Stochastic Models of Evolutionary Dynamics

Stochastic Models of Evolutionary Dynamics

a dissertation presented
by
Ben Michael Edwin Adlam
to
The School of Engineering and Applied Sciences

in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy
in the subject of
Applied Mathematics

Harvard University
Cambridge, Massachusetts
May 2018

Dissertation advisor: Professor Martin Nowak Ben Michael Edwin Adlam

Stochastic Models of Evolutionary Dynamics

Abstract

Stochasticity is a fundamental component of evolution. Many essential evolutionary phenomena cannot be modeled without it. In developing and analyzing stochastic processes that model the dynamics of evolution, this dissertation applies tools from probability theory to study fundamental mathematical principles of evolution. These principles determine how the timeline of macroscopic evolution is constructed by the accumulation of many microscopic changes.

At the microscopic scale, we focus on populations of reproducing individuals. Even under neutral evolution, the complex interaction between mutation and genealogy produces intricate dynamics. In this setting, we prove a very general result about equilibrium frequencies of genotypes, bound the mixing time to equilibrium, and find exact expressions for localization in genotype space for a general class of neutral evolutionary processes.

Population structure is known to affect the dynamics and outcome of evolutionary processes, but analytical results for generic random structures have been lacking. We consider a finite population under constant selection whose structure is given by a variety of weighted, directed, random graphs; vertices represent individuals and edges interactions between individuals. By establishing a robustness result and using large deviation estimates to understand the typical structure of random graphs, we prove that the fixation probability of an invading mutant in a randomly structured population is approximately the same as that of a mutant of equal fitness in a well-mixed population with high probability.

At the macroscopic scale, much is known about the timeline of life and evolution on Earth. However, current mathematical models say very little about evolution on these macroscopic timescales and are limited to describing microscopic evolutionary events, like fixation, that occur over relatively few generations. We describe several mathematical properties of genotype space, which provides the stage for long term evolution. These properties are then incorporated into a model of macroscopic evolution that accumulates many microscopic events. In the weak mutation and weak selection regime, we study the time evolution takes to discover novel functionality. Finally, we describe a mechanism called the regeneration process that suggests how evolution might behave like a tinkerer when innovating.

Contents

1 Introduction **1**
- 1.1 Outline . . . 4

2 Genotype Space **6**
- 2.1 The hypercube . . . 9
- 2.2 The symmetric group . . . 17
- 2.3 Reversibility . . . 20
- 2.4 High-dimensional spaces . . . 23
- 2.5 Adding disorder . . . 31
- 2.6 Fitness . . . 43
- 2.7 One-dimensional projections . . . 51
- 2.8 Continuous genotype spaces . . . 55

3 Evolutionary dynamics **57**
- 3.1 Moran process . . . 58
- 3.2 Wright-Fisher process . . . 64
- 3.3 Evolutionary graph theory . . . 67
- 3.4 General evolutionary processes . . . 80

4 Neutral evolution **83**
- 4.1 Coherence assumption . . . 86
- 4.2 Stationary distribution . . . 87
- 4.3 Marginal mixing times . . . 94
- 4.4 Mixing times . . . 99
- 4.5 Correlations in the stationary distributions . . . 102
- 4.6 Localization in genotype space . . . 106

5 Macroscopic evolutionary dynamics **116**
- 5.1 Defining the model . . . 118
- 5.2 Stationary distributions and mixing times . . . 122
- 5.3 Target sets and hitting times . . . 124
- 5.4 Regeneration process . . . 144

6 Randomly structured populations **148**
- 6.1 Robust isothermal theorem . . . 156
- 6.2 Spectral isothermal theorem . . . 161
- 6.3 Random graphs . . . 165
- 6.4 Optimal fluctuations . . . 176

Appendix A Birth-death chains **184**

Appendix B Concentration inequalities **186**

Appendix C The bond percolated hypercube **190**

Appendix D Notation **194**

References **197**

Acknowledgments

It would be impossible to adequately thank everyone who is also responsible for this dissertation—there are far too many of you, and some would merit me writing its length again. Since I cannot be comprehensive and sufficient in my thanks, I will have to settle on being brief.

Harvard. For giving me a life-changing opportunity nine years ago and then letting me stay. I feel truly privileged to have been able to do something I love for five years, and somehow earn a living doing it.

Prof. Martin Nowak. Your indefatigable enthusiasm and support is a constant reminder of why it is better to cooperate! Thank you for your time, thank you for the office, and thank you for encouraging me to investigate my own ideas.

Members of the Program for Evolutionary Dynamics. Work will never be the same without you, but then it never really was work...

Prof. Krishnendu Chatterjee. For your enthusiastic and fruitful collaborations, supporting my trips to Austria IST, and writing me a letter of recommendation on far too little notice.

Prof. Horng-Tzer Yau and your then-postdocs, Antti Knowles, Alex Bloemendal, and Kevin Schnelli. For sparking my curiosity in probability theory and then showing me how the sausage is made in math research.

My family and friends. I am blessed to have you all in my life. To those of you across the Atlantic, I miss you dearly.

Dad. For your infectious love of life and for always encouraging me to find what makes me happy.

Mum. For always, always, always being there when I need you and loving me more than makes any sense.

Penny. You shared many of my struggles putting this dissertation together, but your inability to take life too seriously reminded me to do the same.

Taylor. I could not have wished for a better partner to go on this journey with. Thank you for the considerable ways you have changed my life and for all the innumerable little ways too.

1

Introduction

Evolution by natural selection can be interpreted very broadly. Any process where information is reproduced imperfectly and with differing rates dependent on that information evolves. Stochasticity is intrinsic to this description, and enters into evolution in two main ways. First, via mutation or incorrect copying of information. Second, in how fitness is evaluated by the environment. The increasing appreciation for the generality of this paradigm of evolution is due partly to the success of describing it mathematically [1]. The specific mechanisms that copy the information in an evolutionary process distinguish evolution dynamics from general stochastic processes. In this dissertation, we examine the component parts of evolution by natural selection mathematically, and eventually integrate them together into models of neutral evolution, evolution over long timescales, and evolution in structured populations. This dissertation is based primarily on the content of four of our papers [2–5] and other currently unpublished results. Two more of our papers [6, 7] are within the field of stochastic evolutionary dynamics, but are outside of the focus of this dissertation.

The dissertation has three main themes neutral evolution, longterm or macroscopic evolution, and randomly structured populations. Neutral evolution and the effects of population structure are standard topics in mathematical biology, but macroscopic evolution requires some preamble.

Modern physics has helped us understand our place in the universe and dramatically changed our perception of time [8]. Cosmological inflation lasted for less than 10^{-32} seconds, yet explains the origin of the large-scale structure of the cosmos. Galaxies and solar systems formed over billions of years. Both processes occurred over vastly different timescales. Mathematical models have been essential for developing our understanding of these timescales.

