frac{S_1}{S_n/n}\right)^3\right] &= n^3\mathbb{E}\left[\left(\frac{S_1}{S_1+S_{2,n}}\right)^3\right] \\ &= n^3\mathbb{E}\left[\left(\frac{S_1}{S_1+S_{2,n}}\right)^3; A_n\right] + n^3\mathbb{E}\left[\left(\frac{S_1}{S_1+S_{2,n}}\right)^3; A_n^c\right] \\ &\leq O(1) + n^3\mathbb{P}(A_n^c).\end{aligned}$$

From Azuma-Hoeffding inequality we know that there exists a k independent of n such that $\mathbb{P}(A_n^c) \leq e^{-kn}$, thus establishing (A.11). ■

A.1.5 Monte Carlo Simulations

We evaluate the conditional expectation of Simpson's Index for fixed time t using Monte Carlo simulations. Based on the representation in (2.5), we first generate M independent copies of the vector (S_1, S_n) denoted by $\{(S_1^{(i)}, S_n^{(i)})\}_{i=1}^M$, and form the estimator

$$\hat{\mu}(n, M) = \frac{n}{M}\sum_{i=1}^{M}\left(\frac{S_1^{(i)}}{S_n^{(i)}}\right)^2,$$

which satisfies $\mathbb{E}[\hat{\mu}(n, M)] = \mathbb{E}[R(t)|N_t = n]$ and $Var[\hat{\mu}(n, M)] = O(1/M)$. If we simulate M_1 copies of N_t, denoted by $\{N_t^{(i)}\}_{i=1}^{M_1}$ and for each realization $N_t = n$ we form the estimator $\hat{\mu}(n, M_2)$, then we have an unbiased estimator of $\mathbb{E}[R(t)]$ via

$$\hat{R}(M_1, M_2) = \frac{1}{M_1}\sum_{j=1}^{M_1}\sum_{n=0}^{\infty} 1_{\{N_t^{(j)}=n\}}\hat{\mu}(n, M_2),$$

since $N_t^{(j)}$ is independent of $\hat{\mu}(n, M_2)$. Note that simulating the mesoscopic model M times and averaging $R(t)$ over those simulations is equivalent to using the estimator $\hat{R}(M, 1)$.

A.2 I_1 Calculations

Recall from Section 2.3.1 that $I(r, t)$ is approximated by (2.8). It is therefore necessary to calculate $\mathbb{P}(D_{ab} \cap E_1)$ and $\mathbb{P}(D_{ab} \cap E_2)$. Also recall that $V_x(t_0)$ is the space-time cone centered at x with radius $c_d t_0$ at time 0 and radius 0 at time t_0. For two points a and b in

our spatial domain we will be interested in the sets

$$D(r,t_0) = V_a(t_0)\Delta V_b(t_0)$$
$$M(r,t_0) = V_a(t_0)\cap V_b(t_0),$$

where $r = \|a-b\|$. We suppress the dependence on a and b in D and M to emphasize that the volume of these sets depends only on the distance $\|a-b\|$. Denote the Lebesgue measure of a set $A \in \mathbb{R}^d \times [0,\infty)$ by $|A|$. In order to calculate $I(r,t)$ it will be necessary to compute $|D(r,t_0)|$ and $|M(r,t_0)|$. Note that

$$|D(r,t_0)| = 2\left(|V_a(t_0)| - |M(r,t_0)|\right). \tag{A.12}$$

In the next two subsections we compute I_1 in one and two dimensions. For ease of notation we define $\mu = u_1 s$. For real number a, define $a^+ = \max\{a,0\}$.

A.2.1 I_1 in 1 dimension

We will first calculate the volumes $|V_a(t_0)|, |M(r,t_0)|$, and from (A.12), $|D(r,t_0)|$. In one dimension these calculations are simple: $|V_x(t_0)| = t_0^2 c_d$ and $|M(r,t_0)| = \frac{[(2t_0c_d - r)^+]^2}{4c_d}$, so we have that:

$$D(r,t_0) = 2t_0^2c_d - \frac{2[(2t_0c_d-r)^+]^2}{4c_d}.$$

Note that if $t_0 < r/(2c_d)$ then the only way sites a and b are the same at time t_0 is if there are zero mutations in $V_a(t_0)\cup V_b(t_0)$, i.e.,

$$I(r,t_0) = \exp\left(-\mu|V_a(t_0)\cup V_b(t_0)|\right) = \exp\left(-2\mu t_0^2 c_d\right).$$

Thus assume for the remainder of the subsection that $t_0 > r/(2c_d)$, in which case

$$|V_a(t_0)\cup V_b(t_0)| = 2t_0^2c_d - \frac{(2t_0c_d-r)^2}{4c_d}.$$

And since the mutations arise according to a Poisson process with parameter μ,

$$\mathbb{P}(E_k) = \frac{(\mu|V_a(t_0)\cup V_b(t_0)|)^k e^{-\mu|V_a(t_0)\cup V_b(t_0)|}}{k!}.$$

From (2.8) it remains to compute $P(D_{ab}|E_1)$ and $P(D_{ab}|E_2)$. Note that if event E_1

occurs, then D_{ab} can only occur if the single mutation occurs in the set $D(r, t_0)$, and therefore

$$\mathbb{P}(D_{ab}|E_1) = \frac{|D(r,t_0)|}{|V_a(t_0) \cup V_b(t_0)|} = 1 - \frac{(2t_0c_d - r)^2}{8t_0^2c_d^2 - (2t_0c_d - r)^2}.$$

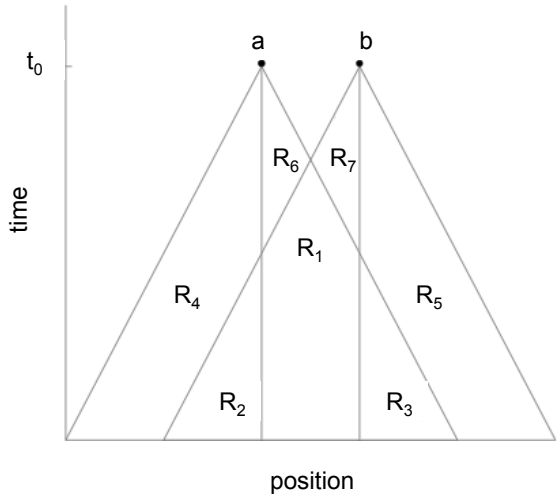


Figure A.1: **Division of $V_a(t_0) \cup V_b(t_0)$ into seven regions.** This division is used when calculating $\mathbb{P}(D_{ab}|E_2)$.

In order to calculate $\mathbb{P}(D_{ab}|E_2)$, we must split $V_a(t_0) \cup V_b(t_0)$ into 7 different regions because the probabilities will differ, depending on the location of the first mutation (as shown in Figure $A.1$). By conditioning on E_2 we assume that two mutations occur in the space-time region $V_a(t_0) \cup V_b(t_0)$. Denote the space-time coordinates of the first mutation by (x_1, t_1).

If (x_1, t_1) occurs outside of $M(r, t_0)$ but between a and b (i.e. in regions R_6 or R_7), then the cells will definitely be different, regardless of where the second mutation occurs. However, if the first mutation occurs in R_i, $1 \leq i \leq 5$, then the location of the second mutation will determine whether the sampled cells are different. Thus each region R_i, $1 \leq i \leq 5$, will have an associated region Z_i that will be used to calculate $\mathbb{P}(D_{ab}|E_2)$. If the first mutation occurs at the point $(x_1, t_1) \in R_i$, then the shape and size of $Z_i(x_1, t_1)$ depends on i and (x_1, t_1).

First, we will consider the regions inside $M(r, t_0)$, which are R_1, R_2, and R_3. For $i =$

$1, 2, 3$, $Z_i(x_1, t_1)$ represents the region in which the occurrence of a second mutation would make the sampled cells different at time t_0, i.e. the two clones will meet between a and b and then each will spread to one of the cells.

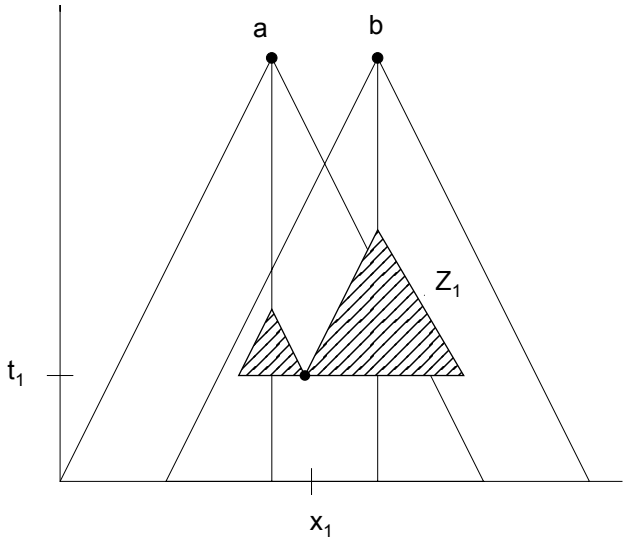


Figure A.2: **Associated region** $Z_1(x_1, t_1)$**.** The region in which the occurrence of a second mutation would make the cells located at a and b different, given that the first mutation occurred in R_1.

If $(x_1, t_1) \in R_1$, then (x_1, t_1) is in $M(r, t_0)$ and between a and b. In this case, $Z_1(x_1, t_1)$ consists of two triangles, whose upper vertices occur at positions a and b (see Figure $A.2$). The base of the triangle on the left is $2(x_1 - a)$, and the base of the triangle on the right is $2(b - x_1)$, so the total area of $Z_1(x_1, t_1)$ is $c_d^{-1}[(x_1 - a)^2 + (b - x_1)^2]$.

If $(x_1, t_1) \in R_2$, then (x_1, t_1) is in $M(r, t_0)$ but to the left of a. In this case, $Z_2(x_1, t_1)$ is a trapezoidal region. This trapezoidal region can be constructed by taking the triangle whose upper vertex is at position b and subtracting the smaller triangle with upper vertex at position a (see Figure $A.3$). The base of the larger triangle is $2(b - x_1)$, and the base of the smaller triangle is $2(a - x_1)$. Hence, the area of $Z_2(x_1, t_1)$ is $c_d^{-1}[(b - x_1)^2 - (a - x_1)^2]$. $Z_3(x_1, t_1)$ is constructed analogously to $Z_2(x_1, t_1)$.

$Z_4(x_1, t_1)$ and $Z_5(x_1, t_1)$ have a slightly different meaning. Given that the first mutation occurs in region 4 or 5, respectively, $Z_4(x_1, t_1)$ and $Z_5(x_1, t_1)$ each represent the region in which the occurrence of a second mutation would make the sampled cells genetically

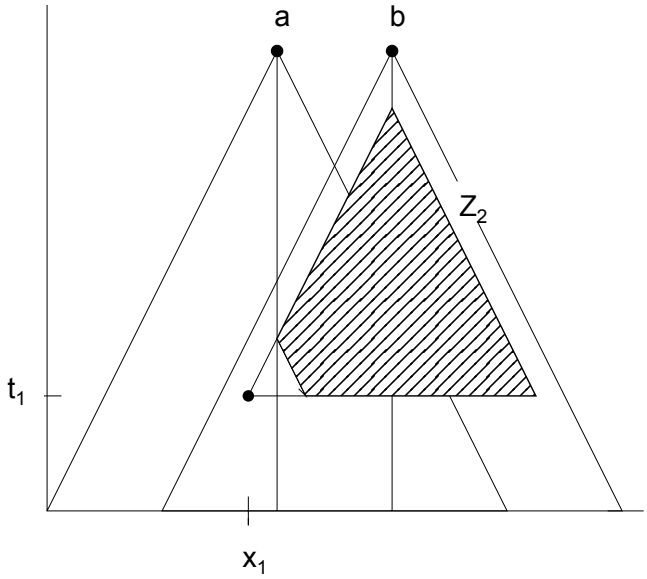


Figure A.3: **Associated region** $Z_2(x_1, t_1)$**.** The region in which the occurrence of a second mutation would make the cells located at a and b different, given that the first mutation occurred in R_2.

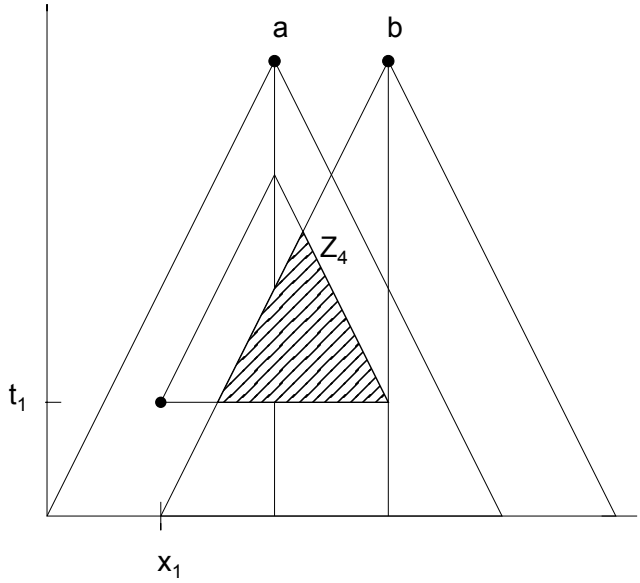


Figure A.4: **Associated region** $Z_4(x_1, t_1)$**.** The region in which the occurrence of a second mutation would make the cells located at a and b the same, given that the first mutation occurred in R_4.

identical.

If $(x_1, t_1) \in R_4$, then (x_1, t_1) is outside of $M(r, t_0)$ and to the left of a. In order for a and b to be the same in this case, the second clone must meet the first clone before it reaches a, and the second clone must spread to b before t_0. Hence, $Z_4(x_1, t_1)$ is a triangle inside $M(r, t_0)$ (see Figure $A.4$). In the next paragraph we will explain how the area of $Z_4(x_1, t_1)$ is calculated.

The distance between the right vertex of $Z_4(x_1, t_1)$ and a is equal to the distance between a and x_1, so the position of that vertex is $a + (a - x_1) = 2a - x_1$. Let V_b' be the portion of V_b that falls between the t-values t_1 and t_0. Then we can find the position of the left vertex of $Z_4(x_1, t_1)$ by considering it as the left corner of V_b'. The height of V_b' is $t_0 - t_1$, so its base is $2c_d(t_0 - t_1)$. Then the left vertex of V_b', and consequently the left vertex of $Z_4(x_1, t_1)$ is $b - c_d(t_0 - t_1)$. Hence, the base of $Z_4(x_1, t_1)$ has length $2a - x_1 - b + c_d(t_0 - t_1)$. Therefore, the area of $Z_4(x_1, t_1)$ is:

$$\frac{(2a - x_1 - b + c_d(t_0 - t_1))^2}{4c_d}.$$

Analogously, the area of $Z_5(x_1, t_1)$ is:

$$\frac{(a - 2b + x_1 + c_d(t_0 - t_1))^2}{4c_d}.$$

In summary, we have the following areas:

$$\begin{aligned}
\text{If } (x_1, t_1) \in R_1, \text{then } |Z_1(x_1, t_1)| &= c_d^{-1}[(x_1 - a)^2 + (b - x_1)^2]. \\
\text{If } (x_1, t_1) \in R_2, \text{then } |Z_2(x_1, t_1)| &= c_d^{-1}[(b - x_1)^2 - (a - x_1)^2]. \\
\text{If } (x_1, t_1) \in R_3, \text{then } |Z_3(x_1, t_1)| &= c_d^{-1}[(x_1 - a)^2 - (x_1 - b)^2]. \\
\text{If } (x_1, t_1) \in R_4, \text{then } |Z_4(x_1, t_1)| &= \frac{(2a - x_1 - b + c_d(t_0 - t_1))^2}{4c_d}. \\
\text{If } (x_1, t_1) \in R_5, \text{then } |Z_5(x_1, t_1)| &= \frac{(a + c_d(t_0 - t_1) - 2b + x_1)^2}{4c_d}.
\end{aligned}$$

Let X_n be the position of the nth mutation. Then:

$$\mathbb{P}(D_{ab}|E_2) = \sum_{i=1}^{3} \mathbb{P}(X_2 \in Z_i | X_1 \in R_i)\mathbb{P}(X_1 \in R_i) + \sum_{i=4}^{5} \mathbb{P}(X_2 \notin Z_i | X_1 \in R_i)\mathbb{P}(X_1 \in R_i) + \sum_{i=6}^{7} \mathbb{P}(X_1 \in R_i).$$

Thus to calculate $\mathbb{P}(D_{ab})$ it remains to calculate $\mathbb{P}(X_2 \in Z_i | X_1 \in R_i)$ and $\mathbb{P}(X_1 \in R_i)$ for $i \in \{1, \ldots, 5\}$.

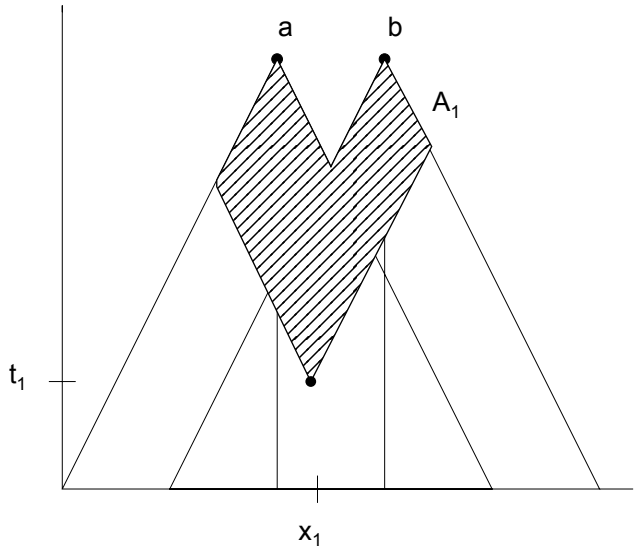


Figure A.5: **Affected region** $A_1(x_1, t_1)$**.** The region inside $V_a \cup V_b$ that is affected by a mutation at $(x, t) \in R_1$, and thus is not susceptible to subsequent mutation.

Let $A_i(x, t)$ be the region inside $V_a \cup V_b$ that is affected by a mutation at $(x, t) \in R_i$. Since type-1 mutations must occur in cells that have not yet mutated, the second type-1 mutation cannot occur inside $A_i(x, t)$.

The area of $A_i(x, t)$ depends on whether (x,t) is in $M(r, t_0), V_a \setminus V_b$, or $V_b \setminus V_a$. The following are the areas $|A_i(x, t)|$, which will be used to calculate $\mathbb{P}(X_2 \in Z_i | X_1 \in R_i)$:

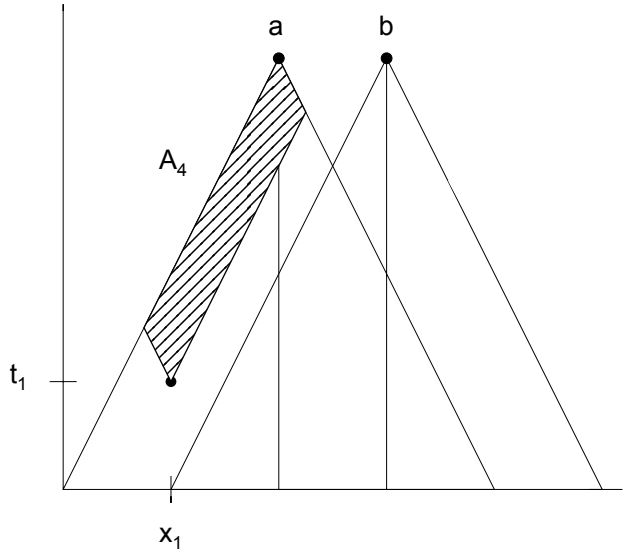


Figure A.6: **Affected region** $A_4(x_1, t_1)$**.** The region inside $V_a \cup V_b$ that is affected by a mutation at $(x, t) \in R_4$, and thus is not susceptible to subsequent mutation.

$$|A_1(x,t)| = |A_2(x,t)| = |A_3(x,t)| =$$
$$c_d(t_0 - t + \frac{r}{2c_d})^2 - \frac{2r^2 + (b - x + c_d(t_0 - t))^2 + (x - a + c_d(t_0 - t))^2}{4c_d}$$
$$|A_4(x,t)| = |A_6(x,t)| = c_d(t_0 - t)^2 - \frac{(x - a + c_d(t_0 - t))^2 + (a - x + c_d(t_0 - t))^2}{4c_d}$$
$$|A_5(x,t)| = |A_7(x,t)| = c_d(t_0 - t)^2 - \frac{(x - b + c_d(t_0 - t))^2 + (b - x + c_d(t_0 - t))^2}{4c_d}.$$

We will explain how $|A_4(x,t)|$ is calculated and leave out the calculations for $|A_1(x,t)|$ and $|A_5(x,t)|$, which can be done similarly.

$|A_4(x,t)|$ is calculated by taking the area of the truncated triangle V_a' (the portion of V_a that lies between times t and t_0) and then subtracting the area of two smaller triangles that are not in A_4 (see Figure A.6). The bases of these triangles lie along line t, between x and the two lower vertices of V_a'. The height of V_a' is $t_0 - t$, so its base is $2c_d(t_0 - t)$. Hence the lower left vertex of V_a' is at position $a - c_d(t_0 - t)$, and the lower right vertex of V_a' is at position $a + c_d(t_0 - t)$. Therefore the base of the left small triangle is $x - a + c_d(t_0 - t)$, so its area is $\frac{(x - a + c_d(t_0 - t))^2}{4c_d}$. The base of the right small triangle is $a + c_d(t_0 - t) - x$,

so its area is $\frac{(a-x+c_d(t_0-t))^2}{4c_d}$. Since $|V_a'| = c_d(t_0-t)^2$, we get the area listed above for A_4 and A_6.

If $X_1 = (x,t) \in R_i$, then $\mathbb{P}(X_2 \in Z_i) = \frac{|Z_i(x,t)|}{|V_a \cup V_b \setminus A_i(x,t)|}$. We can integrate this quantity over the places where the first mutation could have occurred, which is all of R_i, and then divide by $|R_i|$ to get:

$$\mathbb{P}(X_2 \in Z_i | X_1 \in R_i) = \frac{1}{|R_i|} \int_{R_i} \frac{|Z_i(x,t)|}{|V_a \cup V_b \setminus A_i(x,t)|} dx dt.$$

Now it remains to calculate $\mathbb{P}(X_1 \in R_i)$. Since mutations arrive according to a Poisson process, we have $\mathbb{P}(X_1 \in R_i) = \mu(|R_i|)e^{-\mu(|R_i|)}$, and it suffices to know the following areas:

$$\begin{aligned}
|R_1| &= \frac{(2t_0c_d - r)^2}{4c_d} - \frac{(a-b+c_dt_0)^2}{2c_d} \\
|R_2| &= |R_3| = \frac{(a-b+c_dt_0)^2}{4c_d} \\
|R_4| &= |R_5| = \frac{c_dt_0^2}{2} - \frac{(a-b+c_dt_0)^2}{4c_d} \\
|R_6| &= |R_7| = \frac{r^2}{4c_d}.
\end{aligned}$$

The expression for $|R_2|$ listed above is calculated by considering R_2 as a triangle inside V_b. The height of V_b is t_0, so the left vertex is at position $b - c_dt_0$. Then the base of R_2 is $a - b + c_dt_0$, which means its area is $\frac{(a-b+c_dt_0)^2}{4c_d}$.

Then we can use $|R_2|$ to calculate $|R_1|$ and $|R_6|$:

$|R_1| = |M(r,t_0)| - 2|R_2|$, and $|R_4| = \frac{1}{2}|V_a| - |R_2|$. And the height of R_6 is t_0 minus the height of R_2, so $|R_6|$ simplifies to $\frac{r^2}{4c_d}$.

All of the equations above can be used to calculate $\mathbb{P}(D_{ab})$:

$$\begin{aligned}
\mathbb{P}(D_{ab}) \approx &\mathbb{P}(D_{ab}|E_1)\mathbb{P}(E_1) + \mathbb{P}(E_2)\big(\sum_{i=1}^{3} \mathbb{P}(X_2 \in Z_i | X_1 \in R_i)\mathbb{P}(X_1 \in R_i) \\
&+ \sum_{i=4}^{5} \mathbb{P}(X_2 \notin Z_i | X_1 \in R_i)\mathbb{P}(X_1 \in R_i) + \sum_{i=6}^{7} \mathbb{P}(X_1 \in R_i)\big).
\end{aligned}$$

A.2.2 I_1 in 2 dimensions

Similary to the one dimensional case, we will first calculate $|V_a(t_0)|$ and $|M(r,t_0)|$. In the two dimensional setting this is slightly more difficult. First we know that $|V_a(t_0)| = \pi t_0^3 c_d^2/3$, so it remains to find $|M(r,t_0)|$.

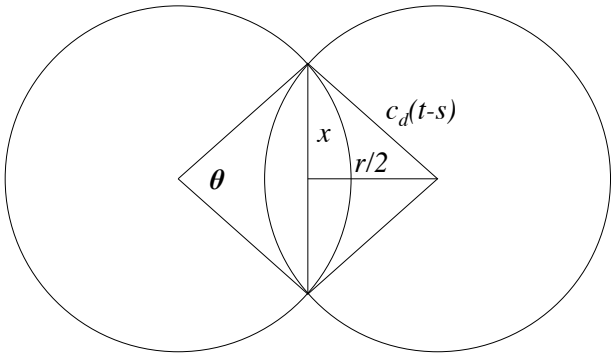


Figure A.7: **2D cross-sectional diagram.** This depicts the overlap of space-time cones, V_a and V_b, at time s.

Observe that if $r > 2c_d t_0$ then $M(r,t_0) = \emptyset$, so we only need to calculate $|M(r,t_0)|$ in the case $r < 2c_d t_0$. If we consider the overlap of space-time cones at the fixed time $s \in [0, t_0 - r/(2c_d)]$ then looking at Figure A.7 it can be seen that half the area of the overlap of their cones at this specific time is given by taking the difference between the area of the circular section with radius $c_d(t-s)$ and angle θ and twice the area of the triangle with side lengths $x, r/2, c_d(t_0 - s)$. The area of the circular section is given by

$$c_d^2(t_0-s)^2 \cos^{-1}\left(\frac{r}{2c_d(t_0-s)}\right),$$

and twice the area of the triangle is given by

$$\frac{r}{2}\sqrt{c_d^2(t_0-s)^2 - r^2/4}.$$

Thus the area of overlap between the two cones at time s is given by

$$a(s) = 2\left(c_d^2(t_0-s)^2 \cos^{-1}\left(\frac{r}{2c_d(t_0-s)}\right) - \frac{r}{2}\sqrt{c_d^2(t_0-s)^2 - r^2/4}\right).$$

The space-time volume of $M(r,t)$ is therefore given by

$$\begin{aligned}|M(r,t_0)| =& \int_0^{t_0-r/2c_d} a(s)ds \\ =& 2\int_0^{t_0-r/2c_d} \left(c_d^2(t_0-s)^2\cos^{-1}\left(\frac{r}{2c_d(t_0-s)}\right) - \frac{r}{2}\sqrt{c_d^2(t_0-s)^2 - r^2/4}\right) ds \\ =& \frac{2}{c_d}\int_{r/2}^{c_dt_0} y^2\cos^{-1}\left(\frac{r}{2y}\right)dy - \frac{r}{c_d}\int_{r/2}^{c_dt_0}\sqrt{y^2-r^2/4}\,dy.\end{aligned}$$

Applying integration by parts to the first integral we see that

$$\frac{2}{c_d}\int_{r/2}^{c_dt} y^2\cos^{-1}\left(\frac{r}{2y}\right)dy = \frac{2c_d^2t_0^3}{3}\cos^{-1}\left(\frac{r}{2c_dt_0}\right) - \frac{r}{3c_d}\int_{r/2}^{c_dt_0}\frac{y^2}{\sqrt{y^2-r^2/4}}dy.$$

Thus we have that

$$\begin{aligned}|M(r,t_0)| =& \frac{2c_d^2t_0^3}{3}\cos^{-1}\left(\frac{r}{2c_dt_0}\right) - \frac{r}{6c_d}\int_{r/2}^{c_dt_0}\frac{16y^2-3r^2}{\sqrt{4y^2-r^2}}dy \\ =& \frac{2c_d^2t_0^3}{3}\cos^{-1}\left(\frac{r}{2c_dt_0}\right) - \frac{rt_0}{3}\sqrt{4c_d^2t_0^2-r^2} - \frac{r^3}{12c_d}\log\left(\frac{r}{2c_dt_0+\sqrt{4c_d^2t_0^2-r^2}}\right),\end{aligned}$$

which we can combine with (A.12) to see that for $r < 2c_dt_0$

$$\begin{aligned}|D(r,t)| =& \frac{2c_d^2t_0^3}{3}\left(\pi - 2\cos^{-1}\left(\frac{r}{2c_dt_0}\right)\right) + \frac{2rt_0}{3}\sqrt{4c_d^2t_0^2-r^2} \\ &+ \frac{r^3}{6c_d}\log\left(\frac{r}{2c_dt_0+\sqrt{4c_d^2t_0^2-r^2}}\right).\end{aligned}$$

With these calculations we see that

$$\begin{aligned}|V_a(t_0)\cup V_b(t_0)| =& \frac{2c_d^2t_0^3}{3}\left(\pi - \cos^{-1}\left(\frac{r}{2c_dt_0}\right)\right) + \frac{rt_0}{3}\sqrt{4c_d^2t_0^2-r^2} \\ &+ \frac{r^3}{12c_d}\log\left(\frac{r}{2c_dt_0+\sqrt{4c_d^2t_0^2-r^2}}\right).\end{aligned}$$

We can now explicitly calculate $\mathbb{P}(D_{ab}|E_1) = \frac{|D(r,t_0)|}{|V_a(t_0)\cup V_b(t_0)|}$. The remainder of this section will deal with the calculation of $\mathbb{P}(D_{ab}|E_2)$.

The approach here will be slightly different from the one-dimensional case because it

is easier to look at the two-dimensional cross sections of $|V_a(t_0) \cup V_b(t_0)|$, rather than the entire three-dimensional space-time cones. Therefore, we will split the cross sections into just two regions, and then when calculating the relevant volumes involved in I_2, we will split the regions into multiple cases. In the end, the process is similar, but the setup will be simpler, and then the volume calculations will be more complicated in the two-dimensional setting.

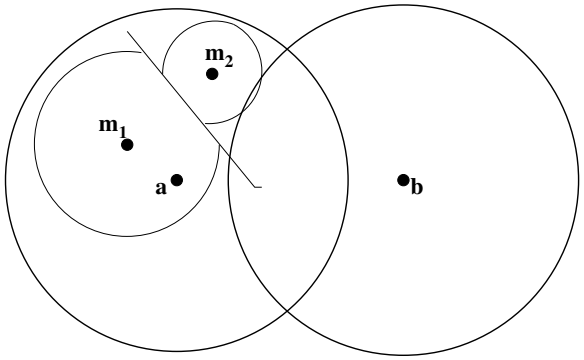


Figure A.8: **2D clone interaction.** When two mutation circles collide, they will continue to expand along the line perpendicular to the line segment joining the two mutation origins.

If two events occur in $V_a(t_0) \cup V_b(t_0)$, then the probabilities will differ, depending on whether the first event occurs in $M(r, t_0)$ or in $D(r, t_0)$. We will assume that if two mutation circles collide, then we can draw a line through that point, perpendicular to the line segment connecting the two mutations (as show in Figure $A.8$). The circles will not extend beyond that line but will continue to expand in all other directions.

If the first event occurs in $M(r, t_0)$ at position (x_1, y_1) at time t_1, then let r_a be the distance between (x_1, y_1) and a, and let r_b be the distance between (x_1, y_1) and b. Then let $C_a(t_1)$ be the cone centered at a that extends to the edge of the expanding clone, so $C_a(t_1)$ will have radius r_a at time t_1 and radius 0 at time $(t_1 + \frac{r_a}{c_d})$. Similarly, $C_b(t_1)$ will be the cone centered at b with radius r_b at time t_1 and radius 0 at time $(t_1 + \frac{r_b}{c_d})$. Cross-sections of these cones are shown in Figure A.9.

If the second mutation occurs outside of $C_a(t_1) \cup C_b(t_1)$, then the first clone will reach both a and b before interacting with the second clone. If the second mutation occurs in $C_a(t_1) \setminus C_b(t_1)$, then the line dividing the two clones will separate a from b, so the second

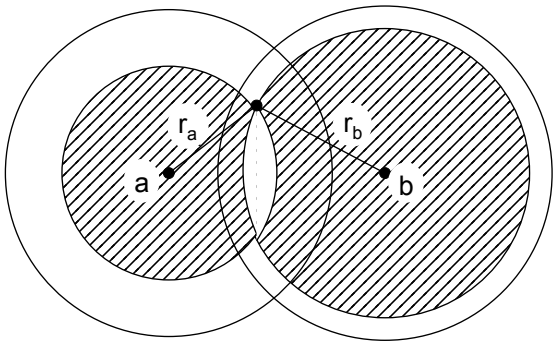


Figure A.9: **Associated region, given an initial mutation in $M(r, t_0)$.** Displayed is the cross-section of the cones $V_a(t_0)$, $V_b(t_0)$, $C_a(t_1)$, and $C_b(t_1)$ at the moment when a mutation occurs in the intersection, $M(r, t_0)$. If a second mutation occurs in the shaded mutation, then the cells located at a and b will be different.

clone will affect a, and the first will affect b, making the two cells different. Similarly, if the second mutation occurs in $C_b(t_1) \setminus C_a(t_1)$, then the first clone will affect a, and the second will affect b.

However if the second mutation occurs in $C_b(t_1) \cap C_a(t_1)$, then both a and b will be on the same side of the line dividing the mutation circles, so the second clone will affect both a and b. Therefore, the two cells will only be different if the second mutation occurs in $C_b(t_1) \triangle C_a(t_1)$.

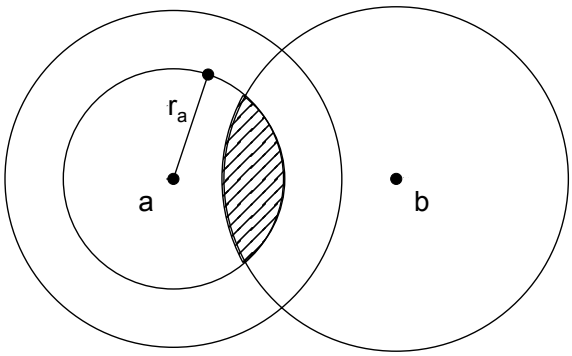


Figure A.10: **Associated region, given an initial mutation in $D(r, t_0)$.** Displayed is the cross-section of the cones $V_a(t_0)$, $V_b(t_0)$, and $C_a(t_1)$ at the moment when a mutation occurs in $D(r, t_0)$. If a second mutation occurs in the shaded mutation, then the cells located at a and b will be the same.

If the first mutation occurs in $D(r, t_0)$, then its position (x_1, y_1) is closer to either a or b. Without loss of generality, say that (x_1, y_1) is closer to a. Again let r_a be the distance between (x_1, y_1) and a, and let $C_a(t_1)$ be the cone centered at a with radius r_a at time t_1 and radius 0 at time $(t_1 + \frac{r_a}{c_d})$. A cross-section of this cone is shown in Figure A.10.

If the second mutation occurs outside of $C_a(t_1)$, then the first mutation will reach a before interacting with the second mutation. Since the first mutation is outside $V_b(t_0)$, it cannot reach b by time t_0, so the sampled cells will be different.

If the second mutation occurs inside $C_a(t)$ but outside $M(r, t)$, then the two mutations will interact before the first mutation reaches a, meaning that the second mutation will affect a. However, the second mutation will not spread to b, since it does not start in $V_b(t_0)$. Hence, the cells located at a and b will only be the same if the second mutation occurs in $C_a(t_1) \cap M(r, t_0)$.

In summary, we have:

$$\begin{aligned}\mathbb{P}(D_{ab}|E_2) =& \mathbb{P}(X_2 \in C_b(t_1) \triangle C_a(t_1) | X_1 \in M(r,t_0)) \mathbb{P}(X_1 \in M(r,t_0)) \\ &+ \mathbb{P}(X_2 \notin C_a(t_1) \cap M(r,t_0) | X_1 \in V_a(t_1) \setminus V_b(t_1)) \mathbb{P}(X_1 \in V_a(t_1) \setminus V_b(t_1)) \\ &+ \mathbb{P}(X_2 \notin C_b(t_1) \cap M(r,t_0) | X_1 \in V_b(t_1) \setminus V_a(t_1)) \mathbb{P}(X_1 \in V_b(t_1) \setminus V_a(t_1)).\end{aligned} \tag{A.13}$$

Since the mutations arise according to a Poisson process, we can use the volume calculations for $M(r, t_0)$ and $V_a(t_1) \setminus V_b(t_1)$ to calculate the following probabilities:

$$\mathbb{P}(X_1 \in M(r,t_0)) = \mu(|M(r,t_0)|) e^{-\mu(|M(r,t_0)|)}$$
$$\mathbb{P}(X_1 \in V_a(t_1) \setminus V_b(t_1)) = \mathbb{P}(X_1 \in V_b(t_1) \setminus V_a(t_1)) = \mu(|V_a(t_1) \setminus V_b(t_1)|) e^{-\mu(|V_a(t_1) \setminus V_b(t_1)|)}$$

Similarly to the 1-D case:

$$\begin{aligned}&\mathbb{P}(X_2 \in C_b(t) \triangle C_a(t) | X_1 \in M(r,t_0)) \\ &= \frac{1}{|M(r,t_0)|} \int_{M(r,t_0)} \frac{|C_b(t) \triangle C_a(t)|}{|V_a(t_0) \cup V_b(t_0) \setminus A(x,y,t)|} dx dy dt,\end{aligned} \tag{A.14}$$

where $A(x, y, t)$ is the cone-shaped region inside $V_a(t_0) \cup V_b(t_0)$ that is affected by a mutation at (x, y, t).

In addition:

$$\begin{aligned}&\mathbb{P}(X_2 \in C_a(t_1) \cap M(r,t_0) | X_1 \in V_a(t_0) \setminus V_b(t_0)) \\ &= \frac{1}{|V_a(t_0) \setminus V_b(t_0)|} \int_{V_a(t_0)\setminus V_b(t_0)} \frac{|C_a(t) \cap M(r,t_0)|}{|V_a(t_0) \cup V_b(t_0) \setminus A(x,y,t)|} dxdydt.\end{aligned} \tag{A.15}$$

We next develop formulas to compute the volumes in the previous two displays. Note that

$$|C_b(t_1) \triangle C_a(t_1)| = |C_b(t_1)| + |C_a(t_1)| - 2|C_b(t_1) \cap C_a(t_1)|,$$

and that $|C_a(t_1)| = \frac{\pi}{3} r_a^2 \left(t_1 + \frac{r_a}{c_d} \right)$, since $C_a(t_1)$ is a cone with radius r_a and height $t_1 + \frac{r_a}{c_d}$. Similarly, $|C_b(t_1)| = \frac{\pi}{3} r_b^2 \left(t_1 + \frac{r_b}{c_d} \right)$.

We next compute $|C_a(t_1) \cap C_a(t_1)|$. A cross-section of $C_a(t_1)$ has radius $r_a - (s - t_1)c_d$ at time s, and a cross-section of $C_b(t_1)$ has radius $r_b - (s - t_1)c_d$ at time s. $C_a(t_1)$ and $C_b(t_1)$ will have a nonempty intersection until $r_a - (s-t_1)c_d + r_b - (s-t_1)c_d = r$, i.e. when $s = \frac{r_a + r_b - r}{2c_d} + t_1$.

If we denote the area of intersection of the cross-sections of $C_a(t_1)$ and $C_b(t_1)$ at time s by $I(s)$, then

$$|C_b(t_1) \cap C_a(t_1)| = \int_{t_1}^{\frac{r_a + r_b - r}{2c_d} + t_1} I(s) ds. \tag{A.16}$$

$I(s)$ is calculated by summing the areas of the two circular segments, each of which can be calculated by subtracting the area of a triangle from the area of a wedge of the circle,

$$\begin{aligned}I(s) = R_a^2(s) \cos^{-1} \left(\frac{d_a(s)}{R_a(s)} \right) - d_a(s) \sqrt{R_a^2(s) - d_a^2(s)} \\ + \mathbb{R}_b^2(s) \cos^{-1} \left(\frac{d_b(s)}{R_b(s)} \right) - d_b(s) \sqrt{R_b^2(s) - d_b^2(s)},\end{aligned}$$

where $R_a(s) = r_a - (s - t_1)c_d$, $R_b(s) = r_b - (s - t_1)c_d$, $d_a(s) = \frac{r^2 - R_b^2(s) + R_a^2(s)}{2r}$, and $d_b = \frac{r^2 + R_b^2 - R_a^2}{2r}$.

In order to compute the quantity $|C_a(t) \cap M(r,t)|$ used in equation (A.15), we need to

first determine when the cross sections of $C_a(t)$ and $V_b(t)$ have nonempty intersection. This occurs when $r_a - (s - t_1)c_d + c_d(t_0 - s) > r$, i.e. when $s < \frac{1}{2}\left(\frac{r_a - r}{c_d} + t_0 + t_1\right)$. Hence:

$$|C_a(t) \cap M(r,t)| = \int_{t_1}^{\frac{1}{2}(\frac{r_a-r}{c_d}+t_0+t_1)} \hat{I}(s)ds, \tag{A.17}$$

where

$$\hat{I}(s) = R_a^2(s)\cos^{-1}\left(\frac{\hat{d}_a(s)}{R_a(s)}\right) - \hat{d}_a(s)\sqrt{R_a^2(s) - \hat{d}_a{}^2(s)}$$
$$+ \hat{R}_b{}^2(s)\cos^{-1}\left(\frac{\hat{d}_b(s)}{\hat{R}_b(s)}\right) - \hat{d}_b(s)\sqrt{\hat{R}_b{}^2(s) - \hat{d}_b{}^2(s)}.$$

R_a is defined above, $\hat{R}_b(s) = c_d(t_0 - s)$, $\hat{d}_a(s) = \dfrac{r^2 - \hat{R}_b(s)^2 + R_a^2(s)}{2r}$, and $\hat{d}_b(s) = \dfrac{r^2 + \hat{R}_b{}^2(s) - R_a^2(s)}{2r}$.