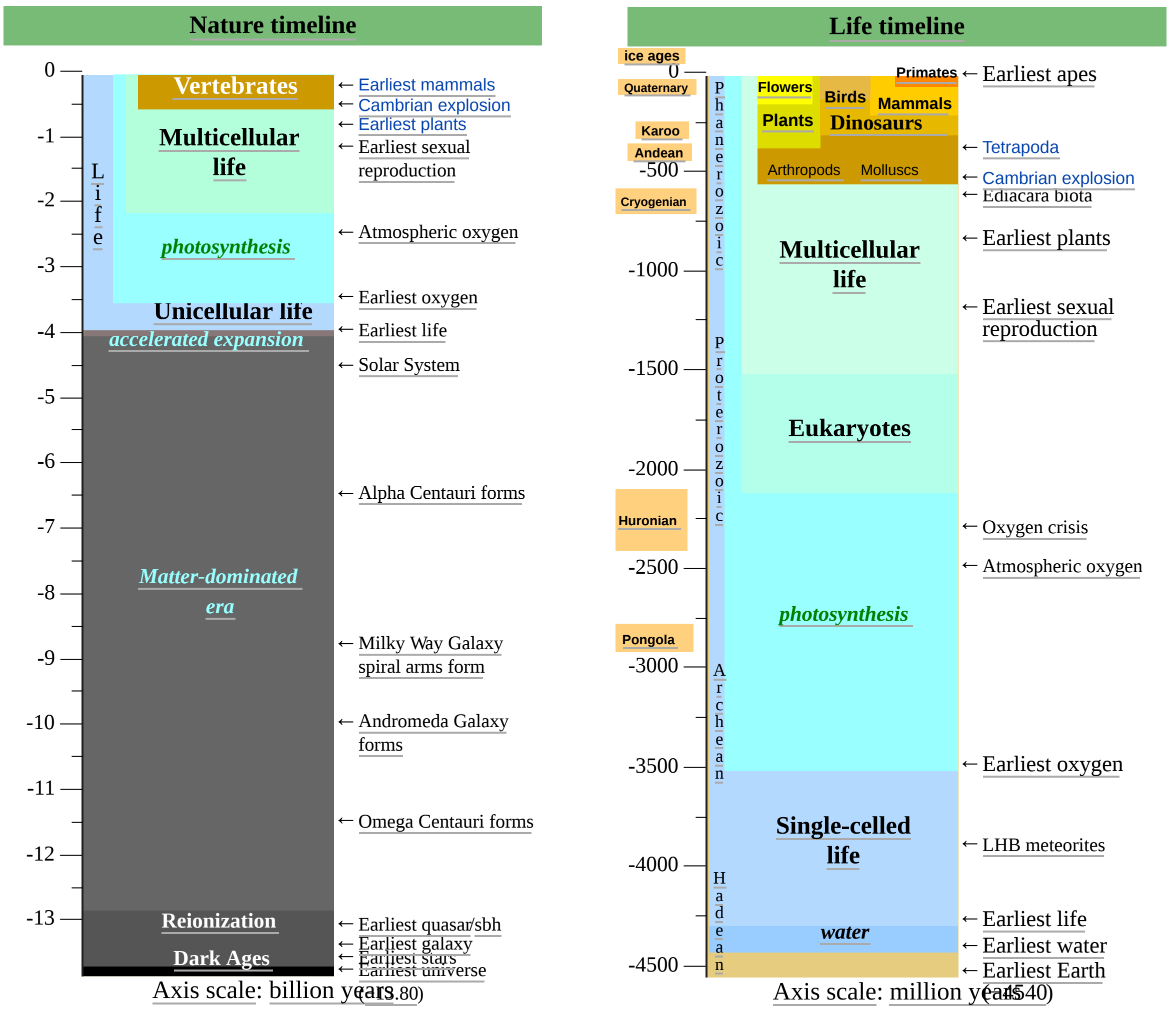

Figure 1.1: Timelines of the physical universe and life [9].

We have learned much about the timeline of evolution from geological and archeological evidence [10]. Life is estimated to have originated as early as 260 million years after the Earth. It took a further 780 million years for photosynthesis to evolve. Was evolution poised to take advantage of an environmental change or can the discovery of novel functionalities take this magnitude of time? Geological evidence suggests that the evolution of eukaryotes and multicellular life is on a similar timescale to photosynthesis. However, the Cambrian explosion, during which the majority of animal phyla evolved, was only 541 million years ago. Is it possible for mathematical models of evolution to explain why the timeline of evolution has this structure?

Immediately following the publication of the Origin of Species, objections were raised that evolution by natural selection could not have had sufficient time to produce the diversity and complexity of life, given contemporary estimates for the age of the earth as approximately 100 million years [11]. Modern physics has since clarified the age of our planet and it is accepted that these 4.54 billion years are adequate for natural selection. However, what are the principle differences in these timescales for evolution [12]?

Our current mathematical models say very little about evolution on these macroscopic timescales and are limited to describing microscopic evolutionary events, like fixation, that occur over relatively few generations [1, 13]. In Chapter 2, we discuss some observations about genotype spaces, which provides the stage for long term evolution. Then, in Chapter 5, we describe a model for macroscopic evolution that accumulates many microscopic events, in the weak mutation and weak selection regime. We find that in this model, evolutionary discoveries can be compared to sampling independently from a specified distribution. Next, we describe a mechanism called the regeneration process that leads to efficient evolution and suggests how evolution might behave as a tinkerer [14].

However, there is a philosophical objection to studying the question of how long evolution takes to discover specific functionalities [15]. Evolution by natural selection does not try to find novel functionalities, it merely describes the competition between types with different rates of replication. Mutations that happen to increase this rate are selected for, but evolution does not try to find such mutations. Asking about the evolution of specific functionalities, is analogous to observing the winning numbers in a lottery, then observing how unlikely those specific numbers were and asking how many resamples it would take to get the same numbers again.

We are sympathetic to this objection, and certainly arguments that use the infeasibility of finding specific sequences in sequence space to conclude evolution by natural selection cannot account for the origin of species in general are preposterous [16]. However, there are at least three reasons questions about discovery times are worth asking. First, the question of how long on average it takes for an evolving population to discover a specific functionality is a well-defined question. Sometimes these question, such as yeast evolving the ability to aerobically metabolize citrate [17,18], seem *post hoc*. Whereas, other functionalities like obtaining energy from the sun's rays (not necessarily using photosynthesis) seem like natural questions to ask *a priori*. Second, if we wish to account for some of the statistical properties of the timeline of evolution, it is unclear how else to approach the question. Third, studying discovery times enables us to understand what affects them, like different mechanisms of mutation and reproduction. Moreover, one can then argue that such mechanisms might be selected for and their effect on discovery times would perhaps explain why.

1.1. Outline

Now we briefly outline the structure of this dissertation. Chapters 2 and 3 are largely expository and lay the ground work for the remaining sections. In Chapter 2, we define and give several significant examples of genotype spaces. A genotype space describes the stage of an evolutionary process, and has three components: a set of labeled genotypes, a rule that describes how they relate to each other through mutation, and a function that assigns a fitness to each genotype. We pay particular attention to high-dimensional genotype spaces, and look in detail at the examples of the hypercube and the symmetric group. We argue that there are biological reasons to expect genotype spaces to be high-dimensional, and demonstrate that many important mathematical properties (in particular, rapid mixing) follow as a consequence of this high-dimensionality. The chapter also contains a long discussion on fitness—what we can learn about it from biological motivations and how it enters into models of genotype space.

Chapter 3 gives many examples of fundamental stochastic models of evolutionary dynamics. We use these models to illustrate the typical question that are asked in the field and motivate several statistics of these processes. We show how these statistics, including fixation probabilities, absorption times, stationary distributions, and other limiting behaviors, can be distorted by population structure. After these examples, we define a very general stochastic model of evolutionary dynamics, which includes the previous examples as special cases and integrates the mutation processes from Chapter 2 into out models.

In Chapter 4, we use the general definition from Chapter 3 to give a simple criterion for an evolutionary process to describe neutral evolution. In this special case of neutral evolution many interesting analytic questions about a process become tractable. Here we address the stationary distribution, mixing times, and localization in genotype space.

Chapter 5 considers evolutionary dynamics in the low mutation and weak selection limit. In this limit, evolution can be tracked over long timescales and described as populations searching genotype space. We ask how long it takes this process to discover various subsets of genotype space. We compare this time to the time random sampling would take to find the subset, and show that in some cases they are equivalent in distribution. In particular, we argue there is a sharp contrast between discovery times that are on average exponential and those that are polynomial. We also identify a simple, biological plausible mechanism, called the regeneration process, that enables discovery in polynomial time.

Finally, in Chapter 6, we return to evolutionary graph theory, which uses weighted, directed graphs to model the effects of populations structure. We generalize the isothermal theorem with a robustness result and then give a sufficient condition for this robustness that can be verified in polynomial time in the size of the graph. We prove that random population structures show behavior that is very close to well-mixed populations with high probability.

2

Genotype Space

The introduction in Chapter 1 described a very general way of thinking about evolution. Any process where information is reproduced imperfectly and with differing rates evolves. While this paradigm is very general, we have to start being more specific to make things interesting. So let us unpack this paradigm by asking four questions. First, we might ask for properties and examples of the evolving information. That is the goal of this chapter. The properties we describe come from thinking through the other components of the paradigm and carefully studying important examples motivated by biology.

Second, what does it mean for information to reproduce? Obviously, the information is not replicating itself—there must be some mechanism that does the copying. The information is not propagated into the future in some arbitrary way, because it is limited by the mechanism that copies it. Often it is reproduced in pieces with a specific structure. Think of how all the DNA in a population is contained within individuals, and it is individuals that actually do the reproducing. This structure informs the way we model the evolution of information. From now on we refer to these pieces or components of information as *genotypes* and the spaces of possible genotypes as a *genotype space*. All the information in an evolutionary process is then a *population* of genotypes.

Third, why and how is the information reproduced imperfectly? The information that is reproduced must resemble the information in the past. This is called heritability in a biological context [19]. Without this, the flow of information is overwhelmed by mutations and we are left with random information and not evolution [20]. Conversely, it is easy to understand why imperfection is necessary as this is the only possible source of novelty. When information is reproduced but with some error, we refer to this as a *mutation*. The particular structure these mutations take is an import aspect of this chapter. For example, think of the many possible errors that can occur when bacterial DNA is duplicated during the interphase of the cell cycle: point mutations, insertions, deletions, duplications frameshifts, reversals *etc*. All of these are a consequence of the mechanism of copying and provide the first entryway for stochasticity into the story [21]. Despite this plethora of possible errata, we can still describe some general properties of these imperfections.