We finally compute $|(V_a(t_0) \cup V_b(t_0)) \setminus A(x,y,t)|$. In pursuit of this, we define U_1 as the region that is affected by the mutation at (x_1, y_1, t_1), i.e.,

$$U_1 = \{(x,y,s) : |(x,y) - (x_1,y_1)| \le c_d(s - t_1), t_1 \le s \le t_0\}.$$

Let $u_1(s)$ be the cross-section of U_1 at time s, i.e.,

$$u_1(s) = \{(x,y) : |(x,y) - (x_1,y_1)| \le c_d(s - t_1)\}.$$

Observe that $A(x_1, y_1, t_1)$ is the region inside $V_a(t_0) \cup V_b(t_0)$ that is affected by a mutation at (x_1, y_1, t_1), so $A(x_1, y_1, t_1) = U_1 \cap (V_a(t_0) \cup V_b(t_0))$. This implies that

$$|(V_a(t_0) \cup V_b(t_0)) \setminus A(x_1,y_1,t_1)| = |V_a(t_0) \cup V_b(t_0)| - |A(x_1,y_1,t_1)|.$$

Then it remains to find $|A(x_1, y_1, t_1)|$. This will be accomplished by looking at the cross-sections of this set for each fixed time s. Define $v_a(s)$ and $v_b(s)$ as the cross sections

of V_a and V_b, respectively, at time s, i.e.,

$$v_a(s) = \{(x, y) : |(x, y) - a| \leq c_d(t_0 - s)\}$$
$$v_b(s) = \{(x, y) : |(x, y) - b| \leq c_d(t_0 - s)\}.$$

If $(x_1, y_1, t_1) \in V_a(t_0) \backslash V_b(t_0)$, then U_1 will not intersect $V_b(t_0)$, so in this case $A(x_1, y_1, t_1)$ $= U_1 \cap V_a(t_0)$. In order to compute the volume of this set we look at the area of the cross-section for each fixed time point. We can divide the interval $[0, t_0]$ intp three distinct intervals, which determine the shape of the area of this cross-section. In the first time interval, the cross-section of U_1 is contained in the cross-section of V_a, so we determine the time interval as shown below,

$$\begin{aligned} & u_1(s) \cap v_a(s) = u_1(s) \\ \iff & u_1(s) \subset v_a(s) \\ \iff & r_a + c_d(s - t_1) < c_d(t_0 - s) \\ \iff & s < \frac{t_1 + t_0}{2} - \frac{r_a}{2c_d}. \end{aligned}$$

In the final interval, the cross-section of V_a is contained in the cross-section of U_1, so we have

$$\begin{aligned} & u_1(s) \cap v_a(s) = v_a(s) \\ \iff & v_a(s) \subset u_1(s) \\ \iff & r_a + c_d(t_0 - s) < c_d(s - t_1) \\ \iff & s > \frac{t_1 + t_0}{2} + \frac{r_a}{2c_d}. \end{aligned}$$

When $\frac{t_1 + t_0}{2} - \frac{r_a}{2c_d} < s < \frac{t_1 + t_0}{2} + \frac{r_a}{2c_d}$, we have

$$\begin{aligned} |u_1(s) \cap v_a(s)| = {} & R_u 2(s) \cos^{-1}\left(\frac{d_u(s)}{R_u(s)}\right) - d_u(s)\sqrt{R_u^2(s) - d_u^2(s)} + \hat{R}_a^2(s) \cos^{-1}\left(\frac{\hat{d}_a(s)}{\hat{R}_a(s)}\right) \\ & - \hat{d}_a(s)\sqrt{\hat{R}_a^2(s) - \hat{d}_a^2(s)}, \end{aligned}$$

where $R_u(s) = c_d(s - t_1)$, $\hat{R}_a(s) = c_d(t_0 - s)$, $d_u(s) = \dfrac{r_a^2 - \hat{R}_a^2(s) + R_u^2(s)}{2r_a}$, and $\hat{d}_a(s) = \dfrac{r_a^2 + \hat{R}_a^2(s) - R_u^2(s)}{2r_a}$.

Thus for $(x_1, y_1, t_1) \in V_a(t_0) \setminus V_b(t_0)$,

$$\begin{aligned} &|A(x_1, y_1, t_1)| \\ &= \int_{t_1}^{\frac{t_1+t_0}{2} - \frac{r_a}{2c_d}} |u_1(s)|ds + \int_{\frac{t_1+t_0}{2} - \frac{r_a}{2c_d}}^{\frac{t_1+t_0}{2} + \frac{r_a}{2c_d}} |u_1(s) \cap v_a(s)|ds + \int_{\frac{t_1+t_0}{2} + \frac{r_a}{2c_d}}^{t_0} |v_a(s)|ds. \end{aligned} \tag{A.18}$$

$|A(x_1, y_1, t_1)|$ is computed analogously when $(x_1, y_1, t_1) \in V_b(t_0) \setminus V_a(t_0)$.

It remains to compute $|A(x_1, y_1, t_1)|$ when $(x_1, y_1, t_1) \in V_b(t_0) \cap V_a(t_0)$. First note that if $(x_1, y_1, t_1) \in V_b(t_0) \cap V_a(t_0)$, then

$$\begin{aligned} & u_1(s) \cap (v_a(s) \cup v_b(s)) = u_1(s) \\ \iff & u_1(s) \subset (v_a(s) \cup v_b(s)) \\ \iff & s < \frac{t_1 + t_0}{2} - \frac{\min\{r_a, r_b\}}{2c_d} \doteq s_1. \end{aligned}$$

Once the cross-sections $v_a(s)$ and $v_b(s)$ are no longer intersecting i.e., when $s > t_0 - \frac{r}{2c_d}$, then

$$|u_1(s) \cap (v_a(s) \cup v_b(s))| = |u_1(s) \cap v_a(s)| + |u_1(s) \cap v_b(s)|.$$

The two quantities $|u_1(s) \cap v_a(s)|$ and $|u_1(s) \cap v_b(s)|$ can be calculated as shown above.

Then for $\dfrac{t_1 + t_0}{2} - \dfrac{\min\{r_a, r_b\}}{2c_d} < s < t_0 - \dfrac{r}{2c_d}$, we have

$$|u_1(s) \cap (v_a(s) \cup v_b(s))| = |u_1(s) \cap v_a(s)| + |u_1(s) \cap v_b(s)| - |u_1(s) \cap v_a(s) \cap v_b(s)|.$$

The quantities $|u_1(s) \cap v_a(s)|$ and $|u_1(s) \cap v_b(s)|$ can be calculated as shown above, and $|u_1(s) \cap v_a(s) \cap v_b(s)|$ can be calculated as shown in [113].

Thus if $(x_1, y_1, t_1) \in V_a(t_0) \cap V_b(t_0)$, then

$$|A(x_1, y_1, t_1)| \tag{A.19}$$
$$= \int_{t_1}^{s_1} |u_1(s)|ds + \int_{s_1}^{t_0 - r/(2c_d)} |u_1(s) \cap (v_a(s) \cup v_b(s))|ds.$$

With (A.18) and (A.19) we can compute $|A(x_1, y_1, t_1)|$ for arbitrary (x_1, y_1, t_1). We can then use $|A(x_1, y_1, t_1)|$ with (A.17) and (A.16) to compute (A.14) and (A.15). Finally we use (A.14) and (A.15) to compute $P(D_{ab}|E_2)$ based on (A.13).

A.3 I_2 Calculations

In this section we describe how to compute $I_2(r, t)$ and $I_2(r, \tau)$. First recall from Section 2.3.2 that R is the radius of the clone, Y is chosen according to a size-biased pick, and X is the distance of p (a point selected at random from Y) from the center of Y.

We first describe how to estimate $I_2(r, t)$ based on (2.11). In particular, we can generate K i.i.d copies of the vector (X, R), denoted by $\{(X_i, R_i)\}_{i=1}^K$. Our method for generating (X_1, R_1) based on the time interval $[0, t]$ is as follows. First generate the arrival times of mutations based on a Poisson process with rate $Nu_1 s$, denote this set of times by $t_1, \ldots, t_n$.. Then for each mutation calculate the size of its family at time t using the formula (2.7), and this gives us the collection of family sizes $Y_{1,1}, \ldots, Y_{1,n}$ of clones $C_{1,1}, \ldots, C_{1,n}$. Choose a clone $C = C_{[1]}$ via a size biased pick from the collection $C_{1,1}, \ldots, C_{1,n}$, and set R to be the radius of C. Let U be a uniform random variable on $[0, 1]$ independent of R and set $X = R\sqrt{U}$. With these samples, form the estimator

$$\hat{I}_2(r, t) = \frac{1}{K} \sum_{i=1}^{K} P(p_2 \in C | R_i, X_i).$$

We can also derive an alternative representation for $P(p_2 \in C)$ that is more suitable for mathematical analysis. Denote the conditional density of X, given $R = y$, by $f_X(x|R = y)$ and the density of R by f_R. It's easy to see that $f_X(x|R = y) = \frac{2x}{y^2}$ for $x \in (0, y)$ and 0

otherwise, and therefore

$$\begin{aligned} P(p_2 \in C) &= \int_r^\infty \int_0^{y-r} \frac{2x}{y^2} f_R(y) dx dy + \int_{r/2}^r \int_{r-y}^y \phi(x,y) \frac{2x}{y^2} f_R(y) dx dy \\ &\quad + \int_r^\infty \int_{y-r}^y \phi(x,y) \frac{2x}{y^2} f_R(y) dx dy \\ &= \int_r^\infty \frac{(y-r)^2}{y^2} f_R(y) dy + \int_{r/2}^r \int_{r-y}^y \phi(x,y) \frac{2x}{y^2} f_R(y) dx dy \\ &\quad + \int_r^\infty \int_{y-r}^y \phi(x,y) \frac{2x}{y^2} f_R(y) dx dy. \end{aligned} \tag{A.20}$$

Define

$$\psi_r(R) = \frac{2}{R^2} \int_{|R-r|}^R x \phi(x,R) dx$$

and

$$\Phi_r(R) = \begin{cases} \frac{(R-r)^2}{R^2} + \psi_r(R), & R \geq r \\ \psi_r(R), & R \in (r/2, r) \\ 0, & R \leq r/2. \end{cases}$$

Therefore we see from (A.20) that we have $\mathbb{P}(p_2 \in C) = \mathbb{E}[\Phi_r(R)]$.

The formula $\mathbb{P}(p_2 \in C) = \mathbb{E}[\Phi_r(R)]$ is difficult to work with, due to the complex distribution of R. However, an interesting observation is that the distribution of R becomes much simpler if we assume that the sampling occurs at the random detection time τ. In this case define $R(\tau)$ to be the radius of the clone that we choose at time τ. Then we can use equation (9) in [6] to see that conditional on $\tau = t$, $R(\tau)$ has density

$$f(x|t) = \frac{\mu \gamma_d x^d}{c_d(1 - e^{-\theta t^{d+1}})} \exp\left[-\frac{\mu \gamma_d r^{d+1}}{c_d(d+1)} \right]$$

for $x \leq c_d t$ and zero otherwise. In the conditional density above $\theta = \mu \gamma c_d^d / (d+1)$. In order to describe the distribution of $R(\tau)$ we then need the distribution of τ, which we can get from (4) of [6]. In particular define

$$\phi(t) = \frac{1}{t} \int_0^t \exp\left(-\theta r^{d+1} \right) dr,$$

and $\lambda = N u_1 s$. Then τ has density

$$f_\tau(t) = \lambda e^{t\lambda(\phi(t)-1)} \left(1 - e^{-\theta t^{d+1}} \right).$$

Therefore we can calculate that

$$\begin{aligned} P\left(R(\tau) > z\right) &= \frac{\mu\gamma_d\lambda}{c_d}\int_{z/c_d}^{\infty}\int_{z}^{c_d t} r^d \exp\left[-\frac{\mu\gamma_d r^{d+1}}{c_d(d+1)}\right] dr e^{t\lambda(\phi(t)-1)}dt \\ &= \lambda\int_{z/c_d}^{\infty}\left(\exp\left(-\theta(z/c_d)^{d+1}\right) - \exp\left(-\theta t^{d+1}\right)\right)e^{t\lambda(\phi(t)-1)}dt \\ &= \exp\left[\lambda(z/c_d)(\phi(z/c_d)-1)\right] - \lambda\left[1-\exp\left(-\theta(z/c_d)^{d+1}\right)\right]\int_{z/c_d}^{\infty} e^{\lambda t(\phi(t)-1)}dt. \end{aligned}$$

Furthermore we can take derivatives to find that $R(\tau)$ has density given by

$$\hat{f}_R(z) = \frac{\lambda\theta(d+1)}{c_d^{d+1}} z^d \exp\left[-\theta\left(\frac{z}{c_d}\right)^{d+1}\right]\int_{z/c_d}^{\infty} e^{t\lambda(\phi(t)-1)}dt.$$

Note that the density $\hat{f}$ is very similar to the Weibull density, and thus we can generate samples from $\hat{f}$ by using the acceptance rejection algorithm with a proposal distribution based on the Weibull distribution. With these samples from the density $\hat{f}$, we can use the function Φ_r to estimate $I_2(r,\tau)$.

Note that when approximating $I_2(r,\tau)$ it is not necessary to simulate the mesoscopic model. We simply generate random variables according to the density $\hat{f}_R$ and then evaluate the function $\Phi_r(R)$. However, approximating $I_2(r,t)$ is a greater computational burden because it requires simulating the mesoscopic model.

A.4 Characteristic length scale based on I_2

In practice, I_2 can help to reduce the number of subsequent biopsy samples that should be taken after a premalignant sample has been found. For example, if a clinician wanted to take samples so that the probability that they come from the original premalignant clone is at most p, then we can define a length $\hat{r}_p \equiv \{\text{argmin}_{r>0} I_2(r,t) < p\}$; samples taken a distance of at least $\hat{r}_p$ away from the original sample should satisfy this requirement. Figure A.11 shows plots of $\hat{r}_{.5}$ as a function of s and as a function of t. We can see in these plots that as the selection strength increases, $\hat{r}_{.5}$ increases. As s increases, mutant clones expand more quickly, so samples must be taken farther away from the premalignant sample in order to guarantee $I_2 < 0.5$. Similarly as the sampling time t increases, we expect the clones to be larger by the time the cells are sampled, so $\hat{r}_{.5}$ increases as well.

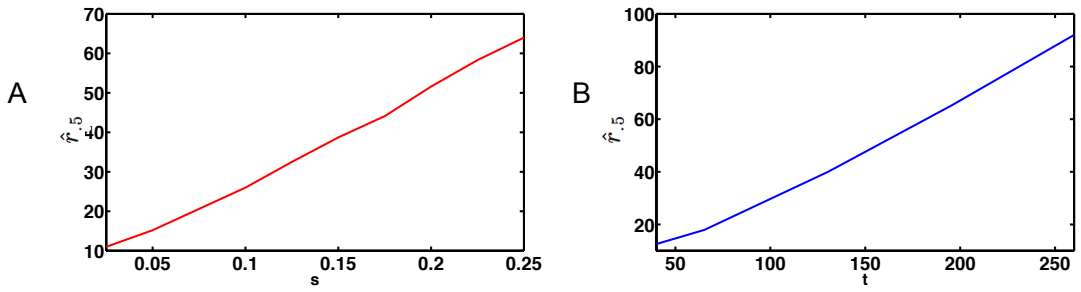


Figure A.11: **2D length-scale $\hat{r}_{.5}$ in 2D for varying parameters.** $\hat{r}_{.5}$ is displayed as a function of (A) selection strength, s, and (B) time of sampling, t. In all panels $N = 2e5$, and $1e4$ Monte Carlo simulations are performed. Unless varied, $s = 0.1$, $u_1 = 7.5e-7$, and t is the median of the detection time τ with $\mu = 2e-6$.

Appendix B

Chapter 3 Appendix

B.1 Local central limit theorem on $\mathbb{Z}^d \times \mathbb{Z}_w$

Let $(Y_t)_{t\geq 0}$ be a simple symmetric random walk on $\mathbb{Z}^d \times \mathbb{Z}_w$, starting at the origin with jump rate α. Let $N(t)$ be the number of jumps Y_t has taken by time t. Let $e_1 = (1, 0, \ldots, 0), \ldots, e_{d+1} = (0, \ldots, 0, 1)$ be the standard basis of unit vectors in $\mathbb{Z}^{d+1}$. Then

$$Y_t = \sum_{j=1}^{N(t)} X(j),$$

where the independent $X(j) = (X_1(j), X_2(j), \ldots, X_{d+1}(j))$ can be e_i or $-e_i$ for $1 \leq i \leq d+1$, each with probability $1/(2d+2)$. The sum is taken mod w in the $d+1$ dimension.

Our approach will be to first establish the LCLT for the embedded discrete time random walk $\{S_n : n \geq 0\}$ with $S_0 = 0$ and for $n \geq 1$

$$S_n = \sum_{j=1}^{n} X(j).$$

Thus for $\mathbf{x} \in \mathbb{Z}^d \times \mathbb{Z}_w$ we are interested in $\lim_{n\to\infty} n^{d/2}\mathbb{P}(S_n = \mathbf{x})$. Of course if w is even the random walk $\{S_n : n \geq 0\}$ has period 2 and the aforementioned limit does not exist. This problem is easily dealt with, but requires some extra notation. In particular we say $\mathbf{x}$ and n have the same parity if $\mathbb{P}(S_n = \mathbf{x}) > 0$.

In order to evaluate the discrete time LCLT, we condition on the number of steps that the discrete time random walk $\{S_n : n \geq 0\}$ takes in each dimension. Let $\hat{N}(n)$ be the

number of steps $\{S_n : n \geq 0\}$ has taken in the first d dimensions in the first n steps, i.e.,

$$\hat{N}(n) = |\{1 \leq j \leq n : X_{d+1}(j) = 0\}|.$$

We define the following notation to denote the probability of taking a step on $\mathbb{Z}^d$

$$\mu_{w,d} = \begin{cases} \frac{2d}{2d+1} & w = 2 \\ \frac{d}{d+1} & w > 2. \end{cases}$$

Then we have $\hat{N}(n) \sim \text{ Bin}(n, \mu_{w,d})$, which implies that for $\epsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}\left(\left|\frac{\hat{N}(n)}{n} - \mu_{w,d}\right| > \epsilon\right) \to 0.$$

Next we define small neighborhoods of the mean of $\hat{N}(n)$, in order to break up $\mathbb{P}(S_n = \mathbf{x})$ into conditional probabilities. For $\nu \in (0, 1/2)$ define

$$A_n(\nu) = \left\{1 \leq j \leq n : |j - n\mu_{w,d}| < n^{1/2+\nu}\right\}.$$

Define $\{\hat{S}_n : n \geq 0\}$ as the simple symmetric random walk on $\mathbb{Z}^d$ and $\{\hat{S}_{n,w} : n \geq 0\}$ as the simple symmetric random walk on $\mathbb{Z}_w$. Therefore, we have the representation $S_n = \left(\hat{S}_{\hat{N}(n)}, \hat{S}_{(n-\hat{N}(n)),w}\right)$ and by the conditional independence of $\hat{S}_{\hat{N}(n)}$ and $\hat{S}_{n-\hat{N}(n),w}$,

$$\mathbb{P}(S_n = \mathbf{x}|\hat{N}(n) = k) = \mathbb{P}(\hat{S}_k = \hat{\mathbf{x}})\mathbb{P}(\hat{S}_{n-k,w} = x_{d+1}),$$

where $\mathbf{x} = (x_1, x_2, \ldots, x_d, x_{d+1}) \in \mathbb{Z}^d \times \mathbb{Z}_w$, and $\hat{\mathbf{x}} = (x_1, x_2, \ldots, x_d) \in \mathbb{Z}^d$.

We will provide explicit bounds for $\mathbb{P}(\hat{S}_{n-\hat{N}(n),w} = x_{d+1})$ below.

Proposition B.1.1. *For positive integer n and $\nu \in (0, 1/2)$ such that $A_n(\nu) \neq \emptyset$ and $\mu_{w,d} + n^{-1/2+\nu} < 1$ choose $k \in A_n(\nu)$ and define $\varepsilon_n = n^{-1/2+\nu}$. If w is odd, then*

$$\frac{1}{w} - \sqrt{w}|\lambda_2|^{n\left(1-\mu_{w,d}-\varepsilon_n\right)} \leq \mathbb{P}(\hat{S}_{n-k,w} = x_{d+1}) \leq \frac{1}{w} + \sqrt{w}|\lambda_2|^{n\left(1-\mu_{w,d}-\varepsilon_n\right)}.$$

If w is even and x_{d+1} and $n-k$ have the same parity then

$$\frac{2}{w} - \sqrt{\frac{w}{2}}|\hat{\lambda}_2|^{n\left(1-\mu_{w,d}-\varepsilon_n\right)} \leq \mathbb{P}(\hat{S}_{n-k,w} = x_{d+1}) \leq \frac{2}{w} + \sqrt{\frac{w}{2}}|\hat{\lambda}_2|^{n\left(1-\mu_{w,d}-\varepsilon_n\right)}.$$

Proof. Let P be the transition matrix for $\hat{S}_{n-\hat{N}(n),w}$. Notice that its stationary distribution is

$$\pi^* = \left(\frac{1}{w}, \frac{1}{w}, \ldots, \frac{1}{w}\right), \tag{B.1}$$

which follows from the fact that $P(x, x+1) = P(x, x-1) = 1/2$ for all x when $w > 2$, and $P(x, x+1) = 1$ when $w = 2$; where addition is carried out modulo w.