Fourth, what determines these differing rates of reproduction? As we have mentioned, there is some mechanism that copies the information. The action of this mechanism is in turn affected by the information it is reproducing—perhaps in some very complex way. Bring to mind the relationship between an organism's DNA, which provides the blueprint for the mechanism that ultimately reproduces the DNA [21, 22]. This complexity is often bundled into the (philosophically challenging but mathematically simple) notion of *fitness*, which is a parameter of evolutionary models [1, 23]. The rates at which different information is reproduced is what *selection* is discriminating, but its lack of omniscient knowledge of fitness values provides a second entryway for stochasticity [12, 24]. Sometimes the rate of reproduction is independent of the information being reproduced. Such processes are called *neutral evolution* [25]. While this assumption leads to a decoupling of the information follow (particularly, how it mutates) and the mechanism of reproduction (which is thus absent of selection), it still has a rich and developed theory [26–28].

As a brief aside, the information we track in an evolutionary process need not be all the information or completely determine the mechanism. This simply adds noise to the process and supports the idea that selection is not an oracle for fitness [12]. For example, an evolutionary model can focus on the dynamics of a particular gene without recording the rest of the genome.

All the information in our evolutionary process is a population of genotypes, where a genotype is something like the smallest piece of reproducing information. After organizing the information in this way, it is natural to ask what states are possible in our evolutionary process and how do we transition from one state to another. All the processes we consider here are Markov chains, so in that terminology: what is the state space and the transition kernel? We answer these questions completely in Chapter 3, but as a precursor we must answer two preliminary questions. What genotypes are possible and how does mutation produce one genotype from another? This moves our focus from genotypes to genotype spaces.

Potentially, we could think of the possible genotypes as a set Γ and assign labels to each genotype $\alpha_1, \alpha_2, \ldots, \alpha_n$. Then we might think of the probability that a particular genotype is produced when it is copied from another [29–31]. Often simply labeling them is suggestive of how they should relate to each other through mutation. Additionally, we could assign to each genotype α a fitness $\mathcal{F}(\alpha)$. Formally, we have the following definition.

Definition 2.1 (Genotype space). *A genotype space* $(\Gamma, \mathcal{M}, \mathcal{F})$ *is a triple such that: (1)* Γ *is a set containing all possible genotypes; (2)* $\mathcal{M}$ *is a stochastic kernel,*

$$\sum_{\beta \in \Gamma} \mathcal{M}(\alpha, \beta) = 1, \tag{2.0.1}$$

where $\mathcal{M}(\alpha, \beta)$ *denotes the probability that genotype* β *is produced when genotype* α *reproduces (when* Γ *is uncountable the definition changes slightly, see Equation* (2.8.1)*); (3)* $\mathcal{F}$ *is a function* $\mathcal{F} : \Gamma \to \mathbb{R}$ *where* $\mathcal{F}(\alpha)$ *is the fitness of genotype* α*. When* $\mathcal{F}$ *is omitted, it is assumed that the genotypes are neutral, that is,* $\mathcal{F}(\alpha) = 1$ *for all* $\alpha \in \Gamma$*.*

Definition 2.2 (Mutation process). *Note that for a genotype space* $(\Gamma, \mathcal{M}, \mathcal{F})$*, the stochasticity of* $\mathcal{M}$ *implies* $(\Gamma, \mathcal{M})$ *forms a Markov chain. We refer to this Markov chain as a mutation process.*

Remark 2.3. It is also possible to use a genotype space $(\Gamma, \mathcal{M}, \mathcal{F})$ to derive an system of ODEs called the quasi-species equations [29]. These equations have many important consequences, but we do not pursue them here. Because the quasi-species equations are deterministic, they cannot capture many important evolutionary phenomena [1]. However, they can be used to describe the evolution of the probability distribution of certain stochastic models.

So far all we have assumed in Definition 2.1 about how genotypes relate to each other is that it is time-homogenous. However, given $\mathcal{M}$ several different perspectives are useful. As we pointed out in Definition 2.2, $\mathcal{M}$ defines a Markov chain or, equivalently, a random walk on a weighted, directed graph. Natural questions then arise: Is the chain irreducible? Is it reversible? What is its stationary distribution? How quickly does it mix? We might also ask about the geometry of the graph: what is its dimension? How does this related to our previous questions about the Markov chain? All of these question have biological interpretations and often they have typical answers.

We start by develop an intuition for many of these questions, by considering some specific and instructive examples of genotype spaces. First, we consider the hypercube in Section 2.1, which is a model for genotypes that store information as fixed, finite length strings of letters from some finite alphabet. Second, we consider the symmetric group in Section 2.2, where a genotype stores the order of a fixed, finite number of distinct things.

In both examples, we observe a number of key properties. The mutation processes are irreducible, reversible, and have uniform stationary distributions to which they rapidly converge. Geometrically, the spaces are high-dimensional, which contributes to the rapid mixing times and leads to a number of biologically interesting properties we outline in Section 2.4. The properties prove crucial to our analysis of evolution over long timescales in Chapter 5. In Section 2.5, we show that these properties are robust to a number of different types of disorder and random perturbation.

Next, in Section 2.6 we consider fitness [32]. We discuss different ways to specify the function $\mathcal{F}$ and how phenotypes acts as an intermediary. In Section 2.7, we discuss projections of the high-dimensional genotype spaces that are useful mathematically and produce other interesting genotype spaces in their own right. Finally, in Section 2.8 we describe a low dimensional, continuous genotype space to contrast with the other examples.

2.1. The hypercube

A familiar example of a genotype space, and one that informs our later definitions and many typical properties, is sequence space. Sequence space is the the set of all proteins of a given length n, where two sequences are neighbors if they differ by a single amino acid. The concept of considering this space in its totality was introduced by John Maynard Smith [33, 34], but it has since been widely utilized theoretically [35–38] and experimentally [39–42].

Some immediate observations are in order. For each amino acid in the sequence there are 20 choices, meaning that as the sequence increases in length, the size of the space grows exponentially. Specifically, the number of sequences of length n in the genotype spaces is 20^n. Meaning that for even modest protein lengths, a vanishingly small number of these proteins will ever physically exist.

However, this seems to be one of the strengths of recording heritable information in this way—a vast potentiality can be described extremely succinctly. If evolution is to produce to complexity, then it needs a vast stage to do this on, but the heritable information has to remain manageable. This is one biological reason we should expect genotype spaces to be high-dimensional [43].

The notion of neighbors we introduced is immediately suggestive of the sorts of mutation we might consider: single amino acid substitutions. Since we are considering proteins of fixed length, we are ruling out deletions and insertions. We also rule out more complicated mutations like duplications of whole subsequences, reversals, and recombinations. but much of the discussion still applies to more complicated models that include more varied mutations. Single amino acid substitutions mutations introduce a natural geometry to the space. Namely, the distance between two protein sequences is just the minimum number of mutations to transform one into another. Since point mutations are symmetric, we end up with a proper distance metric. While this distance ignores the subtly of the rate at which these mutations occur, it is still a useful concept.

Some observations about the geometry of this space are worth pointing out. After introducing a precise metric for the space, we could go on to apply a mathematical definition of dimension. However, in this case, intuition suggests that the dimension of this space is n, as it is a product of n "components," so as we suggested before the dimension of this space can get arbitrarily large as the sequence length increases. High-dimensional spaces have strange, unintuitive behavior [44] that directly influences how evolution takes place in them [20, 45–48]. Each genotype in the space has many neighbors, suggesting that if we imagine a particle moving around by traveling along edges, at each point there are many directions in which the particle can move. This is quite unlike the familiar random walk in 1, 2, or 3 dimensions. Moreover, if the space is the domain of some function, then the condition for a point to be a local maximum of the function becomes increasingly stringent as the dimension increases (as there are more direction the function must be nonincreasing in) [49–51]. Also, continuity assumptions on the function imply that the function is almost constant and its values are highly concentrated about the mean value [44]. The intuition for this observation comes from understanding typical distances in this space: for almost all pairs of points, the distance between them is $\frac{19}{20}n = \mathcal{O}(n)$, which is very small relative to the size of the space. In fact, the diameter of the space is also $n = \mathcal{O}(n)$, so that any point is accessible in relatively few mutations. A function obeying continuity assumptions cannot change much over small distances in our metric. Putting these two observations together, that most points are relatively close and that there are lots of points at the same distance, suggests that simultaneously satisfying the continuity assumption for all points implies the function should be roughly constant.

2.1.1. Formal definitions. Instead of proteins, we could have considered sequences of DNA or RNA, and there would have been little change in our qualitative observations. So more abstractly, we can formalize genotype spaces of this general type. Genomes in these spaces are sequences from some finite alphabet. Without loss of generality, we use $\{1,\ldots,\kappa\}$ as the letter of our alphabet and define the genotype space as

$$\Gamma_n := \{\boldsymbol{\alpha} = (\alpha_1,\ldots,\alpha_n) : \alpha_1,\ldots,\alpha_n \in \{1,\ldots,\kappa\}\}\,. \tag{2.1.1}$$

The metric on our space is Hamming distance, that is,

$$\mathcal{D}\left(\boldsymbol{\alpha},\boldsymbol{\beta}\right) := \sum_{i=1}^{n} \mathbf{1}\left(\boldsymbol{\alpha}(i) \neq \boldsymbol{\beta}(i)\right). \tag{2.1.2}$$

We refer to this space as the *hypercube.* Now we can define precisely the rates of different mutations in the following way: We stick with single point mutations. We control the rate at which some mutation happens with a parameter $\varepsilon \in (0,1)$. If a mutation happens, we sample a coordinate uniformly from $[\![n]\!]$ and then resample that coordinate uniformly from $[\![\kappa]\!]$. Note that even if there is a mutation, it is possible to remain at the same genotype. Mathematically, we have the following definition.

Definition 2.4 (The single point mutation process on the hypercube). *This process is a Markov chain with state space Γ_n defined in* (2.1.1) *and transition kernel $\mathcal{M}$ defined by*

$$\mathcal{M}\left(\boldsymbol{\alpha},\boldsymbol{\beta}\right) := \begin{cases} \frac{\varepsilon}{\kappa n} & \textit{if } \mathcal{D}\left(\boldsymbol{\alpha},\boldsymbol{\beta}\right) = 1 \\ 1-\varepsilon+\frac{\varepsilon}{\kappa} & \textit{if } \mathcal{D}\left(\boldsymbol{\alpha},\boldsymbol{\beta}\right) = 0 \\ 0 & \textit{otherwise} \end{cases}, \tag{2.1.3}$$

where $\mathcal{D}$ is the distance metric defined in (2.1.2)*. This process is also know as a lazy random walk on the hypercube.*

Remark 2.5. Instead of using $\mathcal{D}$ to define $\mathcal{M}$, we could have done the reverse, that is, for all $(\boldsymbol{\alpha},\boldsymbol{\beta})$ such that $\mathcal{M}\left(\boldsymbol{\alpha},\boldsymbol{\beta}\right) > 0$, define $\mathcal{D}\left(\boldsymbol{\alpha},\boldsymbol{\beta}\right) = 1$. For all remaining pairs $(\boldsymbol{\alpha},\boldsymbol{\beta})$, $\mathcal{D}\left(\boldsymbol{\alpha},\boldsymbol{\beta}\right)$ is the minimum path length from $\boldsymbol{\alpha}$ to $\boldsymbol{\beta}$, that is,

$$\mathcal{D}\left(\boldsymbol{\alpha},\boldsymbol{\beta}\right) = \inf_{(\boldsymbol{\alpha}_0,\boldsymbol{\alpha}_1,\ldots,\boldsymbol{\alpha}_k)} \{k : \boldsymbol{\alpha}_0 = \boldsymbol{\alpha},\ \boldsymbol{\alpha}_k = \boldsymbol{\beta},\ \mathcal{D}\left(\boldsymbol{\alpha}_i,\boldsymbol{\alpha}_{i+1}\right) = 1\}\,. \tag{2.1.4}$$

This is simply the graph distance, where each genotype $\boldsymbol{\alpha}$ is a vertex and there is an edge between $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ if $\mathcal{D}(\boldsymbol{\alpha},\boldsymbol{\beta}) = 1$.

2.1.2. Mathematical notions of dimension for finite discrete spaces. With this notion of distance, we can try to understand mathematically how to specify the dimension of the space. We focus on the case $\kappa = 2$. Firstly, the hypercube is naturally defined as a product space: $\{0,1\} \times \cdots \times \{0,1\}$. We are used to dimensions adding in product spaces, and each component in the product clearly has dimension 1, as we need at least 1 dimension to distinguish between the two points. This would suggest that the space has dimension n.

Another property that suggests the space has dimension n is how the volume of a ball grows as its radius increases. Formally, define a closed ball centered at $\boldsymbol{\alpha}$ of radius r by

$$B_{\boldsymbol{\alpha}}(r) := \{\boldsymbol{\beta} \in \Gamma_n : \mathcal{D}(\boldsymbol{\alpha}, \boldsymbol{\beta}) \leqslant r\}\,. \tag{2.1.5}$$

Then, let $r = \theta n$ for some $\theta \in (0,1)$, so we have

$$|B_{\boldsymbol{\alpha}}(r)| = \sum_{k=0}^{\lfloor r \rfloor} \binom{n}{k} = \Theta\left(\left(\frac{1}{\theta^\theta (1-\theta)^{1-\theta}}\right)^n\right) \tag{2.1.6}$$

for all $\boldsymbol{\alpha} \in \Gamma_n$. Note that this behavior only occurs at mesoscopic scales, as at microscopic scales (e.g much less than the diameter of the space n) the volume grows like $\mathcal{O}(n^r)$ and at macroscopic scales (e.g. larger than the diameter of the space n) the ball's volume is limited by the size of Γ_n for fixed n.

This raises an interesting geometric question: suppose we have a set S of cardinality at least $|B_{\boldsymbol{\alpha}}(r)|$, can the boundary of S be smaller than the boundary of $|B_{\boldsymbol{\alpha}}(r)|$, where the boundary of a set is defined as

$$\partial(S) := \{\boldsymbol{\alpha} \in S : \exists \boldsymbol{\beta} \notin S,\, \mathcal{D}(\boldsymbol{\alpha}, \boldsymbol{\beta}) = 1\}? \tag{2.1.7}$$

The answer is no for the hypercube. This is a theorem due to Harper [52] and relates to the theory of expanders [53,54], which we turn on again in Subsection 2.4.5. Expanders are another hallmark of high-dimensional spaces.

The metric dimension is defined as the minimum cardinality of a resolving set $S \subseteq \Gamma_n$, where a set $S = \{\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_k\}$ is resolving if

$$(\mathcal{D}(\boldsymbol{\alpha}, \boldsymbol{\alpha}_1), \ldots, \mathcal{D}(\boldsymbol{\alpha}, \boldsymbol{\alpha}_k)) \neq (\mathcal{D}(\boldsymbol{\beta}, \boldsymbol{\alpha}_1), \ldots, \mathcal{D}(\boldsymbol{\beta}, \boldsymbol{\alpha}_k)) \tag{2.1.8}$$

for all $\boldsymbol{\alpha}, \boldsymbol{\beta} \in \Gamma_n$. Intuitively, the resolving set is like a coordinate system, where each coordinate measures the distance to a point in the resolving set, and since the set is resolving each point in Γ_n is uniquely specified in the coordinate system. It is known that the metric dimension of the hypercube is $2n/\log n$ asymptotically [55–57].

Remark 2.6. A counter-point that shows how divergent different notions of dimension can be is found by considering the Euclidean dimension of the graph. The Euclidean dimension of the graph is given by the smallest integer d such that the graph can be embedded into $\mathbb{R}^d$, that is, all vertices are mapped to a unique point in $\mathbb{R}^d$ such the Euclidean distance between points who are neighbors in the graph is 1. Since the hypercube is bipartite it can certainly be embedded into $\mathbb{R}^4$: all points that have an even number of 1s can be arranged arbitrarily on the circle $C_0 := \{(x, y, z, w) : x^2 + y^2 = 1/2 \text{ and } z = w = 0\}$, and all points with an odd number of 1s can be arranged arbitrarily on the circle $C_1 := \{(x, y, z, w) : z^2 + w^2 = 1/2 \text{ and } x = y = 0\}$. Since neighboring points on the hypercube have a different parity of their number of 1s, we see that their distance apart in the embedding is

$$x^2 + y^2 + z^2 + w^2 = 1/2 + 1/2 = 1. \tag{2.1.9}$$

Thus, this notion of dimension is inadequate for the high-dimensional spaces we are discussing here.