Next, we must determine the convergence rate of $\hat{S}_{n,w}$. We look at two separate cases, for odd w and even w, due to the difference in periodicity. $\hat{S}_{n,w}$ is aperiodic when w is odd, and $\hat{S}_{n,w}$ has period 2 when w is even.

If w is odd, it follows from Proposition 15 in [114] that

$$||P^n(x, \cdot) - \pi^*|| \leq \sqrt{w}|\lambda_2|^n, \tag{B.2}$$

where $||\cdot||$ denotes the total variation norm, and $\lambda_1 = 1$ and $\lambda_2, \lambda_3, \ldots, \lambda_w$ are the other eigenvalues of the transition matrix P, written in decreasing order of absolute value, with $|\lambda_i| < 1$ for $i > 1$.

In the case where w is even if n and x_{d+1} have the same parity, then $\mathbb{P}(\hat{S}_{n,w} = x_{d+1})$ converges to $2 \cdot \frac{1}{w} = \frac{2}{w}$ [115]. Furthermore by modifying the proof of Proposition 15 in [114], we obtain

$$||P^{2n}(0, \cdot) - \pi_1|| \leq \sqrt{\frac{w}{2}}|\hat{\lambda}_2|^n, \tag{B.3}$$

where $\pi_1 = \begin{pmatrix} \frac{2}{w} & 0 & \frac{2}{w} & 0 & \frac{2}{w} & 0 & \ldots & \frac{2}{w} & 0 \end{pmatrix}$ and $\hat{\lambda}_2$ is the second largest eigenvalue of P^2, which satisfies $|\hat{\lambda}_2| < 1$. In addition,

$$||P^{2n+1}(0, \cdot) - \pi_2|| \leq \sqrt{\frac{w}{2}}|\hat{\lambda}_2|^n, \tag{B.4}$$

where $\pi_2 = \begin{pmatrix} 0 & \frac{2}{w} & 0 & \frac{2}{w} & 0 & \frac{2}{w} & \ldots & 0 & \frac{2}{w} \end{pmatrix}$.

If w is odd, then (B.2) implies that

$$\frac{1}{w} - \sqrt{w}|\lambda_2|^{n-k} \leq \mathbb{P}(\hat{S}_{n-k,w} = x_{d+1}) \leq \frac{1}{w} + \sqrt{w}|\lambda_2|^{n-k}.$$

If w is even, then $\mathbb{P}(\hat{S}_{n-k,w} = x_{d+1} | \hat{N}(n) = k) = 0$ if $n - k$ and x_{d+1} have different parity. If $n - k$ and x_{d+1} have the same parity then inequalities (B.3) and (B.4) show that

for even w,

$$\frac{2}{w} - \sqrt{\frac{w}{2}}|\hat{\lambda}_2|^{n-k} \leq \mathbb{P}(\hat{S}_{n-k,w} = x_{d+1}) \leq \frac{2}{w} + \sqrt{\frac{w}{2}}|\hat{\lambda}_2|^{n-k}.$$

Since $k \in A_n(\nu)$ we know that $k \leq n(\mu_{w,d} + \varepsilon_n)$ and the result follows. ■

Next, we obtain bounds for $\mathbb{P}(\hat{S}_k = \hat{\mathbf{x}})$ when $k \in A_n(\nu)$ by using a local central limit theorem on $\mathbb{Z}^d$. For positive integer n define the multivariate normal pdf

$$p_n(\hat{\mathbf{x}}) = \left(\frac{d}{2\pi n}\right)^{d/2} \exp\left[-|\hat{\mathbf{x}}|^2/(2nd)\right],$$

where $|\cdot|$ is the d-dimensional Euclidean norm. Then Theorem 2.1.3 of [116] states that if $\mathbb{P}(\hat{S}_n = \hat{\mathbf{x}}) \neq 0$ there exists a constant c such that

$$|\mathbb{P}\left(\hat{S}_n = \hat{\mathbf{x}}\right) - 2p_n(\hat{\mathbf{x}})| \leq \frac{c}{n^{(d+2)/2}}\left[\left(\frac{|\hat{\mathbf{x}}|^4}{n^2} + 1\right)e^{-|\hat{\mathbf{x}}|^2/(2nd)} + o(1)\right]. \tag{B.5}$$

This of course implies that if n and $\hat{\mathbf{x}}$ have the same parity then

$$n^{d/2}\mathbb{P}(\hat{S}_n = \hat{\mathbf{x}}) = 2\left(\frac{d}{2\pi}\right)^{d/2} + o(1). \tag{B.6}$$

We are now ready to state our discrete time LCLT.

Proposition B.1.2. *If w is odd then*

$$n^{d/2}\mathbb{P}(S_n = \mathbf{x}) = \frac{1}{w}\left(\frac{d}{2\pi\mu_{w,d}}\right)^{d/2} + o(1),$$

as $n \to \infty$.

If w is even and $\mathbb{P}(S_n = \mathbf{x}) > 0$ then

$$n^{d/2}\mathbb{P}(S_n = \mathbf{x}) = \frac{2}{w}\left(\frac{d}{2\pi\mu_{w,d}}\right)^{d/2} + o(1),$$

as $n \to \infty$.

Proof.

Start with the following decomposition,

$$\begin{aligned} n^{d/2}\mathbb{P}(S_n = \mathbf{x}) = {} & n^{d/2}\mathbb{P}(S_n = \mathbf{x}|\hat{N}(n) \in A_n(\nu))\mathbb{P}(\hat{N}(n) \in A_n(\nu)) \\ & + n^{d/2}\mathbb{P}(S_n = \mathbf{x}|\hat{N}(n) \notin A_n(\nu))\mathbb{P}(\hat{N}(n) \notin A_n(\nu)). \end{aligned}$$

Define

$$B_1(n) = n^{d/2}\,\mathbb{P}(S_n = \mathbf{x}|\hat{N}(n) \in A_n(\nu))\mathbb{P}(\hat{N}(n) \in A_n(\nu)),$$

and note that

$$B_1(n) \leq n^{d/2}\mathbb{P}(S_n = \mathbf{x}) \leq B_1(n) + n^{d/2}\mathbb{P}(\hat{N}(n) \notin A_n(\nu)). \tag{B.7}$$

Hoeffding's Inequality [117] gives a bound on the tail of the binomial distribution, to conclude that

$$\mathbb{P}(\hat{N}(n) \notin A_n(\nu)) \leq 2\exp\left(-2n^{2\nu}\right).$$

and we conclude that it suffices to study the large n asymptotics of

$$n^{d/2}\mathbb{P}(S_n = \mathbf{x}|\hat{N}(n) \in A_n(\nu)).$$

Using the law of total probability, we obtain

$$\begin{aligned}
n^{d/2}\mathbb{P}(S_n =&\mathbf{x}|\hat{N}(n) \in A_n(\nu)) = n^{d/2}\mathbb{P}(\hat{S}_{\hat{N}(n)} = \hat{\mathbf{x}}, \hat{S}_{n-\hat{N}(n),w} = x_{d+1}|\hat{N}(n) \in A_n(\nu)) \\
&= n^{d/2}\sum_{k\in A_n(\nu)} \mathbb{P}(\hat{S}_k = \hat{\mathbf{x}}, \hat{S}_{n-k,w} = x_{d+1}|\hat{N}(n) = k)\mathbb{P}(\hat{N}(n) = k|\hat{N}(n) \in A_n(\nu)) \\
&= n^{d/2}\sum_{k\in A_n(\nu)} \mathbb{P}(\hat{S}_k = \hat{\mathbf{x}})\mathbb{P}(\hat{S}_{n-k,w} = x_{d+1})\mathbb{P}(\hat{N}(n) = k|\hat{N}(n) \in A_n(\nu)).
\end{aligned}$$

Without loss of generality assume that $\mathbb{P}(\hat{S}_k = \mathbf{x}) > 0$ for k even, and define

$$A_n(\epsilon, E) = \{k \in A_n(\nu) : k \text{ is even}\}.$$

Then we can apply (B.6)

$$\begin{aligned}
&n^{d/2}\mathbb{P}(S_n = \mathbf{x}|\hat{N}(n) \in A_n(\nu)) \\
&= \sum_{k\in A_n(\epsilon,E)} \left(\frac{n}{k}\right)^{d/2} k^{d/2}\mathbb{P}(\hat{S}_k = \hat{\mathbf{x}})\mathbb{P}(\hat{S}_{n-k,w} = x_{d+1})\mathbb{P}(\hat{N}(n) = k|\hat{N}(n) \in A_n(\nu)) \\
&= \sum_{k\in A_n(\epsilon,E)} \left(\frac{n}{k}\right)^{d/2} (2\left(\frac{d}{2\pi}\right)^{d/2} + o(1))\mathbb{P}(\hat{S}_{n-k,w} = x_{d+1})\mathbb{P}(\hat{N}(n) = k|\hat{N}(n) \in A_n(\nu)).
\end{aligned}$$

Since $k \in A_n(\epsilon, E)$ implies that $n\mu_{w,d} - n^{1/2+\nu} \leq k \leq n\mu_{w,d} + n^{1/2+\nu}$ we obtain

$$\frac{n}{k} = \frac{1}{\mu_{w,d}} + o(1).$$

Therefore

$$\begin{aligned} &n^{d/2}\mathbb{P}(S_n = \mathbf{x}|\hat{N}(n) \in A_n(\nu)) \\ &= 2\left(\frac{d}{2\pi\mu_{w,d}}\right)^{d/2} \sum_{k \in A_n(\epsilon,E)} \mathbb{P}(\hat{S}_{n-k,w} = x_{d+1})\mathbb{P}(\hat{N}(n) = k|\hat{N}(n) \in A_n(\nu)) + o(1). \end{aligned}$$

If w is odd, then Proposition B.1.1 gives

$$\begin{aligned} &n^{d/2}\mathbb{P}(S_n = \mathbf{x}|\hat{N}(n) \in A_n(\nu)) \\ &= \frac{2}{w}\left(\frac{d}{2\pi\mu_{w,d}}\right)^{d/2} \sum_{k \in A_n(\epsilon,E)} \mathbb{P}(\hat{N}(n) = k|\hat{N}(n) \in A_n(\nu)) + o(1). \end{aligned}$$

Similarly, if w is even and $\mathbb{P}(S_n = \mathbf{x}) > 0$ then

$$\begin{aligned} &n^{d/2}\mathbb{P}(S_n = \mathbf{x}|\hat{N}(n) \in A_n(\nu)) \\ &= \frac{4}{w}\left(\frac{d}{2\pi\mu_{w,d}}\right)^{d/2} \sum_{k \in A_n(\epsilon,E)} \mathbb{P}(\hat{N}(n) = k|\hat{N}(n) \in A_n(\nu)) + o(1). \end{aligned}$$

Recall that if $X \sim \text{Bin}(n, p)$ then $\mathbb{P}(X \text{ is even}) = \frac{1}{2} + \frac{1}{2}(1-2p)^n$, therefore

$$\sum_{k \in A_n(\epsilon,E)} \mathbb{P}(\hat{N}(n) = k|\hat{N}(n) \in A_n(\nu)) = \frac{1}{2} + o(1),$$

and the result follows. ■

With the discrete time result established we prove the continuous time result.

Proof.[Proof of Theorem 3.2.2]

For $\nu \in (0, 1/2)$ define the set

$$B_t(\nu) = \{n \geq 1 : |n - \alpha t| \leq t^{1/2+\nu}\}.$$

Using standard large deviation bounds for the Poisson distribution (see e.g.,Theorem 1 of

[118]) we conclude that

$$P(N(t) \notin B_t(\nu)) \leq 2\exp\left[\frac{-t^{2\nu}}{2(\alpha + t^{\nu-1/2})}\right].$$

Therefore there exists a constant c such that

$$\sum_{n\in B_t(\nu)} \mathbb{P}(S_n = \mathbf{x})P(N(t) = n) \leq \mathbb{P}(Y_t = \mathbf{x}) \leq \sum_{n\in B_t(\nu)} \mathbb{P}(S_n = \mathbf{x})P(N(t) = n) + e^{-ct^{2\nu}}.$$

It suffices to find the asymptotics of

$$\begin{aligned}&(\alpha t)^{d/2} \sum_{n\in B_t(\nu)} \mathbb{P}(S_n = \mathbf{x})\mathbb{P}(N(t) = n)\\ &= \alpha^{d/2} \sum_{n\in B_t(\nu)} \left(\frac{t}{n}\right)^{d/2} n^{d/2}\mathbb{P}(S_n = \mathbf{x})\mathbb{P}(N(t) = n);\end{aligned}$$

since $n \in B_t(\nu)$ we know that $t/n = 1/\alpha + o(1)$ and can rewrite above as

$$= \sum_{n\in B_t(\nu)} n^{d/2}\mathbb{P}(S_n = \mathbf{x})\mathbb{P}(N(t) = n) + o(1).$$

If w is odd, the result then follows by using the result from Proposition B.1.2.

If w is even, then assume without loss of generality that $P(S_n = \mathbf{x}) > 0$ for n even and is 0 otherwise. Define the set

$$B_t(\nu, E) = \{n \in B_t(\nu) : n \text{ even integer}\},$$

and then using Proposition B.1.2

$$\begin{aligned}&\sum_{n\in B_t(\nu)} n^{d/2}\mathbb{P}(S_n = \mathbf{x})\mathbb{P}(N(t) = n) + o(1)\\ &= \sum_{n\in B_t(\nu,E)} n^{d/2}\mathbb{P}(S_n = \mathbf{x})\mathbb{P}(N(t) = n) + o(1)\\ &= \frac{2}{w}\left(\frac{d}{2\pi\mu_{w,d}}\right)^{d/2} \sum_{n\in B_t(\nu,E)} \mathbb{P}(N(t) = n) + o(1).\end{aligned}$$

Recall that if $X \sim \text{Pois}(\lambda)$ then $P(X = \text{even}) = 1/2 + e^{-2\lambda}/2$ and therefore as $t \to \infty$

$$P(N(t) \in B_t(\nu, E)) = 1/2 + o(1),$$

and the result follows. ■

B.2 Return time to the origin

Below we prove the return time result, stated in Theorem 3.2.3. As defined in section B.1, let Y_t be a simple random walk on $\mathbb{Z}^d \times \mathbb{Z}_w$ with jump rate α. In this case, we assume $d = 2$. $N(t)$ is the number of jumps Y_t has taken by time t, and recall $N(t) \sim Pois(\alpha t)$.

Proof.[Proof of Theorem 3.2.3]

We start by looking at the return time on S_n, the embedded discrete time walk on $\mathbb{Z}^2 \times \mathbb{Z}_w$. Let τ_0 be the first discrete time step that S_n returns to 0. That is,

$$\tau_0 = \min\{n > 0 : S_n = 0\}.$$

We can use Proposition B.1.2 and a classic result of Dvoretzky and Erdos ([119]) on the return time for a discrete time walk on $\mathbb{Z}^2$ to see that

$$\mathbb{P}(\tau_0 > n) \sim \begin{cases} \frac{4\pi w}{5 \log n} & w = 2 \\ & \text{as } n \to \infty. \\ \frac{2\pi w}{3 \log n} & w > 2. \end{cases} \tag{B.8}$$

Next, we translate this result to the return time asymptotics for the continuous time random walk Y_t. To do this, we start by fixing $\epsilon > 0$. Recall that $T_0 = \inf\{t > 0 : Y_t = 0\}$ and that $N(t)$ is the number of jumps Y_t has taken by time t. We condition $\mathbb{P}(T_0 > t)$ as follows

$$\begin{aligned} \mathbb{P}(T_0 > t) &= \mathbb{P}(\tau_0 > N(t)) \\ &= \mathbb{P}\left(\tau_0 > N(t),\ N(t) \geq \alpha t(1-\epsilon)\right) + \mathbb{P}\left(\tau_0 > N(t),\ N(t) < \alpha t(1-\epsilon)\right) \\ &\leq \mathbb{P}\left(\tau_0 > \alpha t(1-\epsilon)\right) + \mathbb{P}\left(N(t) < \alpha t(1-\epsilon)\right). \end{aligned} \tag{B.9}$$

In addition we have,

$$\begin{aligned} \mathbb{P}\left(\tau_0 > \alpha t(1+\epsilon)\right) &= \mathbb{P}\left(\tau_0 > \alpha t(1+\epsilon), N(t) \leq \alpha t(1+\epsilon)\right) + \mathbb{P}\left(\tau_0 > \alpha t(1+\epsilon), N(t) > \alpha t(1+\epsilon)\right) \\ &\leq \mathbb{P}\left(\tau_0 > N(t)\right) + \mathbb{P}\left(N(t) > \alpha t(1+\epsilon)\right), \end{aligned}$$

and therefore

$$\mathbb{P}\left(\tau_0 > N(t)\right) \geq \mathbb{P}\left(\tau_0 > \alpha t(1+\epsilon)\right) - \mathbb{P}\left(N(t) > \alpha t(1+\epsilon)\right). \tag{B.10}$$

We combine bounds (B.9) and (B.10) to obtain the following,

$$\begin{aligned}\mathbb{P}\left(\tau_0 > \alpha t(1+\epsilon)\right) - \mathbb{P}\left(N(t) > \alpha t(1+\epsilon)\right) &\leq \mathbb{P}\left(\tau_0 > N(t)\right) \\ &\leq \mathbb{P}\left(\tau_0 > \alpha t(1-\epsilon)\right) + \mathbb{P}\left(N(t) < \alpha t(1-\epsilon)\right).\end{aligned}$$

Since $\{N(t) : t \geq 0\}$ is a Poisson process with rate α by the law of large numbers, $\frac{N(t)}{t} \to \alpha$ almost surely as $t \to \infty$. It follows that

$$\lim_{t\to\infty} \mathbb{P}\left(N(t) > \alpha t(1+\epsilon)\right) = \lim_{t\to\infty} \mathbb{P}\left(N(t) < \alpha t(1-\epsilon)\right) = 0$$

By definition, we know $\mathbb{P}\left(\tau_0 > N(t)\right) = \mathbb{P}(T_0 > t)$, so it follows that

$$\lim_{t\to\infty} \mathbb{P}\left(\tau_0 > \alpha t(1+\epsilon)\right) \leq \lim_{t\to\infty} \mathbb{P}(T_0 > t) \leq \lim_{t\to\infty} \mathbb{P}\left(\tau_0 > \alpha t(1-\epsilon)\right).$$

By (B.8), if $w = 2$, we have as $t \to \infty$

$$\mathbb{P}\left(\tau_0 > \alpha t(1+\epsilon)\right) \sim \frac{4\pi w}{5\log(\alpha t)}, \quad \lim_{t\to\infty} \mathbb{P}\left(\tau_0 > \alpha t(1-\epsilon)\right) \sim \frac{4\pi w}{5\log(\alpha t)},$$

and a similar result in the case $w > 2$. Thus, we can conclude that as $t \to \infty$

$$\mathbb{P}(T_0 > t) \sim \begin{cases} \frac{4\pi w}{5\log(\alpha t)} & w = 2 \\ \\ \frac{2\pi w}{3\log(\alpha t)} & w > 2. \end{cases}$$

■

B.3 Error between dual process and BRW

In this section, we show that the error when approximating the pruned dual process $\hat{\zeta}_t^\beta$ with the branching random walk ζ_t^β approaches 0 as $\beta \to 0$.

First we determine a lower bound for the displacement between a parent particle and newborn daughter particle, who paths are independent random walks X^1 and X^2, at the time of successful branching in ζ_t. Let $X_0^1 = 0$, and let X_0^2 be uniformly distributed on the set $\{\pm e_i : i \in 1, 2, 3\}$. Let

$$L = \frac{1}{\sqrt{\beta}\log(1/\beta)}.$$

Note that $\mathbb{P}_x(\cdot) := \mathbb{P}\left(\cdot \middle| |\bar{X}_0| = x\right)$, where $\bar{X}_t = X_t^1 - X_t^2$. Recall that

$$T_0 = \inf\{t > 0 : \bar{X}_t = 0\}.$$

By (3.4), which was obtained using Theorem 3.2.3, we have

$$\begin{aligned}
\mathbb{P}_1\left(|X^1_{\tau(\beta)} - X^2_{\tau(\beta)}| \le L \,\middle|\, T_0 > \tau(\beta)\right) &\le \frac{1}{\mathbb{P}_1(T_0 > \tau(\beta))} \cdot \mathbb{P}_1\left(|X^1_{\tau(\beta)} - X^2_{\tau(\beta)}| \le L\right) \\
&\le \frac{3\log(1/\beta)}{2\pi w} \sum_{|x|\le L} \mathbb{P}_1\left(|X^1_{\tau(\beta)} - X^2_{\tau(\beta)}| = x\right) \\
&\le \frac{3\log(1/\beta)}{2\pi w} \cdot w\left(\frac{1}{\sqrt{\beta}\log(1/\beta)}\right)^2 \frac{3}{4\pi w\, \tau(\beta)} \\
&= \frac{9}{8\pi^2 w\beta\log(1/\beta)} \cdot \beta\sqrt{\log(1/\beta)} \\
&= O\left(\frac{1}{\sqrt{\log(1/\beta)}}\right), \qquad \text{(B.11)}
\end{aligned}$$

where the third inequality follows from the local central limit theorem 3.2.2 on $\mathbb{Z}^2 \times \mathbb{Z}_w$ and from the fact that $L > w$ as $\beta \to 0$.