A consequence of the high-dimensionally of the hypercube is that the distance between two random points drawn from the space is very likely to be close to the average distance between pairs of points. Calculating the average distance between points of the hypercube is simple using the symmetry of the space. There are κ^{2n} pairs of points in the hypercube, so

$$\begin{aligned}
\frac{1}{\kappa^{2n}} \sum_{\boldsymbol{\alpha},\boldsymbol{\beta}\in\Gamma_n} \mathcal{D}(\boldsymbol{\alpha},\boldsymbol{\beta}) &= \frac{1}{\kappa^n} \sum_{\boldsymbol{\alpha}\in\Gamma_n} \mathcal{D}(\mathbf{0},\boldsymbol{\alpha}) \\
&= \frac{1}{\kappa^n} \sum_{\boldsymbol{\alpha}\in\Gamma_n} \sum_{i=1}^{n} \mathbf{1}(\boldsymbol{\alpha}(i) \neq 1) \\
&= \frac{1}{\kappa^n} \sum_{i=1}^{n} \kappa^{n-1}(\kappa - 1) \\
&= n\frac{\kappa - 1}{\kappa}.
\end{aligned} \tag{2.1.10}$$

However, this is not just the distance on average—the distance between almost all points is within $\mathcal{O}(\sqrt{n})$ of this value. This can be expressed by considering a uniform measure on Γ_n and is stated precisely in the following theorem.

Theorem 2.7. *Suppose that $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are drawn uniformly and independently from Γ_n, then*

$$\mathbb{P}\left\{\left|\mathcal{D}(\boldsymbol{\alpha},\boldsymbol{\beta}) - n\frac{\kappa - 1}{\kappa}\right| > \delta\sqrt{n}\right\} \leqslant 2e^{-2\delta^2} \tag{2.1.11}$$

and

$$\mathbb{P}\left\{\mathcal{D}(\boldsymbol{\alpha},\boldsymbol{\beta}) \leqslant (1-\delta)n\frac{\kappa-1}{\kappa}\right\} \leqslant e^{-\frac{\delta(\kappa-1)}{2\kappa}n}. \tag{2.1.12}$$

PROOF. Note that above we found $\mathbb{E}\,\mathcal{D}(\boldsymbol{\alpha},\boldsymbol{\beta}) = n\kappa/(\kappa-1)$. Moreover, since

$$\mathcal{D}(\boldsymbol{\alpha},\boldsymbol{\beta}) = \sum_{i=1}^{n} \mathbf{1}(\boldsymbol{\alpha}(i) \neq \boldsymbol{\beta}(i)), \tag{2.1.13}$$

we find that $\mathcal{D}(\boldsymbol{\alpha},\boldsymbol{\beta})$ is a sum of i.i.d. random variables. Then, applying Lemma B.2, we prove (2.1.11). Equation (2.1.12) follows from Lemma B.3. ∎

The behavior described in Theorem 2.7 is typical of high-dimensional space (see [44] for many more examples).

The dynamical properties of this space are also very nice and amenable to analytic analysis–much of this is due to the reversibility of the mutation process given by Definition 2.4. For a discussion on reversibility see Section 2.3.

LEMMA 2.8. *The single point mutation process on the hypercube is reversible with respect to the uniform distribution and thus its stationary distribution is uniform.*

PROOF. Suppose $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ differ in exactly one coordinate, that is, $\mathcal{D}(\boldsymbol{\alpha},\boldsymbol{\beta}) = 1$. Then

$$\mathcal{M}(\boldsymbol{\alpha},\boldsymbol{\beta}) = \varepsilon\frac{1}{kn} = \mathcal{M}(\boldsymbol{\beta},\boldsymbol{\alpha}). \tag{2.1.14}$$

For $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ that differ in more than one coordinate, the transition probability is 0 in both directions. ∎

REMARK 2.9. Note that we need not have the same mutation rate ε at each coordinate. Instead we specify a mutation rate ε_i for each coordinate, then a mutation occurs with probability ε and if a mutation occurs it happens at coordinate i with probability ε_i. This still leads to a reversible process with respect to the uniform distribution.

DEFINITION 2.10 (RAPID MIXING). *We say a Markov chain $(\Gamma, \mathcal{M})$ is rapidly mixing or mixes rapidly if its mixing time t_{mix} is such that*

$$t_{mix} \leqslant p\left(\log|\Gamma|\right) \tag{2.1.15}$$

for some polynomial p.

The next results bounds the time it takes the single point mutation process on the hypercube to reach its stationary distribution. Note how fast the mixing is relative to the size of the space—the mutation process is rapidly mixing (see Definition 2.10).

LEMMA 2.11. *The mixing time of the single point mutation process on the hypercube is bounded above by*

$$\frac{n}{\varepsilon}\log n. \tag{2.1.16}$$

PROOF. We bound the mixing time using a coupling argument on two processes $\boldsymbol{\alpha}_t$ and $\boldsymbol{\beta}_t$ that marginally are both samples of the mutation process. Let $\boldsymbol{\alpha}_0$ have any initial distribution and suppose that $\boldsymbol{\beta}_0$ is sampled uniformly at random from Γ, that is, in the stationary distribution. Obviously, $\boldsymbol{\beta}_t$ is in the stationary distribution for all t by definition, so once $\boldsymbol{\alpha}_t = \boldsymbol{\beta}_t$, $\boldsymbol{\alpha}_t$ must also be in the stationary distribution. To couple the processes: Sample $\boldsymbol{\beta}_{t+1}$ normally from $\boldsymbol{\beta}_t$. With probability ε, $\boldsymbol{\alpha}_{t+1}$ is equal to $\boldsymbol{\alpha}_t$, otherwise sample a coordinate i uniformly at random from $[\![n]\!]$ and set $\boldsymbol{\alpha}_{t+1}(i) = \boldsymbol{\beta}_{t+1}(i)$ while leaving the other coordinates unchanged. Note that since $\boldsymbol{\beta}_t$ is in the stationary distribution $\boldsymbol{\beta}_t(i)$ is uniform on $[\![\kappa]\!]$ and thus marginally $\boldsymbol{\alpha}_t$ is a sample of the mutation process.

Let T be the coupling time of the two processes, define $T := \min\{t : \boldsymbol{\alpha}_t = \boldsymbol{\beta}_t\}$. We want to bound $\mathbb{P}\{T > t\}$. Define the coupling time of coordinate i by $T_i = \min\{t : \boldsymbol{\alpha}_t(i) = \boldsymbol{\beta}_t(i)\}$, and, since once a coordinate has coupled it stays identical in both process, we have

$$\mathbb{P}\{T > t\} = \mathbb{P}\{T_1 > t, \ldots, T_n > t\} \leqslant \sum_{i=1}^{n} \mathbb{P}\{T_i > t\} = n\left(1 - \frac{\varepsilon}{n}\right)^t, \tag{2.1.17}$$

since each coordinate couples with probability ε/n at each step if it has not already. Thus,

$$\mathbb{P}\left\{T > \frac{n}{\varepsilon}(\log n + c)\right\} \leqslant e^{-c}, \tag{2.1.18}$$

which completes the proof.

As an aside, we note that this bound on T is optimal: Further define the random time that exactly $n-k$ coordinates are the same by

$$S_k := \min\{t : \mathcal{D}(\boldsymbol{\alpha}_t, \boldsymbol{\beta}_t) = k\} \tag{2.1.19}$$

Note that $T = S_0$ and that, since once a coordinate has coupled it stays identical in both process, the sequence

$S_0, \ldots, S_n$ is strictly decreasing. Moreover, given $\mathcal{D}(\boldsymbol{\alpha}_t, \boldsymbol{\beta}_t) = k$, the probability that $S_{k-1} = t+1$ is constant, that is, $S_{k-1} - S_k$ is a geometric random variable with success probability $\varepsilon k/n$. Thus, $\mathbb{P}\{T > t\}$ is largest when $\mathcal{D}(\boldsymbol{\alpha}_t, \boldsymbol{\beta}_t) = n$. Therefore, in this case,

$$\mathbb{E}S = \sum_{k=1}^{n} \mathbb{E}\left(S_{k-1} - S_k\right) = \sum_{k=1}^{n} \frac{1}{\varepsilon}\frac{n}{k} \leqslant C\frac{n \log n}{\varepsilon}. \tag{2.1.20}$$

■

Remark 2.12. Again a result is possible in the varying mutation regime. One can simply replace ε in (2.1.16) by the minimum ε_i. In some cases this upper bound is optimal, but it can be improved depending on exactly how the ε_i are distributed.

As mentioned in the remarks above, it is possible to define many different natural mutation processes on the same genotype space. Another important example on the hypercube is defined below.

Definition 2.13 (The independent point mutation process on the hypercube). *This process is a Markov chain with state space Γ_n defined in* (2.1.1) *and transition kernel $\mathcal{M}$ defined by*

$$\mathcal{M}\left(\boldsymbol{\alpha}, \boldsymbol{\beta}\right) := \left(\frac{\varepsilon}{\kappa}\right)^{\mathcal{D}(\boldsymbol{\alpha},\boldsymbol{\beta})} \left(1 - \varepsilon + \frac{\varepsilon}{\kappa}\right)^{n-\mathcal{D}(\boldsymbol{\alpha},\boldsymbol{\beta})}, \tag{2.1.21}$$

where $\mathcal{D}$ is the distance metric defined in (2.1.2). *This process is a type of product chain.*

Remark 2.14. Note that in Equation (2.1.21) we used Hamming distance to define the transition kernel for the mutation process. Suppose instead we started with the transition kernel and used it to define a notion of distance $\mathcal{D}$ between genotypes-exactly as in Remark 2.5. In this model, we would end up with the distance

$$\mathcal{D}(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \mathbf{1}\left(\boldsymbol{\alpha} \neq \boldsymbol{\beta}\right). \tag{2.1.22}$$

This distance metric does a poor job of conveying distance between mutations in these dynamics, because while it is possible to mutate from any genotype to another (that is, $\mathcal{M}(\boldsymbol{\alpha}, \boldsymbol{\beta}) > 0$ for all $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$), some mutation are much more likely than others. This suggests that the strategy in Remark 2.5 is reasonable when the probabilities of different mutations are of the same order, and otherwise, distance should be weighted in some way inversely to their liklihood.

A particular choice of ε is $\varepsilon = \tilde{\varepsilon}/n$, as in this case the the number of point mutations, $\mathcal{D}(\boldsymbol{\alpha}_t, \boldsymbol{\alpha}_{t+1})$, is approximately Poissonian with parameter $\tilde{\varepsilon}$. As before, the process is reversible with respect to the uniform distribution, since $\mathcal{D}$ is symmetric, and we have the following result on its mixing time.

LEMMA 2.15. *The mixing time of the independent point mutation process on the hypercube is bounded above by*

$$\frac{n}{\varepsilon} \log n. \tag{2.1.23}$$

PROOF. As in the proof of Lemma 2.11, we can couple the process to a process started in the uniform distribution. The process $\boldsymbol{\beta}_t$ starts in the uniform distribution and is updated normally. For each time step and independently for each coordinate i in $[\![n]\!]$, with probability ε set $\boldsymbol{\alpha}_{t+1}(i) = \boldsymbol{\beta}_{t+1}(i)$—otherwise leave the coordinate unchanged.

Again consider the coupling time T and the coupling times of each coordinate, then

$$\mathbb{P}\{T > t\} = \mathbb{P}\{T_1 > t, \dots, T_n > t\} \leqslant \sum_{i=1}^{n} \mathbb{P}\{T_i > t\} = n(1-\varepsilon)^t, \tag{2.1.24}$$

since each coordinate couples with probability ε at each step if it has not already. Thus,

$$\mathbb{P}\left\{T > \frac{1}{\varepsilon}(\log n + c)\right\} \leqslant e^{-c}. \tag{2.1.25}$$

In particular, the choice of $\varepsilon = \tilde{\varepsilon}/n$ yields a mixing time of the same order as that of the single point mutation process. ■

REMARK 2.16 (HISTORICAL). The hypercube has been studied extensively in mathematics. Random walks on the hypercube are a classical topic [58, 59] and their mixing times were one of the first examples of the cut-off phenomena [60–62]. It has even been suggested before that carefully working out the hypercubes properties should be a priority due to its potential impact on biology [63].

2.2. THE SYMMETRIC GROUP

In Section 2.1, we saw an example of a high-dimensional genotype space. To emphasize the importance of high-dimensional genotype spaces, we present another important example in this section. We see that many of the key properties of the space are similar: the space is high-dimensional in a formal sense, each point has many neighbors, the diameter of the space is small relative to its size, most pairs of points are a typical distance from one another, and the mixing time of the mutation process is fast. Although we have already developed some biological intuition as to why we might expect these properties in genotype spaces, it is still useful to see supporting examples.

A useful technique in studying closely related species, called comparative chromosome mapping examines the order of the genes on a particular chromosome [64, 65]. This ordering can change over time due to mutations that reverse the order of particular segments of the chromosome, but species whose gene orderings are "closer" to one another are thought to be more closely related. Once we choose some arbitrary labeling (from $[\![n]\!]$) of the genes, we can specify orderings of these genes using permutations–in this section, each permutation α is a genotype. The space of permutation of the letters $[\![n]\!]$ is the familiar symmetric group:

$$\Gamma_n := \{(i_1, \ldots, i_n) : i_1, \ldots, i_n \in [\![n]\!] \text{ and } i_j \neq i_k \text{ for all } j \neq k\}\,. \tag{2.2.1}$$

So the size of the space Γ_n is $n! \sim \sqrt{2\pi n}(n/e)^n$, where the asymptotic form is given by Stirling's formula.

We mentioned before that in this model mutations can reverse the order of segments of the chromosome and thus change the order of the genes, but to define a particular geometry or dynamics on the space, we need to be more specific. There are several different specific models for mutations on this space and we introduce two of them in this section.

First, we consider mutations that are less biologically plausible in our initial discuss of gene reordering on chromosomes, but easier to study mathematically. A transposition is a permutation that swaps the order of two letters and we denote a transposition of the letters i and j by σ_{ij}. In this model, a mutation happens with probability ε. If there is a mutation, then a transposition is sampled randomly by choosing $i, j \in [\![n]\!]$ independently and uniformly at random. This transposition is then applied to the current genotype to obtain the new genotype. Note that the order that i and j are sampled does not matter, and that it is possible to have $i = j$, in which case the mutation leaves the genotype unchanged. Mathematically, we have the following definition.

Definition 2.17 (The random transposition mutation process). *This process is a Markov chain with state* *space* Γ_n *defined in* (2.2.1) *and transition kernel* $\mathcal{M}$ *defined by*

$$\mathcal{M}(\alpha,\beta) := \begin{cases} \frac{2\varepsilon}{n^2} & \text{if } \beta = \sigma_{ij}(\alpha) \text{ and } i \neq j \\ 1-\varepsilon+\frac{\varepsilon}{n} & \text{if } \alpha = \beta \\ 0 & \text{otherwise} \end{cases}, \tag{2.2.2}$$

where σ_{ij} *is some transposition [66].*

Note that, just as in Remark 2.5, we could define a distance metric using the dynamics above, that is, the distance $\mathcal{D}(\alpha,\beta)$ between two permutations α and β is simply the number of transpositions to turn one into the other. With this distance metric, we immediately see that each point in the space has $\mathcal{O}(n(n-1)/2)$ neighbors. The maximum distance between two points in the space can be found by noting that $\mathcal{D}(\alpha,\beta) = \mathcal{D}\left(\beta^{-1}\alpha,\iota\right)$, where β^{-1} is the inverse permutation of β and ι is the identity permutation $(1,\ldots,n)$. Then one permutation furtherest from ι is $(2,\ldots,n,1)$ at distance n. Thus, the diameter of the space is $\mathcal{O}(n)$, which is small relative to its size.

The random transposition mutation process on the symmetric group is irreducible and aperiodic, since any permutation can be generated by transpositions. Moreover, the process is reversible with respect to the uniform distribution, so its stationary distribution is simply the uniform distribution on Γ_n. The mixing time of this process can be bounded by

$$\frac{2}{\varepsilon} n \log n + \mathcal{O}(n) \tag{2.2.3}$$

using a coupling argument just as we did in the proof of Lemma 2.11 (see chapter 9 of [67] for details). Just like the single point mutation process on the hypercube, the mixing time is rapid relative to the size of the space.

Now we turn to a more realistic model of mutation for gene orders. Define an n-reversal as a permutation that transposes k and $i+j-k$ for $k \in [\![i,j]\!]$ and leaves all other letters fixed. In this new model, a mutation happens with probability ε. If there is a mutation, then an n-reversal is sampled randomly by choosing $i,j \in [\![0,n]\!]$ independently and uniformly at random. This n-reversal is then applied to the current genotype to obtain the new genotype. Note that the order that i and j are sampled does not matter, and that it is possible to have $i=j$, in which case the mutation leaves the genotype unchanged.

Definition 2.18 (The n-reversal chain). *This process is a Markov chain with state space* Γ_n *defined in* (2.2.1) *and transition kernel $\mathcal{M}$ defined by*

$$\mathcal{M}(\alpha, \beta) := \begin{cases} \frac{2\varepsilon}{(n+1)^2} & \text{if } \beta = \xi_{ij}(\alpha) \text{ and } i \neq j \\ 1 - \varepsilon + \frac{\varepsilon}{n+1} & \text{if } \alpha = \beta \\ 0 & \text{otherwise} \end{cases}, \tag{2.2.4}$$

where ξ_{ij} is an n-reversal.

Note that as before, the n-reversal chain is reversible with respect to the uniform distribution. In [68, 69], the mixing is found to be bounded by $(2 + 10^{-10})n \log n/\varepsilon$. Again, the process is rapidly mixing.