Similarly, for any two particles with independent random walk paths X^1 and X^2, such

that $|\bar{X}_0| = x$,

$$\begin{aligned}\mathbb{P}_x\left(|X^1_{\tau(\beta)} - X^2_{\tau(\beta)}| \le L\right) &\le \sum_{|y|\le L} \mathbb{P}_x\left(|X^1_{\tau(\beta)} - X^2_{\tau(\beta)}| = y\right) \\ &\le w\left(\frac{1}{\sqrt{\beta}\log(1/\beta)}\right)^2 \frac{3}{4\pi w\,\tau(\beta)} \\ &= O\left(\frac{1}{(\log(1/\beta))^{3/2}}\right). \end{aligned} \tag{B.12}$$

Error components.

Possible error between $\hat{\zeta}^\beta_t$ and ζ^β_t arises from the following groups:

1. Any particle that coalesces with a particle other than its parent by time $\tau(\beta)$ after its birth (and descendants of such an ancestor in ζ^β_t)

2. Any particle that coalesces with another particle after surviving until $\tau(\beta)$ (and its descendants in ζ^β_t)

Recall that p_i is the index of the parent of the particle with path $\hat{Z}^i$, and that $(\hat{b}_j)_{j\in\mathbb{Z}^+}$,$(b_j)_{j\in\mathbb{Z}^+}$ are the branching times in $\hat{\zeta}_t$ and ζ_t, respectively. Also recall that $\hat{D}_t(i)$, $D_t(i)$ are the sets of all descendants of $\hat{Z}^i, Z^i$, respectively, by time t.

Each ancestor i (with path Z^i) in group 1 or group 2 has an associated $t_{e,i} < h(\beta)T$ and $j(i) \le i$ such that:

(A1) $Z^i_t = \hat{Z}^{j(i)}_t = \infty \;\; \forall t \in [0, b_i]$

(A2) $\hat{Z}^{j(i)}_t = \infty \; \forall t \in [t_{e,i}\,, h(\beta)T]$

(A3) $Z^i_t \in \zeta^\beta_t \; \forall t \in [b_i\,, h(\beta)T]$

The following property also holds for particles i in group 1:

(A4) $\hat{Z}^{j(i)}_t = \infty \; \forall t \in [0, h(\beta)T]$.

Let

$$E_1(\beta) = \left\{ D^\beta_T(i), Z^{i,\beta} \in Z^\beta : Z^i \text{ satisfy (A1)-(A4)} \right\}.$$

Ancestor particles i in group 2 also satisfy the following properties:

(B4) $t_{e,i} \in [b_i, h(\beta)T]$

(B5) $Z^i_t = \hat{Z}^{j(i)}_t \; \forall t \in [b_i\,, t_{e,i}]$.

Let

$$E_2(\beta) = \left\{ D_T^\beta(i), Z^{i,\beta} \in Z^\beta : Z^i \text{ satisfy (A1)-(A3), (B4)-(B5) } \right\}.$$

Note that for ease of notation, we will suppress the β notation in $E_1(\beta), E_2(\beta)$ for the remainder of the section.

Inductive argument.

We will use (B.11) and (B.12) in an inductive proof showing the error between $\hat{Z}^\beta$ and Z^β is small as β approaches 0.

First, in order to make this error precise, recall that

$$d_{p,T}(Z^\beta, \hat{Z}^\beta) = \sup_i \left\{ d_T(Z^{i,\beta}, \hat{Z}^{i,\beta}) \right\},$$

where d_T and $d_{p,T}$ are the metrics for the Skorokhod topology $\mathbf{\hat{D}}$ and the product Skorokhod topology $\mathcal{D}$, as defined in section 3.3.6.

Then given $\delta > 0$, we have

$$\mathbb{P}\left(d_{p,T}(Z^\beta, \hat{Z}^\beta) > \delta \right) = \mathbb{P}(|E_1 \cup E_2| > 0).$$

In order to show that this probability converges to 0 as $\beta \to 0$, we will first prove a lemma about the hitting time for two random walks.

Recall that $\bar{X}_t = X_t^1 - X_t^2$ for X^1, X^2 independent simple random walks, and that $L = \frac{1}{\sqrt{\beta}\log(1/\beta)}$. We define

$$\gamma(\beta) = \mathbb{P}\left(\bar{X}_t = 0 \text{ for some } t \le h(\beta)T \;\middle|\; \bar{X}_0 = x_0, |x_0| > L - 1 \right). \tag{B.13}$$

Let $\tilde{X}_t$ be the projection of $\bar{X}_t$ onto $\mathbb{Z}^2$. That is,

$$X_t^1 = (x_t^1, y_t^1, z_t^1), X_t^2 = (x_t^2, y_t^2, z_t^2) \implies \tilde{X}_t = \left(x_t^1 - x_t^2, y_t^1 - y_t^2 \right),$$

and let

$$\tilde{\gamma}(\beta) = \mathbb{P}\left(\tilde{X}_t = 0 \text{ for some } t \le h(\beta)T \;\middle|\; \tilde{X}_0 = x_0, |x_0| > L - w - 1 \right).$$

Note that $|\bar{X}_t| \ge |\hat{\tilde{X}}_t|$, so $\gamma \le \hat{\gamma}$.

Lemma B.3.1. *Let $h(\beta) = \frac{1}{\beta}\log\left(\frac{1}{\beta}\right)$, and let $\tilde{X}_t$ be a random walk on $\mathbb{Z}^2$ with jump*

distribution p. Then

$$\lim_{\beta \to 0} \hat{\gamma}(\beta) = 0.$$

Proof.

Adapted from the paper by Durrett and Zahle [120].

Let $a(x) = \sum_{k=0}^{\infty} [p_k(0) - p_k(x)]$ be the potential kernel for $\tilde{X}_t$. Then we have

$$\begin{aligned}
\sum_{y \in \mathbb{Z}^2} p(y-x)a(y) \; - a(x) &= \sum_{y \in \mathbb{Z}^2} \sum_{k=0}^{\infty} p(y-x)[p_k(0) - p_k(y)] - \sum_{k=0}^{\infty} [p_k(0) - p_k(x)] \\
&= \lim_{N \to \infty} \sum_{y \in \mathbb{Z}^2} \sum_{k=0}^{N} p(y-x)[p_k(0) - p_k(y)] - \lim_{N \to \infty} \sum_{k=0}^{N} [p_k(0) - p_k(x)] \\
&= \lim_{N \to \infty} \sum_{y} p(y-x) \sum_{k=0}^{N} p_k(0) - \lim_{N \to \infty} \sum_{y} \sum_{k=0}^{N} p(y-x) p_k(y) - \\
&\quad \lim_{N \to \infty} \sum_{k=0}^{N} [p_k(0) - p_k(x)] \\
&= \lim_{N \to \infty} \sum_{k=0}^{N} p_k(0) - \lim_{N \to \infty} \sum_{k=0}^{N} p_{k+1}(x) - \lim_{N \to \infty} \sum_{k=0}^{N} [p_k(0) - p_k(x)] \\
&= \lim_{N \to \infty} \left[\sum_{k=0}^{N} p_k(x) - \sum_{k=0}^{N} p_{k+1}(x) \right] \\
&= p_0(x) \\
&= \delta(0, x) \qquad \text{(B.14)}
\end{aligned}$$

Then for a random walk S_n with jump distribution p, where $\hat{t} = \min\{n > 0 : S_n = 0\}$, we have

$$\begin{aligned}
\mathbb{E}[a(S_{n+1}) | S_1, \ldots, S_n] &= \sum_{x \in \mathbb{Z}^2} a(S_n + x) p(x) \\
&= \sum_{y \in \mathbb{Z}^2} a(y) p(y - S_n) \\
&= a(S_n) + \delta(0, S_n), \qquad \text{(B.15)}
\end{aligned}$$

where the last equality follows from (B.14). Thus, $a(S_{n \wedge \hat{t}})$ is a martingale.

In Theorem 2 of [121], Fukai and Uchiyama show that

$$a(x) = C \log |x| + O(1). \tag{B.16}$$

Let

$$R = \frac{1}{\sqrt{\beta}} \log(1/\beta),$$

and let $B(R) = \{x \in \mathbb{R}^2 : |x| \leq R\}$, $B(1) = \{x \in \mathbb{R}^2 : |x| \leq 1\}$. Then define

$$\lambda(\beta) = \mathbb{P}(T_1 < T_R | \tilde{X}_0 = x_0, |x_0| = L - w - 1),$$

where

$$T_1 = \min\{t > 0 : |\tilde{X}_t| \leq 1\}, \;\; T_R = \min\{t > 0 : |\tilde{X}_t| \geq R\}.$$

Let $S = \min\{T_1, T_R\}$. $a(\tilde{X}_{t \wedge T_0})$ is a martingale by (B.15), and S is a stopping time, so by the optional sampling theorem and (B.16), we obtain

$$\begin{aligned}
a(x_0) &= \mathbb{E}[a(\tilde{X}_S)] \implies \\
\log(L - w - 1) + O(1) &= \lambda(\log 1 + O(1)) + (1 - \lambda)(\log R + O(1)) \implies \\
\frac{1}{2} \log(1/\beta) - \log\log(1/\beta) + O(1) &= (1 - \lambda) \left(\frac{1}{2} \log(1/\beta) + \log\log(1/\beta) + O(1) \right) \implies \\
\frac{1}{2} \log(1/\beta) - \frac{1 - \lambda}{2} \log(1/\beta) &= (2 - \lambda) \log\log(1/\beta) + O(1) \implies \\
\frac{\lambda}{2} \log(1/\beta) &= (2 - \lambda) \log\log(1/\beta) + O(1) \\
&\leq 2 \log\log(1/\beta) + O(1)
\end{aligned}$$

Therefore,

$$\lim_{\beta \to 0} \lambda(\beta) \leq \lim_{\beta \to 0} \frac{4 \log\log(1/\beta) + O(1)}{\log(1/\beta)} = 0. \tag{B.17}$$

We use this to prove that $\tilde{\gamma} \to 0$ as $\beta \to 0$. Note that

$$\tilde{\gamma} \leq \lambda + \mathbb{P}\left(\inf_{t \leq h(\beta) T} |\tilde{X}_t| \leq 1 \middle| |\tilde{X}_0| \leq R \right). \tag{B.18}$$

Let $\tilde{Y}_t := \tilde{X}_t - \tilde{X}_0$. Then $\tilde{Y}_0 = 0$ and

$$\mathbb{P}\left(\inf_{t\le h(\beta)T} |\tilde{X}_t| \le 1 \middle| |\tilde{X}_0| \le R\right) = \mathbb{P}\left(\sup_{t\le h(\beta)T} |\tilde{Y}_t| \ge R-1 \middle| \tilde{Y}_0 = 0\right).$$

Since $\tilde{Y}_t$ is a symmetric random walk, $|\tilde{Y}_t|$ is a submartingale. Then by Doob's martingale inequality,

$$\begin{aligned}
\mathbb{P}\left(\sup_{t\le h(\beta)T} |\tilde{Y}_t| \ge R-1 \middle| \tilde{Y}_0 = 0\right) &\le \frac{1}{R-1}\mathbb{E}[|\tilde{Y}_{h(\beta)T}|] \\
&\le \frac{1}{R-1}\left(\mathbb{E}\left[\left(\tilde{Y}^{(1)}_{h(\beta)T}\right)^2 + \left(\tilde{Y}^{(2)}_{h(\beta)T}\right)^2\right]\right)^{1/2} \\
&= \frac{\sqrt{2}}{R-1}\left(\mathbb{E}\left[\left(\tilde{Y}^{(1)}_{h(\beta)T}\right)^2\right]\right)^{1/2} \\
&= \frac{C\sqrt{2h(\beta)T}}{R-1} \\
&= O\left(\frac{1}{\log(1/\beta)}\right)^{1/2}
\end{aligned}$$

Then by (B.18), we have

$$\tilde{\gamma}(\beta) \le \lambda(\beta) + O\left(\frac{1}{\log(1/\beta)}\right)^{1/2},$$

and by (B.17) we can conclude that

$$\lim_{\beta\to 0} \tilde{\gamma}(\beta) = 0.$$

■

Since $\gamma \le \hat{\gamma}$, we also have

$$\lim_{\beta\to 0} \gamma(\beta) = 0. \tag{B.19}$$

We use (B.19) in the following theorem to show that $\mathbb{P}(|E_1 \cup E_2| > 0) \to 0$ as $\beta \to 0$.

Lemma B.3.2. *Let $h(\beta) = \frac{1}{\beta}\log(1/\beta)$, with $\beta > 0$, and let $\delta > 0$. Let $\hat{\zeta}^\beta_t$ be a branching coalescing random walk with branching rate $\beta h(\beta)$, and let ζ^β_t be a branching random walk with branching rate μ_β, as described by (3.3). Then*

$$\lim_{\beta\to 0} \mathbb{P}\left(d_{p,T}(Z^\beta, \hat{Z}^\beta) > \delta\right) = 0.$$

Proof.

Recall that $\mathbb{P}\left(d_{p,T}(Z^\beta, \hat{Z}^\beta) > \delta\right) = \mathbb{P}(|E_1 \cup E_2| > 0)$. Therefore it suffices to show that $\lim_{\beta\to 0} \mathbb{P}(|E_1 \cup E_2| > 0) = 0$.

Let $\epsilon > 0$, and let $\hat{N}_T = |\hat{\zeta}_T^\beta|$, $N_T = |\zeta_T^\beta|$. Recall ζ_t^β has branching rate $\mu_\beta \leq \frac{2\pi w}{3}$ by (3.3), so

$$\mathbb{E}[\hat{N}_T] \leq \mathbb{E}[N_T] \leq K \exp\left(\frac{2\pi w}{3} T\right).$$

Hence, $\exists M_\epsilon$ such that $\mathbb{P}(N_T > M_\epsilon) < \epsilon$. We will assume from now on that $\hat{N}_T \leq M_\epsilon$.

We will show by induction that $\lim_{\beta\to 0} \mathbb{P}(|E_1 \cup E_2| > 0) = 0$. Recall $L = \dfrac{1}{\sqrt{\beta}\log(1/\beta)}$.

Suppose $|\hat{\zeta}_0^\beta| = K$ with $|\hat{Z}_0^i - \hat{Z}_0^j| \geq L$ for all pairs $\hat{Z}_0^i, \hat{Z}_0^j \in \hat{\zeta}_0$. Note that $\hat{\zeta}_0 = \zeta_0$. The probability of coalescence between each pair $i, j \leq K$ of initial particles is

$$\mathbb{P}(\exists t \in [0, h(\beta)T] \text{ s.t. } \hat{Z}_t^i = \hat{Z}_t^j) \leq \gamma,$$

as defined in (B.13).

There are at most $\frac{M_\epsilon^2}{2}$ pairs of particles, so

$$\sum_{i,j\leq K} \mathbb{P}\left(\exists t \in [0, h(\beta)T] \text{ s.t. } \hat{Z}_t^i = \hat{Z}_t^j\right) \leq \frac{K^2\gamma}{2} \leq \frac{M_\epsilon^2\gamma}{2}. \tag{B.20}$$

Let $\hat{Z}^{K+1}$ be the path of the first particle that is born that does not coalesce with its parent before surviving for time $\tau(\beta)$.

$$|\hat{Z}^{K+1}_{b_{(K+1)}} - \hat{Z}^{p_{(K+1)}}_{b_{(K+1)}}| = 1 \implies |\hat{Z}^{K+1}_{b_{(K+1)}} - \hat{Z}^{j}_{b_{(K+1)}}| > L - 1 \quad \forall \left(j : j < K, j \neq p_{(K+1)}\right).$$

Then we have

$$\sum_{j\leq K, j\neq p_{(K+1)}} \mathbb{P}\left(\exists t \in [b_{(K+1)}, h(\beta)T] \text{ s.t. } \hat{Z}_t^{K+1} = \hat{Z}_t^j\right) \leq K\gamma \leq M_\epsilon\gamma. \tag{B.21}$$

Hence, we have

$$\mathbb{P}(\exists Z^{i,\beta} \in Z^\beta \text{ with } i \leq K+1 : Z^{i,\beta} \text{ satisfy (A1)-(A4)}) +$$
$$\mathbb{P}(\exists Z^{i,\beta} \in Z^\beta \text{ with } i \leq (K+1) : Z^{i,\beta} \text{ satisfy (A1)-(A3), (B4)-(B5)}) \quad \leq \frac{M_\epsilon^2 \gamma}{2} + M_\epsilon \gamma.$$

Result (B.11) implies that

$$\mathbb{P}\left(\left|\hat{Z}^{K+1}_{b_{(K+1)}+\tau(\beta)} - \hat{Z}^{p_{(K+l)}}_{b_{(K+1)}+\tau(\beta)}\right| \leq L\right) = O\left(\frac{1}{\sqrt{\log(1/\beta)}}\right).$$

Additionally, by (B.12), $\forall\left(j : j \leq K, j \neq p_{(K+1)}\right)$,

$$\mathbb{P}\left(\left|\hat{Z}^{K+1}_{b_{(K+1)}+\tau(\beta)} - \hat{Z}^{j}_{b_{(K+1)}+\tau(\beta)}\right| \leq L\right) = O\left(\frac{1}{(\log(1/\beta))^{3/2}}\right).$$

Thus for sufficiently small β, we have $K+1$ particles separated by a distance of at least L at time $b_{(K+1)} + \tau(\beta)$, with high probability. We repeat this argument at most M_ϵ times, so it follows that

$$\mathbb{P}(|E_1(\beta) \cup E_2(\beta)| > 0) \leq M_\epsilon \left(\frac{M_\epsilon^2 \gamma(\beta)}{2} + M_\epsilon \gamma(\beta) + O\left(\frac{1}{\sqrt{\log(1/\beta)}}\right)\right).$$

Therefore, by (B.19),

$$\lim_{\beta \to 0} \mathbb{P}(|E_1(\beta) \cup E_2(\beta)| > 0) = 0, \tag{B.22}$$

completing the proof. ■

B.4 Lower bound proof

In this section, we prove Lemma B.4.1, which provides a lower bound for the propagation speed. We restate the lemma here.

Lemma B.4.1. *Let $\{\xi_{t,\beta} : t \geq 0\}$ be a biased voter model on $\mathbb{Z}^2 \times \mathbb{Z}_w$ with fitness advantage $\beta > 0$ and define $\tau_\emptyset = \min\{t : \xi_t^0 = \emptyset\}$, Then for any $\epsilon > 0$*

$$\lim_{\beta \to 0} \lim_{t \to \infty} \mathbb{P}\left(\xi_{t,\beta} \cap \{x_\varepsilon(t;\beta)\} \neq \emptyset | \tau_\emptyset = \infty\right) = 1,$$

where $x_\varepsilon(t;\beta) = \lfloor a(w,\beta) t e_1 (1-\varepsilon) \rfloor$.

Proof.

First we will approximate the branching Brownian motion χ_t with branching rate

$$\mu = \begin{cases} \frac{4\pi w}{5} & w = 2 \\ \frac{2\pi w}{3} & w > 2. \end{cases}$$

using a block construction. This is a modified version of the construction in [120]. Let

$$\rho = \begin{cases} \sqrt{\frac{4\mu}{5}} & w = 2 \\ \sqrt{\frac{2\mu}{3}} & w > 2. \end{cases} \tag{B.23}$$

Then let

$$\mathcal{L}_0 = [-\rho L, \rho L]^2 \times \mathbb{R}_w, \quad \mathcal{L}_m = 2m\rho L e_1 + \mathcal{L}_0 \text{ for } m \in \mathbb{Z},$$

where L is a large constant. Let $\bar{\chi}_t$ be the modified version of χ_t, in which particles are killed when they land outside of $[-4\rho L, 4\rho L]^2 \times \mathbb{R}_w$. We modify the process in this way so that we can compare it to a 1-dependent percolation.

The notation $\bar{\chi}_t^x$ indicates that $\bar{\chi}_0 = \{x\}$. For any $A \in \mathbb{R}^2 \times \mathbb{R}_w$, let

$$\bar{\chi}_t^x(A) := |\bar{\chi}_t^x \cap A|.$$

Let $\bar{M}_t^x$ be a Brownian motion starting at x that is killed when it lands outside $[-4\rho L, 4\rho L]^2 \times \mathbb{R}_w$. The particles in $\bar{\chi}_t$ move as independent Brownian motion, so it follows that

$$\mathbb{E}[\bar{\chi}_t^x(A)] = \exp(\mu t)\, \mathbb{P}(\bar{M}_t^x \in A). \tag{B.24}$$

Due to the scaling invariance of Brownian motion, we have

$$\mathbb{P}(\bar{M}_{L^2}^x \in \mathcal{L}_0) = \mathbb{P}^{x/L}\left(M_s \in D_\rho \text{ for } s \le 1, M_1 \in [-\rho, \rho]^2 \times \mathbb{R}_w\right),$$

where M_s is a Brownian motion with $\sigma^2 = \frac{1}{3}$, and $D_\rho = [-4\rho, 4\rho]^2 \times \mathbb{R}_w$.

Thus, we have

$$\liminf_{L\to\infty} \inf_{x \in \mathcal{L}_0} \mathbb{P}(\bar{M}_{L^2}^x \in \mathcal{L}_1) \ge \inf_{y \in [-\rho,\rho]^2 \times \mathbb{R}_w} \mathbb{P}^y\left(M_s \in D_\rho \text{ for } s \le 1, M_1 \in [\rho, 3\rho]^2 \times \mathbb{R}_w\right) > 0,$$

and analogously, $\liminf_{L\to\infty} \inf_{x\in\mathcal{L}_0} \mathbb{P}(\bar{M}^x_{L^2} \in \mathcal{L}_{-1}) > 0$.