Remark 2.19. We have only give two examples of mutation processes on the symmetric group, but the properties we have seen are conjectured to hold much more generally. For example, considering the diameter of the space is a standard question in group theory—we are simply asking about the diameter a Cayley graph of the group. The diameter of such graphs is conjectured to be at most $\mathcal{O}(n^2)$ [70]. Another conjecture concerns the mixing times of such processes on the symmetric group: they have been conjectured to be at most $\mathcal{O}(n^3 \log n)$ [71]. Thus, in general, we might expect the diameter and mixing time to be small relative to the size of the space. In Subsection 2.4.5, we return to random transposition processes on the symmetric group and show that even when we restrict which transpositions are possible the process still mixes rapidly.

Remark 2.20 (Historical). Again the symmetric group is a classic object of study in algebra and combinatorics [72]. Random walks on the symmetric were first motivated by card shuffling and date back at least to Markov [73] and Poincare [74]. Again their mixing times have been the object of a great deal of study [62, 66, 75, 76]. Durrett is associated with the mathematical analysis of the biologically motivated n-reversal chain [68, 69]. Geometrical properties of these processes on the symmetric group are considered in [77].

2.3. Reversibility

The models for mutation we have considered so far have all been reversible. This section contains a brief mathematical discussion about reversibility and why we might expect it biologically. Reversibility requires that if we sample a Markov chain $(x_t)_{t\in\mathbb{N}}$ up to any finite time t and consider the joint distribution of the sample $(x_0, \ldots, x_t)$, then it should be the same as the reverse of this sequence $(x_t, \ldots, x_0)$. Specifically, for $t = 1$, for reversibility we require

$$\pi(\alpha)\mathcal{M}(\alpha, \beta) = \mathbb{P}\{x_0 = \alpha\}\mathbb{P}\{x_1 = \beta | x_0 = \alpha\} = \mathbb{P}\{x_0 = \alpha, x_1 = \beta\} = \mathbb{P}\{x_1 = \alpha, x_0 = \beta\} = \pi(\beta)\mathcal{M}(\beta, \alpha) \quad (2.3.1)$$

for all $\alpha, \beta \in \Gamma$, which is exactly the condition we have been using to verify reversibility in the mutation processes above. We use condition (2.3.1), as it implies reversibility for sequences of any length:

$$\begin{aligned}
\mathbb{P}\{x_0 = \alpha_0, \ldots, x_t = \alpha_t\} &= \mathbb{P}\{x_0 = \alpha_0\}\mathbb{P}\{x_1 = \alpha_1 | x_0 = \alpha_0\} \cdots \mathbb{P}\{x_t = \alpha_t | x_0 = \alpha_0, \ldots, x_{t-1} = \alpha_{t-1}\} \\
&= \pi(\alpha_0)\mathcal{M}(\alpha_0, \alpha_1) \cdots \mathcal{M}(\alpha_{t-1}, \alpha_t) \\
&= \mathcal{M}(\alpha_1, \alpha_0)\,\pi(\alpha_1)\mathcal{M}(\alpha_1, \alpha_2) \cdots \mathcal{M}(\alpha_{t-1}, \alpha_t) \\
&= \mathcal{M}(\alpha_1, \alpha_0)\,\mathcal{M}(\alpha_2, \alpha_1) \cdots \mathcal{M}(\alpha_t, \alpha_{t-1})\pi(\alpha_t) \\
&= \mathbb{P}\{x_t = \alpha_0, \ldots, x_0 = \alpha_t\}. \qquad (2.3.2)
\end{aligned}$$

Assume that the Markov chain is irreducible and aperiodic. Note that the stationarity of π for $\mathcal{M}$ is immediate from condition (2.3.1)—simply sum over α to see

$$\sum_{\alpha \in \Gamma} \pi(\alpha)\mathcal{M}(\alpha, \beta) = \sum_{\alpha \in \Gamma} \pi(\beta)\mathcal{M}(\beta, \alpha) = \pi(\beta). \quad (2.3.3)$$

2.3.1. Random walks on graphs and symmetric Markov chains. A special case of reversible Markov chains are those that can be defined as a random walk on a weighted graph. That is consider a graph with vertex set Γ and associate a symmetric weight $\mathcal{W}(\alpha, \beta) = \mathcal{W}(\beta, \alpha) \geqslant 0$ for each pair of vertices $\alpha, \beta \in \Gamma$. Then define the following transition kernel for a Markov chain with state space Γ:

$$\mathcal{M}(\alpha, \beta) := \frac{\mathcal{W}(\alpha, \beta)}{\sum_{\beta'} \mathcal{W}(\alpha, \beta')}. \quad (2.3.4)$$

Note that under the definition in (2.3.4), the Markov chain is irreducible exactly when the graph is connected and aperiodic when it is not bipartite. One easily finds that the process is reversible with respect to the distribution

$$\pi(\alpha) = \frac{\sum_{\beta} \mathcal{W}(\beta, \alpha)}{\sum_{\alpha', \beta} \mathcal{W}(\alpha', \beta)}, \quad (2.3.5)$$

since

$$\pi(\alpha)\mathcal{M}(\alpha,\beta) = \frac{\sum_\beta \mathcal{W}(\beta,\alpha)}{\sum_{\alpha',\beta}\mathcal{W}(\alpha',\beta)}\frac{\mathcal{W}(\alpha,\beta)}{\sum_{\beta'}\mathcal{W}(\alpha,\beta')} = \pi(\beta)\mathcal{M}(\beta,\alpha). \tag{2.3.6}$$

This special case includes the single point mutation process on the hypercube and the random transposition mutation process. In fact, their mutation kernels are actually symmetric, that is

$$\mathcal{M}(\alpha,\beta) = \mathcal{M}(\beta,\alpha) \tag{2.3.7}$$

for all $\alpha,\beta \in \Gamma$. Symmetric mutation kernels are a special case of random walks on graphs, where the quantity

$$\sum_\beta \mathcal{W}(\alpha,\beta) \tag{2.3.8}$$

does not depend on α. Obviously, this is a stronger condition than reversibility and it implies reversibility with respect to the uniform distribution. Symmetry is a common feature of models of genotype space but there is no reason for this to always be so.

2.3.2. Kolmogorov condition for reversibility. While symmetric mutation kernels cannot always be expected, one can give an argument for reversibility in general. Consider a path through genotype space that starts and ends at the same genotype, that is, a cycle $(\alpha_0,\alpha_1,\ldots,\alpha_{k-1},\alpha_0)$. By iterating the reversibility condition (2.3.1) over the cycles, we obtain the Kolmogorov condition for reversibility:

$$\begin{aligned}\pi(\alpha_0)\mathcal{M}(\alpha_0,\alpha_1)\cdots\mathcal{M}(\alpha_{k-1},\alpha_0) &= \mathcal{M}(\alpha_1,\alpha_0)\pi(\alpha_1)\mathcal{M}(\alpha_1,\alpha_2)\cdots\mathcal{M}(\alpha_{k-1},\alpha_0)\\ &= \mathcal{M}(\alpha_1,\alpha_0)\mathcal{M}(\alpha_2,\alpha_1)\cdots\mathcal{M}(\alpha_0,\alpha_{k-1})\pi(\alpha_0).\end{aligned} \tag{2.3.9}$$

Equivalently,

$$\mathcal{M}(\alpha_0,\alpha_1)\cdots\mathcal{M}(\alpha_{k-1},\alpha_0) = \mathcal{M}(\alpha_0,\alpha_{k-1})\mathcal{M}(\alpha_{k-1},\alpha_{k-2})\cdots\mathcal{M}(\alpha_1,\alpha_0). \tag{2.3.10}$$

This implies that if we are watching a mutation process over time, the process transitions the cycle in one direction just as frequently as the other.

Moreover, this condition implies that the expected time to travel from state α to state β is the same as the expected time to travel from state β to α (See Section 5.3). We can interpret this condition biologically. *A priori*, mutation has to bias for one genotype over another. We might take this condition as a component of the assumption that mutation is unstructured.

All of the mutation processes we consider are reversible and we take reversibility as an assumption in many of our theorems. It is key to making our models tractable, as it allows the use of many tools from the theory of Markov chains. We feel this approach is justified on three fronts: 1) reversible mutation processes include many significant examples such as the hypercube and the symmetric group; 2) there is a biological argument, detailed above, for reversibility; 3) reversibility is a common assumption in such biological models and is common in the literature due to its implication on model tractability [78–80].