Then by (B.24), we can find L large enough so that

$$\inf_{x\in\mathcal{L}_0} \mathbb{E}[\bar{\chi}^x_{L^2}(\mathcal{L}_i)] \geq 2 \quad \text{for } i = -1, 1 \tag{B.25}$$

$\mathbb{E}[\chi^x_{L^2}(\mathbb{R}^2 \times \mathbb{R}_w)^2]$ is a finite number that depends on L, as shown in [99], so we will say

$$\mathbb{E}[\chi^x_{L^2}(\mathbb{R}^2 \times \mathbb{R}_w)^2] = C_L < \infty.$$

For $\epsilon > 0$, let A be a set of sites, such that $|A| \geq \dfrac{C_L}{\epsilon}$ and $A \subset [-\rho L, \rho L]^2 \times \mathbb{R}_w$.

If $\bar{\chi}_t$ starts with one particle at each site in A, then we have

$$\mu := \mathbb{E}[\bar{\chi}^A_{L^2}(\mathcal{L}_i)] \geq 2|A| \quad \text{for } i = -1, 1. \tag{B.26}$$

Note that $\mathbb{E}[\bar{\chi}^x_{L^2}(\mathcal{L}_i)^2] \leq \mathbb{E}[\chi^x_{L^2}(\mathbb{R}^2 \times \mathbb{R}_w)^2] = C_L$. Due to the independence of each particle in $\bar{\chi}$, we obtain the following:

$$\begin{aligned} \mathrm{Var}[\bar{\chi}^A_{L^2}(\mathcal{L}_i)] &= \sum_{x\in A} \mathrm{Var}[\bar{\chi}^x_{L^2}(\mathcal{L}_i)] \\ &\leq \sum_{x\in A} (C_L - 2^2) \\ &\leq |A| C_L \quad \text{for } i = -1, 1. \end{aligned} \tag{B.27}$$

Let $K \leq |A|$ for $A \subset \mathcal{L}_0$. It follows from (B.26) and (B.27) that

$$\begin{aligned} \mathbb{P}\left(\bar{\chi}^A_{L^2}(\mathcal{L}_i) < K\right) &\leq \mathbb{P}\left(\left(\mu - \bar{\chi}^A_{L^2}(\mathcal{L}_i)\right) > 2|A| - K\right) \\ &\leq \frac{|A| C_L}{(2|A| - K)^2} \\ &\leq \frac{|A| C_L}{|A|^2} \\ &= \frac{C_L}{|A|} \\ &\leq \frac{C_L}{K} \quad \text{for } i = -1, 1, \end{aligned} \tag{B.28}$$

where the second inequality follows from Chebyshev's Inequality.

Let $\epsilon_1 > 0$, with ϵ_1 chosen small enough so that it satisfies the assumption in (2) on

page 1030 of [62] using $1-p=\epsilon_1$ and $q<1-\epsilon$. Since $|A|\geq \dfrac{C_L}{\epsilon}$, we can choose $K\leq |A|$ large enough so that

$$\mathbb{P}(\bar{\chi}^A_{L^2}(\mathcal{L}_i)<K)\leq \epsilon_1/2 \quad \text{for } i=-1,1. \tag{B.29}$$

In other words, by considering $\mathcal{L}_i$ "open" if it contains at least K particles, we have shown $\bar{\chi}_t$ dominates 1-dependent oriented percolation that has open sites with probability $1-\epsilon_1/2$ [122].

Extension to Dual Process.

We have shown via Theorems 3.3.1 and B.3.2 that $\hat{Z}^\beta \Longrightarrow Y$, where $\hat{Z}^\beta$ is the rescaled pruned dual process on $\mathbb{R}^2\times\mathbb{R}_w$, and Y_T is a branching Brownian motion with branching rate μ.

We show previously that

$$\bar{\chi}_0(\mathcal{L}_0)\geq K \implies \mathbb{P}\left(\bar{\chi}^A_{L^2}(\mathcal{L}_i)<K\right)<\epsilon \quad \text{for } i=-1,1.$$

Now we will show that the same holds true for $\bar{\zeta}^\beta$, a modified version of the pruned dual process, in which particles are killed outside of $[-4\rho L,4\rho L]^2\times\mathbb{R}_w$.

Recall that $\hat{\mathbf{D}}$ is the space of Cadlag paths, modified to account for path values of ∞, and

$$\mathcal{D}=\{(x_1,x_2,\ldots):\exists k_0 \ \ s.t. \ \forall k\leq k_0, x_k\in\hat{\mathbf{D}}, \ \ \forall k>k_0, x_k=\infty\}.$$

Let F_m be a function on $\mathcal{D}$ that counts the paths that stay in $[-4\rho L,4\rho L]^2\times\mathbb{R}_w$ and end up in $\mathcal{L}_m$ at time L^2. More precisely,

$$F_m:\mathcal{D}\longrightarrow\mathbb{N}, \qquad F_m((x_1,x_2,\ldots))=\sum_i \mathbb{1}_{G_i\cap\{x_i(L^2)\in\mathcal{L}_m\}}, \tag{B.30}$$

where $G_i=\{x_i(t)\in[-4\rho L,4\rho L]^2\times\mathbb{R}_w\cup\{\infty\} \text{ for } 0\leq t\leq L^2\}$.

Since $\hat{Z}^\beta\Longrightarrow Y$ and F_m is continuous with respect to the limit distribution, the continuous mapping theorem implies

$$F_m(\hat{Z}^\beta)\Longrightarrow F_m(Y) \quad \text{as } \beta\to 0.$$

Therefore, $\exists \beta > 0$ such that

$$\mathbb{P}(\bar{\zeta}^{\beta,A}_{L^2}(\mathcal{L}_i) < K) \leq \mathbb{P}(\bar{\chi}^{A}_{L^2}(\mathcal{L}_i) < K) + \epsilon_1/2, \tag{B.31}$$

Then since we showed in (B.29) that $\mathbb{P}(\bar{\chi}^{A}_{L^2}(\mathcal{L}_i) < K) \leq \epsilon_1/2$, we can conclude that $\mathbb{P}(\bar{\zeta}^{\beta,A}_{L^2}(\mathcal{L}_i) < K) \leq \epsilon_1$. Hence, $\bar{\zeta}_t$ dominates 1-dependent oriented percolation with density $1 - \epsilon_1$.

K particles in $\mathcal{L}_0$.

We assume that the BBM χ starts with one particle at the origin. Since the displacement of Brownian motion is normally distributed, the probability density function for the position of a single particle at time t is

$$f_t(x) = \frac{1}{2\pi\sigma^2 t} \exp\left(-\frac{|x|^2}{2\sigma^2 t}\right), \tag{B.32}$$

where $|x| > w$.

Recall the branching rate for χ is

$$\mu = \begin{cases} \frac{4\pi w}{5} & w = 2 \\ \frac{2\pi w}{3} & w > 2, \end{cases}$$

and that the infinitesimal variance for χ is

$$\sigma^2 = \begin{cases} \frac{2}{5} & w = 2 \\ \frac{1}{3} & w > 2. \end{cases}$$

.

By (B.24), we have

$$\begin{aligned}
\mathbb{E}\left[\bar{\chi}_t(\mathcal{L}_0)\right] &= \exp(\mu t)\mathbb{P}\left(\bar{M}^0_t \in \mathcal{L}_0\right) \\
&= \exp(\mu t)\mathbb{P}\left(M^0_t \in \mathcal{L}_0, M^0_s \in [-4\rho L, 4\rho L]^2 \times \mathbb{R}_w \ \forall s \leq t\right) \\
&\geq \exp(\mu t)\left(1 - \left(\mathbb{P}\left(M^0_t \in \mathcal{L}^C_0\right) + \mathbb{P}\left(\exists s \leq t : M^0_s \notin [-4\rho L, 4\rho L]^2 \times \mathbb{R}_w\right)\right)\right) \\
&\geq \exp(\mu t)\left(1 - \left(2 \cdot \mathbb{P}\left(|M_{t,1}| > \rho L\right) + 2 \cdot \mathbb{P}\left(\exists s \leq t : |M^0_{s,1}| > 4\rho L\right)\right)\right),
\end{aligned}$$

where M^0_t is a Brownian motion with components $M^0_t = (M_{t,1}, M_{t,2}, M_{t,3})$.

Using (B.32) and the definition of ρ in (B.23), for $w > 2$, we have

$$\begin{aligned}
\mathbb{E}\left[\bar{\chi}_t(\mathcal{L}_0)\right] &\geq \exp(\mu t)\left(1 - 2\int_{\rho L}^{\infty} \frac{3}{2\pi t}\exp\left(-\frac{3x^2}{2s}\right)dx - 2\int_0^t \int_{4\rho L}^{\infty} \frac{3}{2\pi s}\exp\left(-\frac{3x^2}{2s}\right)dxds\right) \\
&\geq \exp(\mu t)\left(1 - \frac{3}{\pi t}\int_{\rho L}^{\infty} \frac{x}{\rho L}\exp\left(-\frac{3x^2}{2t}\right)dx - \int_0^t \frac{3}{\pi s}\int_{4\rho L}^{\infty} \frac{x}{4\rho L}\exp\left(-\frac{3x^2}{2s}\right)dxds\right) \\
&= \exp(\mu t)\left(1 - \frac{1}{\pi\rho L}\exp\left(-\frac{3\left(\rho L\right)^2}{2t}\right) - \int_0^t \frac{1}{4\pi\rho L}\exp\left(-\frac{3\left(4\rho L\right)^2}{2s}\right)ds\right) \\
&= \exp(\mu t)\left(1 - \frac{1}{\pi L\rho}\exp\left(-\frac{\mu L^2}{t}\right) - \int_0^t \frac{1}{4\pi L\rho}\exp\left(-\frac{16\mu L^2}{s}\right)ds\right) \\
&\geq \exp(\mu t)\left(1 - \frac{1}{\pi L\rho}\exp\left(-\frac{\mu L^2}{t}\right) - \frac{t}{\pi L\rho}\exp\left(-\frac{16\mu L^2}{t}\right)\right)
\end{aligned}$$

The inequality is analogous for $w = 2$.

Thus, using the same argument shown in (B.28), for any $\epsilon_2 > 0$, we can choose L and K such that, for all w,

$$\mathbb{P}(\bar{\chi}_L(\mathcal{L}_0) \geq K) \geq 1 - \epsilon_2/2.$$

Then by using the function F_m defined by (B.30) and the continuous mapping theorem again, we can conclude that there exists $\beta > 0$ such that

$$\begin{aligned}
\mathbb{P}(\bar{\zeta}_L^{\beta,0}(\mathcal{L}_0) \geq K) &\geq \mathbb{P}(\bar{\chi}_L(\mathcal{L}_0) \geq K) - \epsilon_2/2 \\
&\geq 1 - \epsilon_2.
\end{aligned} \tag{B.33}$$

In [40], they define the following quantity:

$$\begin{aligned}
\sigma_R = \min\left\{t : \xi_t^0 \supset B_{x,R} \text{ for some } x \in \mathbb{Z}^2 \times \mathbb{Z}_w\right\} \\
(= \infty \text{ if no such } x, t \text{ exist}),
\end{aligned}$$

where $B_R = \{(x_1, x_2, x_3) \in \mathbb{Z}^2 \times \mathbb{Z}_w : |x_1|, |x_2| \leq R\}$, and $B_{x,R} = B_R + x$ for $x = (x_1, x_2, 0)$.

Lemma 1 in [40] states that $\mathbb{P}(\sigma_R < \infty | \tau_\emptyset = \infty) = 1$ for all $R > 0$, and the proof holds on $\mathbb{Z}^2 \times \mathbb{Z}_w$, so we have

$$\mathbb{P}(\sigma_{\rho L^2} < \infty | \tau_\emptyset = \infty) = 1. \tag{B.34}$$

Speed using block construction.

Recall $\mathcal{L}_0 = [-\rho L, \rho L]^2 \times \mathbb{R}_w$ and $\mathcal{L}_m = 2m\rho L e_1 + \mathcal{L}_0$ for $m \in \mathbb{Z}$. Due to the scaling in

our construction, let $L = \sqrt{h(\beta)}$. We will use S to denote the time $\sigma_{\rho L^2}$.

Recall that B.31 shows that $\bar{\zeta}_t^{\bar{\beta}}$ dominates a 1-dependent oriented percolation. To make our relation to this percolation precise, for $A \subset \mathcal{L}_0$, and for $m, n \geq 0$, and $m + n$ even, we let

$$\bar{\zeta}_{nL^2}^{\bar{\beta},A}(\mathcal{L}_m) \geq K \implies \theta(m,n) = 1, \tag{B.35}$$

and

$$W_n = \{m : \theta(m,n) = 1\}.$$

Let

$$r_n = \sup\{m : \theta(m,n) = 1\},$$

so $\mathcal{L}_{r_n}$ is the farthest block from the origin containing at least K particles by time nL^2. W_n is known as 1-dependent oriented percolation, and it has open sites with probability $1 - \epsilon_1$.

Result (2) on page 1030 of [62] says that for $q < 1$ and $\lambda > 0$, $\mathbb{P}(\frac{r_n}{n} \leq q) \leq Ce^{-\lambda n}$ on the set where the oriented percolation does not die out. In our case, with $q < 1 - \epsilon$, this result implies that

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{r_n}{n} \geq 1 - \epsilon\right) = 1. \tag{B.36}$$

Recall that

$$a(w,\beta) = \begin{cases} \sqrt{\frac{16\pi w\beta}{25\log(1/\beta)}} & w = 2 \\ \sqrt{\frac{4\pi w\beta}{9\log(1/\beta)}} & w > 2 \end{cases} = \frac{\rho}{\sqrt{h(\beta)}}.$$

For $x_\varepsilon(t;\beta) = \lfloor a(w,\beta)te_1(1-\varepsilon)\rfloor$, $S = \sigma_{\rho L^2}$, and $\tau_\emptyset^A = \min\{t : \xi_t^A = \emptyset\}$, we see that

$$\begin{aligned}
\mathbb{P}(\xi_t^0 \cap \{x_\varepsilon(t;\beta)\} \neq \emptyset | \tau_\emptyset = \infty) &= \mathbb{P}(\xi_t^0 \cap \{x_\varepsilon(t;\beta)\} \neq \emptyset,\ S < \infty | \tau_\emptyset = \infty) \\
&\quad + \mathbb{P}(\xi_t^0 \cap \{x_\varepsilon(t;\beta)\} \neq \emptyset,\ S = \infty | \tau_\emptyset = \infty) \\
&= \mathbb{P}(\xi_t^0 \cap \{x_\varepsilon(t;\beta)\} \neq \emptyset | \tau_\emptyset = \infty,\ S < \infty)\mathbb{P}(S < \infty | \tau_\emptyset = \infty) \\
&\quad + \mathbb{P}(\xi_t^0 \cap \{x_\varepsilon(t;\beta)\} \neq \emptyset | \tau_\emptyset = \infty,\ S = \infty)\mathbb{P}(S = \infty | \tau_\emptyset = \infty)
\end{aligned}$$

Then by (B.34), we have

$$\begin{aligned}\lim_{\beta\to 0}\lim_{t\to\infty}\mathbb{P}(\xi_t^0\cap\{x_\varepsilon(t;\beta)\}\neq\emptyset|\tau_\emptyset=\infty) &= \lim_{\beta\to 0}\lim_{t\to\infty}\mathbb{P}(\xi_t^0\cap\{x_\varepsilon(t;\beta)\}\neq\emptyset|\tau_\emptyset=\infty,\ S<\infty)\\ &\geq \lim_{\beta\to 0}\lim_{t\to\infty}\mathbb{P}\left(\xi_t^{B_{\rho L^2}}\cap\{x_\varepsilon(t+S;\beta)\}\neq\emptyset|\tau_\emptyset^{B_{\rho L^2}}=\infty\right)\\ &= \lim_{\beta\to 0}\lim_{t\to\infty}\frac{\mathbb{P}\left(E_{\rho L^2}(\epsilon,\beta)\right)-\mathbb{P}\left(E_{\rho L^2}(\epsilon,\beta)|\tau_\emptyset^{B_{\rho L^2}}<\infty\right)\mathbb{P}\left(\tau_\emptyset^{B_{\rho L^2}}<\infty\right)}{\mathbb{P}\left(\tau_\emptyset^{B_{\rho L^2}}=\infty\right)},\end{aligned}$$

where $E_{\rho L^2}(\epsilon,\beta)=\left\{\xi_t^{B_{\rho L^2}}\cap\{x_\varepsilon(t+S;\beta)\}\neq\emptyset\right\}$. If we set $L=\sqrt{h(\beta)}$, then by the gambler's ruin formula, if $w>2$,

$$\begin{aligned}\mathbb{P}\left(\tau_\emptyset^{B_{\rho L^2}}<\infty\right) &= (1+\beta)^{-|B_{\rho L^2}|}\\ &= (1+\beta)^{-\frac{16\pi w\log(1/\beta)}{9\beta}}.\end{aligned}$$

Similarly, if $w=2$, $\mathbb{P}\left(\tau_\emptyset^{B_{\rho L^2}}<\infty\right)=(1+\beta)^{-\frac{4^3\pi w\log(1/\beta)}{25\beta}}$.

Note that $\lim_{\beta\to 0}(1+\beta)^{-\frac{16\pi w\log(1/\beta)}{9\beta}}=\lim_{\beta\to 0}(1+\beta)^{-\frac{4^3\pi w\log(1/\beta)}{25\beta}}=0$, so we have

$$\lim_{\beta\to 0}\lim_{t\to\infty}\mathbb{P}\left(\tau_\emptyset^{B_{\rho L^2}}<\infty\right)=\lim_{\beta\to 0}\mathbb{P}\left(\tau_\emptyset^{B_{\rho L^2}}<\infty\right)=0,$$

and

$$\lim_{\beta\to 0}\lim_{t\to\infty}\mathbb{P}\left(\tau_\emptyset^{B_{\rho L^2}}=\infty\right)=1\quad\text{for all } w.$$

This implies that

$$\begin{aligned}
\lim_{\beta\to 0}\lim_{t\to\infty}\mathbb{P}(\xi_t^0(\beta)\cap\{x_\varepsilon(t;\beta)\}\neq\emptyset|\tau_\emptyset=\infty) &\geq \lim_{\beta\to 0}\lim_{t\to\infty}\mathbb{P}\left(\xi_t^{B_{\rho L^2}}\cap\{x_\varepsilon(t+S;\beta)\}\neq\emptyset\right)\\
&= \lim_{\beta\to 0}\lim_{t\to\infty}\mathbb{P}(\tilde{\zeta}_t^{x_\varepsilon(t+S;\beta)}\cap B_{\rho L^2}\neq\emptyset)\\
&\geq \lim_{\beta\to 0}\lim_{t\to\infty}\mathbb{P}(\bar{\zeta}_t^0\cap(B_{\rho L^2}+x_\varepsilon(t+S;\beta))\neq\emptyset)\\
&= \lim_{\beta\to 0}\lim_{n\to\infty}\mathbb{P}(\bar{\zeta}_{nL^2}^{\beta,0}\cap(B_{\rho L}+x_\varepsilon(nL^2+S;\beta))\neq\emptyset)\\
&= \lim_{\beta\to 0}\lim_{n\to\infty}\big(\mathbb{P}(\bar{\zeta}_{nL^2}^{\beta,0}\cap(B_{\rho L}+x_\varepsilon(nL^2+S;\beta))\neq\emptyset,\ \bar{\zeta}_L^{\beta,0}(\mathcal{L}_0)\geq K)\\
&\qquad + \mathbb{P}(\bar{\zeta}_{nL^2}^{\beta,0}\cap(B_{\rho L}+x_\varepsilon(nL^2+S;\beta))\neq\emptyset,\ \bar{\zeta}_L^{\beta,0}(\mathcal{L}_0)<K)\big)\\
&\geq \lim_{\beta\to 0}\lim_{n\to\infty}\mathbb{P}(\bar{\zeta}_{nL^2}^{\beta,0}\cap(B_{\rho L}+x_\varepsilon(nL^2+S;\beta))\neq\emptyset|\bar{\zeta}_L^{\beta,0}(\mathcal{L}_0)\geq K)\mathbb{P}(\bar{\zeta}_L^{\beta,0}(\mathcal{L}_0)\geq K)\\
&\geq \lim_{\beta\to 0}\lim_{n\to\infty}\mathbb{P}(\bar{\zeta}_{nL^2}^{\beta,A}\cap(B_{\rho L}+x_\varepsilon(nL^2+L+S;\beta))\neq\emptyset)(1-\epsilon_2)\\
&\geq \lim_{\beta\to 0}\lim_{n\to\infty}\mathbb{P}\left(\bar{\zeta}_{nL^2}^{\beta,A}(\mathcal{L}_{\bar{m}})\geq K\right)(1-\epsilon_2),
\end{aligned}$$

by (B.33), where $A\subset\mathcal{L}_0$, $K=|A|$, and $\bar{m}=\left\lceil\dfrac{a(w,\beta)(nL^2+L+S)(1-\epsilon)}{2\rho L}\right\rceil$.

Then we use (B.35) and (B.36) to show that for $n>0$,

$$\begin{aligned}
\lim_{\beta\to 0}\lim_{t\to\infty}\mathbb{P}(\xi_t^0(\beta)\cap\{x_\varepsilon(t;\beta)\}\neq\emptyset|\tau_\emptyset=\infty) &\geq \lim_{\beta\to 0}\lim_{n\to\infty}\mathbb{P}(\theta(\bar{m},n)=1)(1-\epsilon_2)\\
&\geq \lim_{\beta\to 0}\lim_{n\to\infty}\mathbb{P}(r_n\geq\bar{m})(1-\epsilon_2)\\
&= \lim_{\beta\to 0}\lim_{n\to\infty}\mathbb{P}\left(r_n\geq\frac{a(w,\beta)(nL^2+L+S)}{\rho L}\right)(1-\epsilon_2)\\
&= \lim_{\beta\to 0}\lim_{n\to\infty}\mathbb{P}\left(r_n\geq\frac{(nL^2+L+S)(1-\epsilon)}{L^2}\right)(1-\epsilon_2)\\
&= \lim_{\beta\to 0}\lim_{n\to\infty}\mathbb{P}\left(\frac{r_n}{n}\geq\left(1+\frac{1}{nL}+\frac{S}{nL^2}\right)(1-\epsilon)\right)(1-\epsilon_2)\\
&= \lim_{\beta\to 0}\lim_{n\to\infty}\mathbb{P}\left(\frac{r_n}{n}\geq 1-\epsilon\right)(1-\epsilon_2)\\
&= 1-\epsilon_2,
\end{aligned}$$

■

B.5 Shape theorem extension to $\mathbb{Z}^2 \times \mathbb{Z}_w$

In this section, we describe the necessary modifications to extend the shape theorem of a biased voter model on $\mathbb{Z}^d$, proven by Bramson and Griffeath in [39, 40], to a biased voter model on $\mathbb{Z}^2 \times \mathbb{Z}_w$.

First, we adjust the definition of a box, B_R, and a ball, D_R, of radius R, which are used repeatedly within the proofs in [39, 40], to hold in the setting $\mathbb{Z}^2 \times \mathbb{Z}_w$. For $R \geq w$, let

$$B_R := \{(x_1, x_2, x_3) \in \mathbb{Z}^2 \times \mathbb{Z}_w : |x_1| \leq R, |x_2| \leq R\},$$

and let

$$D_R := \{(x_1, x_2, x_3) \in \mathbb{Z}^2 \times \mathbb{Z}_w : x_1^2 + x_2^2 \leq R^2\}.$$

Note that a box translated by x is defined as $B_{x,R} = B_R + x$ for $x \in \mathbb{Z}^2$. We mention later the reason for which D_R is defined as w stacked two-dimensional balls, rather than as a cut-out of a three-dimensional ball.

The primary modification occurs in the proof of Proposition 2, which is used within the proof of Proposition 1 in [40]. They define a Markov chain (X_t^x), embedded in the dual process $(\tilde{\zeta}_t^x)$, which has uniformly positive drift toward the origin. We modify the jump rates for (X_t^x) on $\mathbb{Z}^2 \times \mathbb{Z}_w$ for any $y = (y_1, y_2, y_3), z = (z_1, z_2, z_3) \in \mathbb{Z}^2 \times \mathbb{Z}_w$, as indicated below.

First let $w > 2$. If $y_3 < \frac{w}{2}$:

$$y \to z \text{ at rate } \begin{cases} (1+\beta)/6 \text{ if } ||z-y|| = 1 & \text{and } \textit{either } ||z|| < ||y|| \\ & \textit{or } y_i = 0, z_i = 1 \text{ for } 1 \leq i \leq 2 \\ 1/6 \text{ if } ||z-y|| = 1 & \text{and the other above condition does not hold.} \end{cases}$$

If $y_3 \geq \frac{w}{2}$:

$$y \to z \text{ at rate } \begin{cases} (1+\beta)/6 \text{ if } ||z-y|| = 1 & \text{and } \textit{either } z_3 = y_3 + 1, \\ & \textit{or } z_3 = y_3, \ ||z|| < ||y||, \\ & \textit{or } y_i = 0, z_i = 1 \text{ for } 1 \leq i \leq 2 \\ 1/6 \text{ if } ||z-y|| = 1 & \text{and the other above condition does not hold.} \end{cases}$$

Next let $w = 2$. Then the jump rates are defined as follows.

$$y \to z \text{ at rate } \begin{cases} (1+\beta)/5 \quad \text{if } ||z-y|| = 1 & \text{and } \textit{either } z_3 = 1, y_3 = 0, \\ & \textit{or } z_3 = y_3,\ ||z|| < ||y||, \\ & \textit{or } y_i = 0, z_i = 1 \text{ for } 1 \le i \le 2 \\ 1/5 \quad \text{if } ||z-y|| = 1 & \text{and the other above condition does not hold.} \end{cases}$$

For all w, no transitions take place, outside of those listed above. With these modifications, (X_t^x) has uniformly positive drift toward the origin off of some ball $D_\gamma \subset \mathbb{Z}^2 \times \mathbb{Z}_w$, where γ is a positive constant, as is the case in $\mathbb{Z}^2$. In this case, for large λ, the drift of $||X_t^x||$ toward 0 is minimized over states in $\mathbb{Z}^2 \times \mathbb{Z}_w - D_\gamma$ at sites located on the axes of $\mathbb{Z}^2$ (i.e. sites (y_1, y_2, y_3) in which $y_1 = 0$ or $y_2 = 0$), and this minimal drift is asymptotically $\frac{\beta}{4}$ as $\lambda \to \infty$, in contrast to $\frac{\beta}{2d}$ in the $\mathbb{Z}^d$ case.

For $w > 2$, exactly half of the nearest neighbors z for any site y satisfy the condition indicating a transition rate of $(1+\beta)/6$, so as was the case in [40], the rate at which (X_t^x) leaves any site is $(2+\beta)/2$. Hence, for states in $\mathbb{Z}^2 \times \mathbb{Z}_w - D_\gamma$, the minimal expected displacement toward 0 resulting from each jump, is asymptotically

$$\frac{\beta}{4} \cdot \frac{2}{2+\beta} = \frac{\beta}{2(2+\beta)},$$

in contrast to $\dfrac{\beta}{d(2+\beta)}$ in the $\mathbb{Z}^d$ case.

For $w = 2$, the rate at which (X_t^x) leaves a site on sheet 0 is $\left(\frac{2}{5}(1+\beta) + \frac{3}{5}\right)$, and the rate at which (X_t^x) leaves a site on sheet 1 is $\left(\frac{3}{5}(1+\beta) + \frac{2}{5}\right)$. Therefore, the minimal expected displacement toward 0, for states in $\mathbb{Z}^2 \times \mathbb{Z}_w - D_\lambda$, is asymptotically

$$\frac{\beta}{4} \Big/ \left(\frac{1}{2} \cdot \frac{3(1+\beta)+2}{5} + \frac{1}{2} \cdot \frac{2(1+\beta)+3}{5} \right) = \frac{\beta}{2(2+\beta)}.$$

Thus, for all w, as is done in [40], we define a family of continuous time processes on $[-\gamma, \infty) \subset \mathbb{R}^1$,

$$Z_t^{(x,\alpha)} = ||X^x_{2t/(2+\beta)}|| - \gamma,$$

where $\alpha = ||x|| - \gamma$. In our case, for large γ,$(Z_t^{(x,\alpha)})$ has minimal drift toward 0 of $\mu + \epsilon$,

where $\mu > 0$, $\epsilon > 0$ are chosen so that

$$0 < \mu < \mu + \epsilon < \frac{\beta}{2(2+\beta)}.$$

Note that we have $\frac{1}{2}$ in place of $\frac{1}{d}$ in the μ estimate from the $\mathbb{Z}^d$ case. The constant μ is used within the proof of Proposition 1 in [40], in which they estimate the probability that the BVM, initially covering box B_R, eventually contains a box growing approximately linearly in t. The μ is used to define a sequence of time intervals s_k, and the proof considers the probability that $\xi_t^{B_R}$ covers $B_{R(k)}$ for $t \in [t_k, t_{k+1}]$ (where $t_k = \sum_{j=1}^{k} s_j$).

The family of processes $Z_t^{(x,\alpha)}$ is defined so that it can be used to obtain the following inequality for $x \in B_{R'}$,

$$\mathbb{P}(x \notin \xi_t^{B_R}) \leq \max_{\alpha \leq \sqrt{d}R'} \mathbb{P}(Z_t^{(x,\alpha)} \geq R - \gamma), \tag{B.37}$$

where $d = 2$ in our case, consistent with the use of 2 for d in the estimate for μ. The use of $d = 2$ is required so that it can be shown that the right-hand side of the inequality above, with $R' = R(k)$, is on the order of $\exp(-\gamma R(k))$. This is used to prove Proposition 2 in [40]. Note that in Lemma 2 and Proposition 2, the $(2R(k))^d$ should be $w(2R(k))^2$, due to the shape of a box in $\mathbb{Z}^2 \times \mathbb{Z}_w$, dictating the need for $d = 2$ within the proofs of Lemma 2 and Proposition 2. Thus, it was imperative that D_R be defined as a set of stacked discs, so that with $d = 2$, the following inclusions hold:

$$D_{R'} \subset B_{R'} \subset D_{\sqrt{d}R'},$$

giving rise to (B.37).

In part II of the shape-theorem [39], Bramson and Griffeath show that satisfying a given set of regularity assumptions is sufficient to apply Richardson's proof technique for the Williams-Bjerknes model in [123]. Thus, they are able to conclude that the rate of expansion of the biased voter model is indeed linear. Durrett and Griffeath generalize this proof to other contact processes in [124].

The BVM on $\mathbb{Z}^2 \times \mathbb{Z}_w$ satisfies the necessary properties, outlined in [124], to be considered a growth model. The proof that these properties are sufficient to guarantee linear asymptotic growth holds in this case, using our definition for D_R. This can be verified using the following observation. If we let $\hat{D}_R^d$ denote an R-ball in $\mathbb{Z}^d$, then $|D_R| = w|\hat{D}_R^2|$, and $|\delta D_R| = w|\delta \hat{D}_R^2|$, where δD_R is used to denote the boundary of D_R. Thus, such quantities, particularly

relevant within the proof of Proposition 1 in [124], can be expressed in terms of $d = 2$, consistent with the relationship between our defined D_R and B_R in $\mathbb{Z}^2 \times \mathbb{Z}_w$.

Appendix C

Chapter 4 Appendix

Figure C.1 depicts the set of all possible birth events that occur throughout the evolution of the tumor, as described in Section 4.2. Each row in the figure corresponds to division of each type of cell, and the rate with which each birth event occurs is provided below the image.

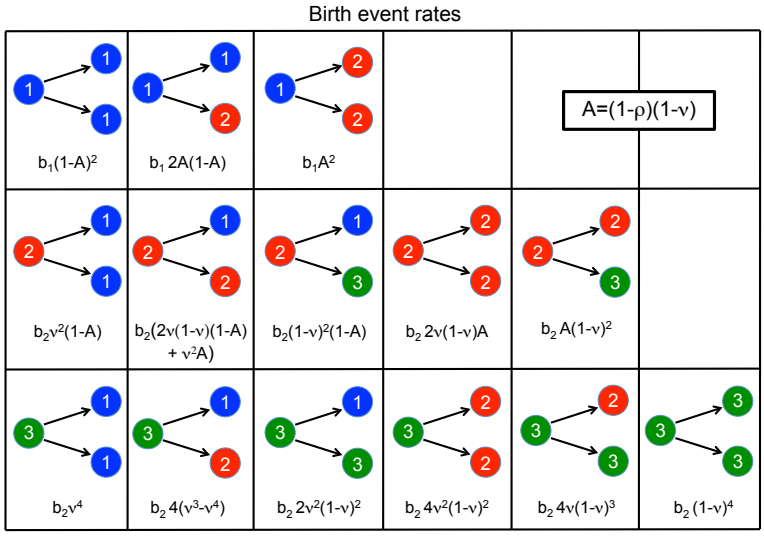


Figure C.1: **The set of possible birth events when a type-1, type-2, or type-3 cell divides.** Each birth event is displayed with the corresponding rate at which it occurs. Note that we let $A = (1-\rho)(1-\nu)$, where ρ, ν denote the probability of maintenance methylation and *de novo* methylation, respectively, at each CpG site during cell replication.

C.1 Parameterization

Detection thresholds and surgical removal proportion

The clinical data described in section 4.3 is used determine the tumor size detection thresholds D_1 and D_2, at diagnosis and recurrence, respectively. To convert this to approximate cell population numbers for calibrating D_1 in the model, we use a standard approximation for the number of cells in a spherical tumor of radius 0.5 cm (volume $\frac{\pi}{6}$ cm^3) is 10^9 cells [125]. After removing outliers, the average tumor radius at the time of detection was $r_1 = 0.142$ cm (volume 0.012 cm^3 assuming a tumor spheroid). Thus, applying the conversion factor above we obtain $D_1 = 2.298 \times 10^7$ cells. Similarly, the average tumor radius at the time of recurrence is $r_2 = 0.1286$ cm, resulting in the estimate $D_2 = 1.702 \times 10^7$ cells. The observed effect $D_2 < D_1$ may reflect the fact that patients previously diagnosed and treated for GBM are under closer scrutiny for recurrence than for the initial diagnosis.

After detection, surgery is performed on the tumor. The patient data describing the tumor radius size after surgery is displayed in a pie chart in Figure 4.6c. Since a large number of patients were reported with 0 mm of tumor remaining after surgery, and there are a few large outliers, we felt the median would provide a better representation of the data than the average. Thus we found a median of 0.1 mm of tumor remaining after surgery and then calculated the number of tumor cells, using the same techniques shown when calculating D_1, D_2. After dividing this number by D_1, we determined that the proportion of cells remaining after surgery is $p_s = 2.237 \times 10^{-7}$.

C.1.1 Methylation parameters

Literature estimates of DNA methylation and demethylation rates are conflicting, and they can vary over several orders of magnitude. For example, in [86], an estimate of 2×10^{-5} per CpG site per cell division is used for both the methylation and demethylation rates in their model. In [126], the authors use demethylation and methylation rates for each CpG site on the order of 10^{-4} and 10^{-3}, respectively, and in [127] the authors estimate demethylation and methylation rates on the order of 10^{-3} and 10^{-2}, respectively. These estimates are derived from observed methylation patterns obtained through genomic sequencing. Due to the discrepancy in these reported estimates of methylation rates, we chose instead to model the more specific mechanisms of *de novo* and maintenance methylation and then use these processes to calculate the rates of conversion between cell types.

In [1] the authors used a steady state solution for their Markov chain model in conjunction with observed methylation data from the promoter region in human lymphocytes to estimate the *de novo* and maintenance methylation probabilities, ν and ρ, respectively [128]. In our model, we use these estimates of

$$\begin{aligned} \nu &= 0.09 \\ \rho &= 0.95, \end{aligned} \tag{C.1}$$

which will serve as the baseline methylation parameter values.

C.1.2 Intrinsic birth and death rates

Here we describe the parametrization of the birth and death rates in the pretreatment phase and recovery periods, when no treatment is administered.

We use the PDX cell line experiments described in Section 4.3 to determine the cell viability of MGMT^+ and MGMT^- cells when exposed to TMZ, and also to calibrate untreated birth and death rates for both types of cells. In particular, live cell counts from three DMSO groups are used to determine the net growth rate for each group as shown below. Since the cell population grows exponentially, for each live cell count $\tilde{L}$, the corresponding net growth rate $\tilde{\lambda}$ is determined by $\tilde{\lambda} = \ln(\tilde{L}/48000)/8 \quad \text{day}^{-1}$.

The *in vitro* net growth rates and dead cell counts are combined to determine a death rate for each DMSO group, using the following relationship between the death rate $\tilde{c}$ and the dead cell count $\tilde{D}$ after 8 days,

$$\tilde{D} = \int_0^8 \tilde{c} * 48000 e^{s\tilde{\lambda}} ds \quad \text{cells.}$$

In the equation above, $48000e^{s\tilde{\lambda}}$ is the expected number of total cells in one DMSO group at time s. We assume that the death rate is equal for all cell types in the absence of drug, so we multiply the expected number of cells at time s by the death rate, and then integrate that product as s varies between 0 and 8 days, to obtain the number of dead cells after 8 days have passed.

In vitro birth and death rates differ from the those *in vivo*, but we assume that there is scaling factor that we can multiply *in vitro* birth and death rates by to determine the corresponding *in vivo* rates. Given this assumption, the ratio of death rate to net growth rate *in vitro* should be equal to the ratio *in vivo*. Thus, after averaging the ratios for each of the three groups, we obtained a mean ratio of 0.0356. We assume that the death rates

for all cell types are equal in the absence of TMZ and that the birth rates differ. Using the in vivo net growth rate $\lambda = 0.0897$ day^{-1}, obtained from the patient data depicted in Figure 4.6b, we determined that the death rates in the absence of therapy are

$$c_1 = c_2 = \lambda * 0.0356 = 0.0032 \quad \text{day}^{-1}. \tag{C.2}$$

Then it remains to determine the birth rates, b_1 and b_2, for type-1 and type-2/3 cells, respectively. We use the multitype branching process mean to calculate these initial birth rates. In Section C.3, we show that

$$\mathbb{E}\left(\begin{bmatrix} X_1(t) & X_2(t) & X_3(t) \end{bmatrix} \mid \begin{bmatrix} X_1(0) & X_2(0) & X_3(0) \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \end{bmatrix}\right) = \begin{bmatrix} 1 & 0 & 0 \end{bmatrix} \mathbf{M}(t).$$

where, in the absence of therapy,

$$\mathbf{M}(t) = \exp(\mathbf{G}t),$$

and $\mathbf{G}$ is the infinitesimal generator of the multitype branching process, defined in Section C.3. All of the entries in $\mathbf{G}$ are expressions of ν, ρ, d_1, d_2, b_1, and b_2.

We use the overall net growth rate rate $\lambda = 0.0897$/cell/day and the quantity D_1 to determine the following estimation of the time Δt_1 between the emergence of the first cancer cell and tumor detection.

$$\Delta t_1 = \frac{\ln(2.298 \times 10^7)}{0.0897} = 188.96 \text{ days}. \tag{C.3}$$

Using Δt_1, the methylation probabilities ν, ρ in (C.1), and the death rates in (C.2), we can solve for b_1 and b_2 given the following constraints,

$$\begin{aligned} &m_{1,1}(\Delta t_1) + m_{1,2}(\Delta t_1) + m_{1,3}(\Delta t_1) = D_1 \\ &(m_{1,2}(\Delta t_1) + m_{1,3}(\Delta t_1)) = 0.2387\,(m_{1,1}(\Delta t_1) + m_{1,2}(\Delta t_1) + m_{1,3}(\Delta t_1))\,, \end{aligned}$$

where $m_{i,j}(\Delta t_1)$ is the entry in the i-th row and j-th column of $\mathbf{M}(t)$ when $t = \Delta t_1$. The first constraint follows from the fact that D_1 is the expected tumor size at the time of detection, Δt_1. The second constraint follows from our experimental measurements of the frequency of cells expressing MGMT within each of the groups that were not exposed to TMZ. We averaged the frequency of MGMT$^+$ cells in the three DMSO groups to determine that the proportion of hemimethylated and unmethylated cells in the absence of TMZ

should be 0.2387, implying the second constraint. We solved the system of equations above to determine the initial birth rates,

$$\begin{aligned} b_1 &= 0.0927 \quad \text{day}^{-1} \\ b_2 &= 0.0938 \quad \text{day}^{-1}. \end{aligned} \tag{C.4}$$

C.1.3 Birth and death rates during TMZ treatment

Recall that α, β are the radiosensitivity parameters used in linear-quadratic model. The parameter ratio α/β indicates the fractionation sensitivity of the tumor cells, and a standard estimate for α/β in most tumors is 10 [129, 130]. Thus we use typical estimates of $\alpha = 0.1$ Gy^{-1} and $\beta = 0.01$ Gy^{-2} as baseline values in our model.

In order to determine the contribution of chemotherapy to type-1 and type-2/3 death rates, we first describe the concentration of TMZ as a function of time. We model the plasma concentration of TMZ using an exponential decay function $C(t) = C_0 e^{-kt}$, where t is the time since the last TMZ dose. The parameter $k = \dfrac{\ln 2}{t_{1/2}}$, where $t_{1/2}$ is the half-life of TMZ. We used linear regression of pharmacokinetic data to approximate the parameter $C_0(Z)$ in μM, as shown in Figure C.2, which is the maximum plasma concentration of TMZ in μM, as a function of the administered dose Z of TMZ in mg/m^2 of body-surface area. We obtained the following linear relationship:

$$C_0(Z) = 0.28Z \ \ \mu\text{M}.$$

We also used this pharmacokinetic data to approximate $t_{1/2} = 0.074$ days [131, 132, 133, 134, 135].

Due to the toxicity of TMZ, the cell death rates fluctuate during treatment, but the birth rates remain fixed. In order to determine the death rates as a function of the plasma concentration of TMZ, we need the cell viability functions for cells expressing MGMT and cells with no MGMT expression.

We determined the cell viability functions using data obtained from the MTT assays, described in Section subsec:data. in which one group of GBM6 cells was not exposed to TMZ, and seven other groups of GBM6 cells were exposed to varying concentrations of TMZ. After 8 days the number of MGMT^- (type-1) cells and number of MGMT^+ (type-2+type-3) cells were counted for each group. We used the MGMT^- and MGMT^+ cell counts obtained in the experiments to calculate the cell viability for methylated and hemi/unmethylated cells when exposed to each of the TMZ concentrations tested in the assay.

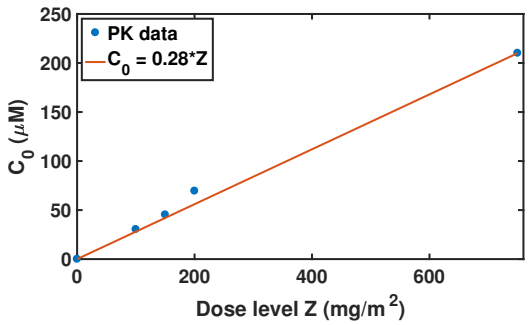


Figure C.2: **Maximum plasma concentration C_0 of TMZ, in μM, as a function of admistered dose.** The plot depicts C_0 for various administered doses Z of TMZ in mg/m^2 of body-surface area, obtained from pharmacokinetic data, and the linear fit of these data points.

Using these data points for each concentration of TMZ, we fit the cell viability functions for type-1 and type-2 cells to two Hill equations. The Hill equations are of the form $\frac{1}{1+\left(\frac{C}{E}\right)^H}$, where E is the EC_{50} value and H is known as the Hill coefficient. We denote these cell viability functions $v_1(C)$ and $v_2(C)$ for type-1 and type-2/3 cells, respectively, at concentration C. The best fit equations, displayed in Figure C.3, are

$$v_1(C) = \frac{1}{1+\left(\frac{C}{20.301}\right)^{1.476}},$$

$$v_2(C) = \frac{1}{1+\left(\frac{C}{57.305}\right)^{2.096}}.$$

In the cell viability experiment, the number of type-1 cells exposed to concentration C after 8 days is

$$v_1(C)(P_0 e^{8\lambda_1}),$$

where P_0 is the initial number of cells tested in the experiment, and $\lambda_1 = b_1 - c_1$ is the type-1 net growth rate in the absence of treatment. Note that $P_0 e^{8\lambda_1}$ is the expected number of type-1 cells after 8 days, in the absence of treatment. Thus, the net growth rate for type-1 cells exposed to concentration C is

$$\frac{\ln(v_1(C)e^{8\lambda_1})}{8} \quad \text{day}^{-1},$$

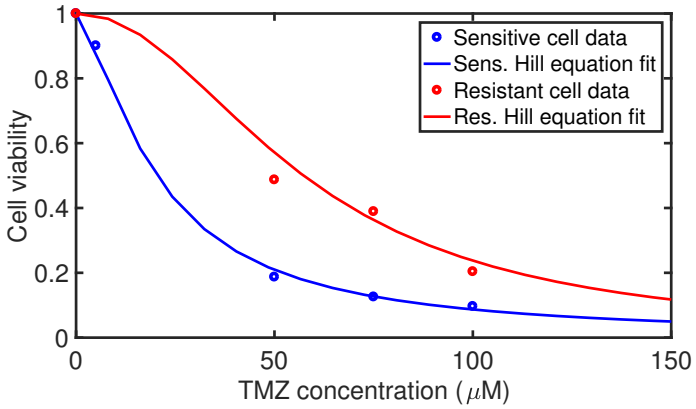


Figure C.3: **Cell-viability functions.** The plot depicts experimental cell viability data points for type-1 and type-2/3 cells exposed to TMZ, and the cell viability curves to which we fit the data.

which implies that the type-1 death rate component due to TMZ is

$$g_1(C(t)) = \frac{1}{8}\ln\left(1+\left(\frac{C(t)}{20.301}\right)^{1.476}\right) \quad \text{day}^{-1}, \tag{C.5}$$

where t is the time since the most recent dose of TMZ. Analogously the type-2/3 death rate component due to TMZ is

$$g_2(C(t)) = \frac{1}{8}\ln\left(1+\left(\frac{C(t)}{57.305}\right)^{2.096}\right) \quad \text{day}^{-1}. \tag{C.6}$$

C.2 Robustness to variation in parameters

In Section 4.4, we claim that a reduction in methylation percentage between tumor diagnosis and recurrence may result from TMZ's inhibition of maintenance methylation. We varied each of the model parameters independently to test the robustness of the claim to variability in the parameters.

As each parameter varies, we investigate the change in methylation percentage between tumor diagnosis and recurrence in the case in which TMZ does not impact methylation ($\rho_z = 0.95$) and in the case in which TMZ causes a reduction in maintenance methylation ($\rho_z = 0.5$). Figure C.4 shows this comparison as b_1, b_2, d_1, and d_2 vary, and Figure C.5 shows the comparison as ρ and ν vary. Note that as ρ varies, we let $\rho_z = \rho$. We observe that there is minimal change in methylation between detection and recurrence when $\rho_z = 0.95$,

except in the case with no *de novo* methylation ($\nu = 0$). However it is highly unlikely that the DNMT3a/b methyltransferases will be completely inactive throughout the entire evolution of the tumor. Thus our claim that selection alone does not explain the observed methylation shift is robust to variation of the model parameters within a reasonable range.

Additionally in the case in which TMZ inhibits maintenance methylation ($\rho_z = 0.5$), we observe a significant decrease in methylation percentage between tumor diagnosis and recurrence as each parameter varies, except when ρ is near 0.5. This is expected, since $\rho_z = 0.5$, so TMZ does not significantly impact the maintenance methylation when ρ is near 0.5 in the absence of drug. Hence, whenever TMZ has a sizable impact on the maintenance methylation rate, we observe a downward shift in methylation between detection and recurrence, consistent with clinical studies. Thus, our conjecture that the clinically observed methylation shift results from TMZ's impact on maintenance methylation is robust to parameter variability.

C.3 Population Means

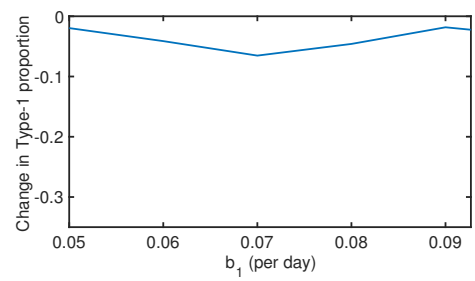
We can calculate the expected number of type-1 and type-2 cells at time t in terms of the birth and death rates. We will use this calculation and experimental data to determine the values for b_1 and b_2 in our model. For ease of notation, we use the following notation for the components of each offspring distribution p_1, p_2, p_3.

$$\begin{array}{lll} \mu_1^{1,1} = p_1((2,0,0)) & \mu_2^{1,1} = p_2((2,0,0)) & \mu_3^{1,1} = p_3((2,0,0)) \\ \mu_1^{1,2} = p_1((1,1,0)) & \mu_2^{1,2} = p_2((1,1,0)) & \mu_3^{1,2} = p_3((1,1,0)) \\ \mu_1^{2,2} = p_1((0,2,0)) & \mu_2^{1,3} = p_2((1,0,1)) & \mu_3^{1,3} = p_3((1,0,1)) \\ & \mu_2^{2,2} = p_2((0,2,0)) & \mu_3^{2,2} = p_3((0,2,0)) \\ & \mu_2^{2,3} = p_2((0,1,1)) & \mu_3^{2,3} = p_3((0,1,1)) \\ & & \mu_3^{3,3} = p_3((0,0,2)) \end{array}$$

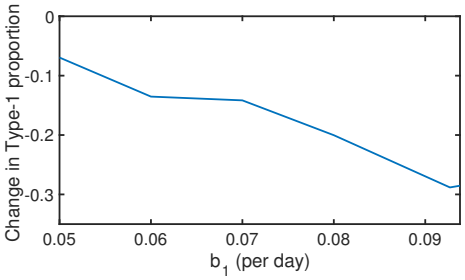
The offspring distributions p_1, p_2, p_3 are defined in terms of ρ, ν in (4.1), (4.2), and (4.3), respectively. For each component listed above, we use the following notation to denote the rate at which a type-i cell gives rise to a type-j and type-k cell:

$$u_i^{j,k}(t) := b_i \mu_i^{j,k}(t),$$

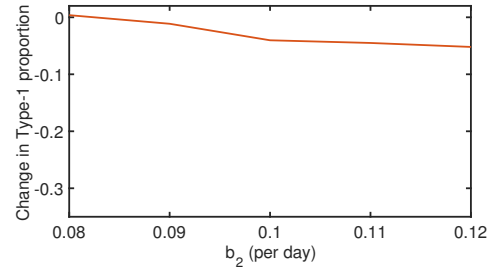
where $1 \leq i, j, k \leq 3$. Note that in our case, $b_3 = b_2$. Additionally, if $\mu_i^{j,k}$ is not listed above, then we let $\mu_i^{j,k} = 0$. We let T_i denote the sum of the event rates for cell type-i, as defined



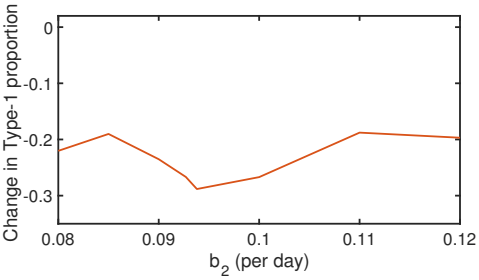
(a) Vary b_1.



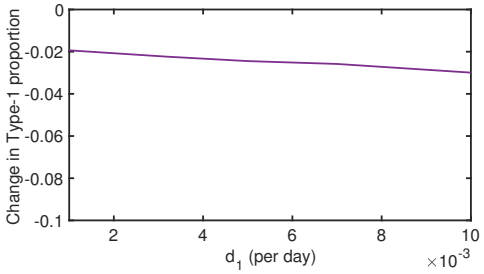
(b) Vary b_1 ($\rho_z = 0.5$).



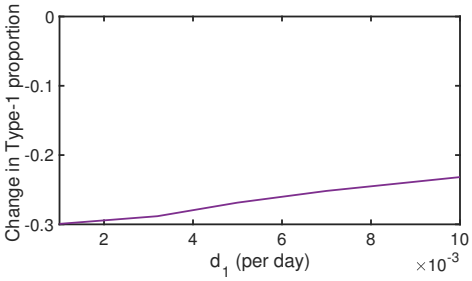
(c) Vary b_2.



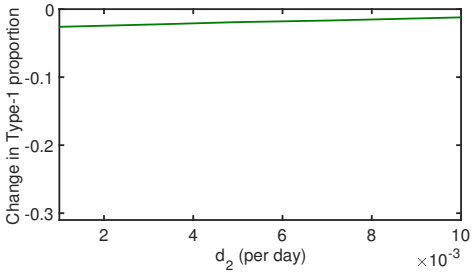
(d) Vary b_2 ($\rho_z = 0.5$).



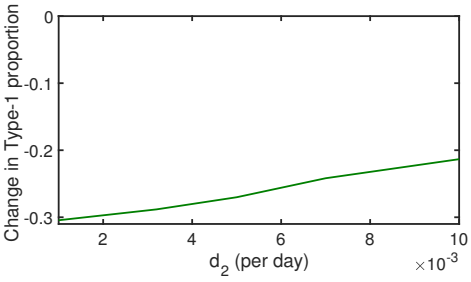
(e) Vary d_1.



(f) Vary d_1 ($\rho_z = 0.5$).

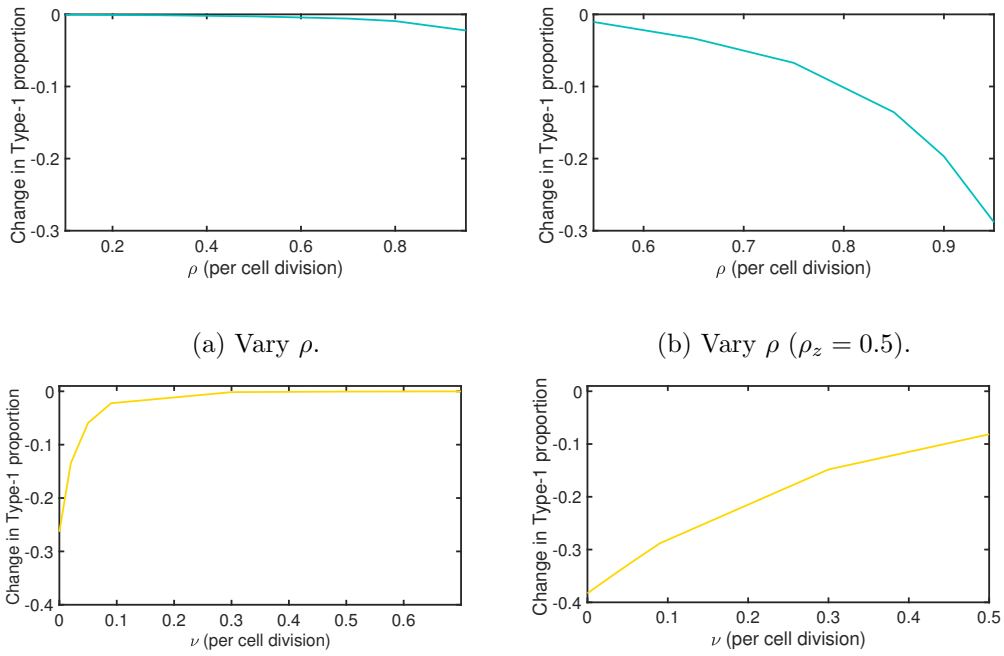


(g) Vary d_2.



(h) Vary d_2 ($\rho_z = 0.5$).

Figure C.4: **Change in methylation percentage as birth and death rates vary.** The plots show the change in proportion of type-1 cells between tumor detection and recurrence. In the baseline case, in which TMZ does not impact maintenance methylation ($\rho_z = 0.95$), the change in proportion is shown as (a) b_1, (c) b_2, (e) d_1, and (g) d_2 vary. In the case in which $\rho_z = 0.5$, we plot the change in proportion as (b) b_1, (d) b_2, (f) d_1, and (h) d_2 vary.



(a) Vary ρ.

(b) Vary ρ ($\rho_z = 0.5$).

(c) Vary ν.

(d) Vary ν ($\rho_z = 0.5$).

Figure C.5: **Change in methylation percentage as methylation probabilities vary.** The plots depict the change in proportion of type-1 cells between tumor detection and recurrence. In the baseline case, in which TMZ does not impact maintenance methylation ($\rho_z = 0.95$), the change in proportion is shown as (a) ρ and (c) ν vary. In the case in which $\rho_z = 0.5$, we plot the change in proportion as (b) ρ and (d) ν vary.

below,

$$T_i(t) := d_i(t) + \sum_{1\le j,k\le 3} u_i^{j,k}(t), \quad 1 \le i \le 3.$$

In our case, $d_3 = d_2$. We write the rates as functions time, since ν, ρ, d_1, d_2 may vary in the absence and presence of therapy. Let $\mathbf{s} = (s_1, s_2, s_3)$, where $0 \le s_1, s_2, s_3 \le 1$. The following are the probability generating functions for type-1, type-2, and type-3 cells, respectively:

- $F_1(\mathbf{s}, t) = \mathbb{E}(\mathbf{s}^{X_1(t)}|X_1(0) = 1) = \sum_{j=0}^{\infty} P(X_1(t) = j|X_1(0) = 1)\mathbf{s}^j$,
- $F_2(\mathbf{s}, t) = \mathbb{E}(\mathbf{s}^{X_2(t)}|X_2(0) = 0) = \sum_{j=0}^{\infty} P(X_2(t) = j|X_2(0) = 0)\mathbf{s}^j$,
- $F_3(\mathbf{s}, t) = \mathbb{E}(\mathbf{s}^{X_3(t)}|X_3(0) = 0) = \sum_{j=0}^{\infty} P(X_3(t) = j|X_3(0) = 0)\mathbf{s}^j$.

The following are the probability generating functions of the offspring distributions:

$$\begin{aligned}
f_1(\mathbf{s}, t) &= (T_1(t))^{-1}(d_1(t) + u_1^{1,1}(t)s_1^2 + u_1^{1,2}(t)s_1s_2 + u_1^{2,2}(t)s_2^2),\\
f_2(\mathbf{s}, t) &= (T_2(t))^{-1}(d_2(t) + u_2^{1,1}(t)s_1^2 + u_2^{1,2}(t)s_1s_2 + u_2^{1,3}(t)s_1s_3 + u_2^{2,2}(t)s_2^2) + u_2^{2,3}(t)s_2s_3),\\
f_3(\mathbf{s}, t) &= (T_3(t))^{-1}\cdot\\
&\qquad (d_3(t) + u_3^{1,1}(t)s_1^2 + u_3^{1,2}(t)s_1s_2 + u_3^{1,3}(t)s_1s_3 + u_3^{2,2}(t)s_2^2) + u_3^{2,3}(t)s_2s_3 + u_3^{3,3}(t)s_3^2).
\end{aligned}$$

The associated infinitesimal generating functions are

$$\begin{aligned}
g_1(\mathbf{s}, t) &= T_1(t)(f_1(\mathbf{s}, t) - s_1),\\
g_2(\mathbf{s}, t) &= T_2(t)(f_2(\mathbf{s}, t) - s_2),\\
g_3(\mathbf{s}, t) &= T_3(t)(f_3(\mathbf{s}, t) - s_3).
\end{aligned}$$

Let $\mathbf{M}(t) = \{m_{ij}(t); 1 \le i, j \le 3\}$ be the mean matrix of this two-type branching process.

Using the backward Kolmogorov equation, we obtain the infinitesimal generator for the branching process:

$$\mathbf{G}_t = \left[T_i(t)\left(\frac{\partial f_i(\mathbf{s}, t)}{\partial s_j}\bigg|_{\mathbf{s}=(1,1,1)} - \delta_{ij}\right)\right]_{1\le i,j\le 3} =$$

$$\begin{bmatrix} 2u_1^{1,1}(t) + u_1^{1,2}(t) - T_1(t) & u_1^{1,2}(t) + 2u_1^{2,2}(t) & 0 \\ 2u_2^{1,1}(t) + u_2^{1,2}(t) + u_2^{1,3}(t) & u_2^{1,2}(t) + 2u_2^{2,2}(t) + u_2^{2,3}(t) - T_2(t) & u_2^{1,3}(t) + u_2^{2,3}(t) \\ 2u_3^{1,1}(t) + u_3^{1,2}(t) + u_3^{1,3}(t) & u_3^{1,2}(t) + 2u_3^{2,2}(t) + u_3^{2,3}(t) & u_3^{1,3}(t) + u_3^{2,3}(t) + 2u_2^{3,3}(t) - T_3(t) \end{bmatrix}$$

Then $\mathbf{M}(t)$ must satisfy the equation:

$$\frac{d\mathbf{M}(t)}{dt} = \mathbf{G}_t \mathbf{M}(t).$$

Therefore we have

$$\mathbf{M}(t) = \exp\left(\int_0^t \mathbf{G}_s \, ds\right),$$

and

$$\mathbb{E}\left(\begin{bmatrix} X_1(t) & X_2(t) & X_3(t) \end{bmatrix} \mid \begin{bmatrix} X_1(0) & X_2(0) & X_3(0) \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \end{bmatrix}\right) = \begin{bmatrix} 1 & 0 & 0 \end{bmatrix} \mathbf{M}(t).$$

Table C.1: **Baseline parameters.** These are calibrated using experimental and clinical data in Section C.1.

Parameter	**Description**	**Value**
D_1	Initial tumor detection size	$2.298 * 10^7$ cells
D_2	Recurrent tumor detection size	$1.702 * 10^7$ cells
p_s	Proportion of cells remaining after surgery	$2.237 * 10^{-7}$
Δt_1	Time between first cancer cell and initial detection	188.96 days
λ	Overall net growth, in the absence of treatment	0.0897 day^{-1}
ν	Probability of de novo methylation of a previously unmethylated site	0.09 /cell division
ρ	Probability of maintenance methylation of a previously methylated site	0.95 /cell division
b_1, b_2	Type-1, type-2 birth rates	$b_1 = 0.0927, b_2 = 0.0938$ day^{-1}
c_1, c_2	Type-1, type-2 death rates, in the absence of treatment	$c_1 = c_2 = 0.0032$ day^{-1}
α, β	Radiosensitivity parameters	$\alpha = 0.1$ Gy^{-1}, $\beta = 0.01$ Gy^{-2}
$C(t)$	Concentration of TMZ, at time t after administered dose Z mg/m^2	$C(t) = 0.28Z \exp\left(-t\frac{\ln 2}{0.074}\right)$ μM
$v_1(C)$	Type-1 cell viability (as a function of TMZ concentration)	$\left(1 + \left(\frac{C(t)}{20.301}\right)^{1.476}\right)^{-1}$
$v_2(C)$	Type-2 cell viability (as a function of TMZ concentration)	$\left(1 + \left(\frac{C(t)}{57.305}\right)^{2.096}\right)^{-1}$
$g_1(C(t))$	Type-1 death rate component due to TMZ	$\frac{1}{8}\ln\left(1 + \left(\frac{C(t)}{20.301}\right)^{1.476}\right)$ day^{-1}
$g_2(C(t))$	Type-2 death rate component due to TMZ	$\frac{1}{8}\ln\left(1 + \left(\frac{C(t)}{57.305}\right)^{2.096}\right)$ day^{-1}
$h(D(t))$	Death rate component due to radiation (as a function of the total Gray delivered in $[0, t]$)	$(\alpha + 2\beta D(t))\frac{d}{dt}D(t)$ day^{-1}
p_1, p_2, p_3	Type-1, type-2, type-3 offspring distributions	See (4.1), (4.2), (4.3)