However, there are some important examples that are not reversible. One simple example is to consider mutations that delete whole segments of the genotype. In this case the probability of deleting some gene say is positive ($\mathcal{M}(\alpha, \beta) > 0$) but the probability of this gene arising from nothing (say there is not another copy of the gene that can arise through gene duplication) is zero ($\mathcal{M}(\beta, \alpha) = 0$), in which case the equation (2.3.1) cannot be satisfied:

$$\pi(\alpha)\mathcal{M}(\alpha, \beta) > 0 = \pi(\beta)\mathcal{M}(\beta, \alpha), \tag{2.3.11}$$

So long as there are other mutations that ensure that the mutation process is irreducible (and thus assigns positive probability to each genotype in the stationary distribution, $\pi(\alpha) > 0$).

2.4. High-dimensional spaces

In this section, we consider again some of the properties we have seen in the case of the hypercube in Section 2.1 and the symmetric group in Section 2.2. We argue that there are biological reasons to expect these properties in genotype spaces, and that these properties are all suggestive of high-dimensionality. We saw that under most understandings of dimensionality, the spaces we previously considered are high-dimensional. In the second part of this section, we review some geometrical techniques for bounding the mixing times of a Markov chain. These techniques are particularly effective in producing small bounds when the Markov chain is high-dimensional. Together this suggests that we should expect the mutation processes on genotype spaces to be rapidly mixing.

2.4.1. Each genotype has many neighbors. We saw that both the single-point mutation process on the hypercube and the random transposition mutation process on the symmetric group yield many possible mutations for each genotype. The degree of the hypercube is n and the degree of the symmetric group is $n(n-1)/2$ under transpositions. What is important is that the number of neighbors grows arbitrarily large as the space gets larger. This is a feature of high-dimensional spaces in general. While we have not yet introduced fitness, biologically it could be argued that helps evolution avoid becoming stuck in local maxima [43, 81, 82].

2.4.2. Volumes grow rapidly as distances increase. For the hypercube, an easy estimate (2.1.6) showed that the volume of a closed ball of radius r grows like C_r^n when $r = \Theta(n)$. This order of growth for the volume of a closed ball is exactly what we find in n-dimensional Euclidean space $\mathbb{R}^n$. So again this sort of property is suggestive of high-dimensional space. A similar estimate, which shows the dimension of the space growing with n, is possible for the symmetric group under transpositions. This property also implies that the boundary of a closed ball is large compared to the total volume of the ball. In this way, the number of genotypes accessible by a fixed number of mutations increases rapidly with the number of mutation.

2.4.3. Short distances between points. A related property to the growth of volume is that the distances between points is small relative to the size of the space. The size of the hypercube is κ^n and the size of the symmetric group is $n! \sim \sqrt{2\pi n}(n/e)^n$. However, we saw in the hypercube that the distance between most points is $(\kappa-1)/\kappa n$ in Theorem 2.7, and that its diameter is n. Similarly in the symmetric group under transpositions, we saw its diameter is $n-1$. In both cases, the diameter of the space grows logarithmically in the size of the space. The property that all points are close together even though the space is large is also suggestive of a high-dimensional space.

We can make a biological argument for why we might expect this property in genotype spaces. The further two genotypes are from each other the longer it takes for mutations to produce one from the other, even in the presence of a strong selective advantage. An evolutionary system that can travel along this selective gradient quickly could outcompete a different system whose speed is limited by mutation [83]. Moreover, in large genotype spaces where the maximal distance between genotypes does not grow slowly, many genotypes could be practically inaccessible to each other with mutation in realistic timeframes. High-dimensional spaces are the only way to increases the potential genotypes that evolution can explore. In this sense, high-dimensional spaces can store lots of genotypes but represent them compactly.

2.4.4. Product spaces. While not all genotype spaces need be product spaces, this is a way to obtain high-dimensional spaces, as the total dimension of the space is given by the sum of the dimensions of its components. Moreover, biologically this may make sense if the genotype has different modular components [84, 85].

2.4.5. Mixing times. In Lemmas 2.11 and 2.15, we found that the mixing time of the hypercube, for both the single (see Definition 2.4) and independent (see Definition 2.13) point mutation processes, is bounded above by $n \log n/\varepsilon$. Similarly, we saw that the symmetric group under random transpositions (see Definition 2.17) has a mixing time bounded by $\frac{2}{\varepsilon} n \log n + \mathcal{O}(n)$ and for the n-reversal chain (see Definition 2.18), the mixing time is bounded by $\mathcal{O}\left(n^3 \log n\right)$. While the bounds vary, they still imply rapid mixing. We do not claim that these bounds (or any we obtain later) are optimal. Often optimal bounds are unnecessary as the size of the space completely dominates the mixing time. Finding optimal upper and lower bounds on mixing times is an active area of research and often requires advanced and careful estimates [67]. What we are interested in are techniques to prove upper bounds on mixing times that show mixing is rapid, and we prioritize the generality and robustness of these techniques over their ability to produce optimal bounds. We feel that general and robust techniques are easier to interpret and apply biologically.

In general, we might expect rapid mixing in high-dimensional genotype spaces [86]. To make this intuition more precise. We introduce two general bounds on the mixing times of Markov chains. These bounds are both in terms of the eigenvalues of the Markov chain, specifically the spectral gap, which contain lots of information about the geometry of the Markov chain [67]. Let $\mathcal{M}$ be the transition kernel for an irreducible, aperiodic, reversible Markov chain with state space Γ. Since it is irreducible and aperiodic it has a unique stationary distribution π with respect to which it is reversible. Then define

$$\mathcal{Q}(\alpha, \beta) := \pi(\alpha)\mathcal{M}(\alpha, \beta) = \pi(\beta)\mathcal{M}(\beta, \alpha) \tag{2.4.1}$$

for each $\alpha, \beta \in \Gamma$. Let $\lambda_1, \ldots, \lambda_K$ denote the eigenvalues of $\mathcal{M}$ ordered nonincreasingly, where $K := |\Gamma|$. Also denote the eigenvector associated to λ_k by u_k. By the Perron-Frobenius theorem, $|\lambda_i| \leqslant 1$ for all i. In fact, because $\mathcal{M}$ is stochastic $\lambda_0 = 1$. Irreducibility implies $\lambda_i < 1$ for all $i < 1$. Aperiodicity implies $\lambda_K > -1$. Now, define

$$\lambda_* := \max\{|\lambda_i| : \text{for } i > 1\} \tag{2.4.2}$$

and define the spectral gap as $\gamma_* := 1 - \lambda_*$. Our observations above imply that the gap is positive—$\gamma_* > 0$. In particular, for reversible chains, define $\gamma_* = 1 - \lambda_2$. Spectral information about $\mathcal{M}$ can be used to bound the point-wise convergence of the distribution of a Markov chain's distribution to its stationary distribution, since

$$\mathbb{P}_\alpha\left\{\alpha_t = \beta\right\} = \mathcal{M}^t(\alpha,\beta) = \pi^{-1/2}(\alpha)\pi^{1/2}(\beta)\sum_{k=1}^{n} u_\alpha(k)u_\beta(k). \tag{2.4.3}$$

This immediately implies

$$\left|\mathbb{P}_\alpha\left\{\alpha_t = \beta\right\} - \pi(\beta)\right| = C\left(1-\lambda_2\right)^t \tag{2.4.4}$$

for some constant C not dependent on t. For reversible chains, it is easy to use this spectral information to bound the mixing time. We denote the relaxation time of a Markov chain with $t_{\text{rel}} := 1/\gamma_*$.

THEOREM 2.21. *Let $\mathcal{M}$ be the transition kernel for an irreducible, aperiodic Markov chain with state space Γ. Assume the Markov chain is reversible with respect to π. Denote the relaxation time of $\mathcal{M}$ by t_{rel}, then*

$$t_{rel} \leqslant t_{mix} \leqslant t_{rel}\log\frac{1}{\min_{\alpha\in\Gamma}\pi(\alpha)}. \tag{2.4.5}$$

For a proof, see [67].

Now we turn to the geometric bounds on the spectral gap. Define the quantity

$$\Phi_* := \min_{A:\pi(A)\leqslant 1/2}\frac{\mathcal{Q}(A,A^c)}{\pi(A)}, \tag{2.4.6}$$

where $\pi(A) := \sum_{\alpha\in A}\pi(\alpha)$ and

$$\mathcal{Q}(A,B) := \sum_{\alpha\in A}\sum_{\beta\in B}\mathcal{Q}(\alpha,\beta). \tag{2.4.7}$$

The quantity Φ_* has many names—including Cheeger's constant, the bottleneck ratio, and the isoperimetric ratio. Intuitively, Φ_* measures the maximum flow of probability in the Markov chain when it is in the stationary distribution out of all subsets A normalized by their size under the measure π. If the flow is large for all subsets, this suggests that the probability flows rapidly around the states of the Markov chain without any bottlenecks, in which case, the mixing time should be fast. If we think of $\pi(A)$ measuring the volume of A and $\mathcal{Q}(A,A^c)$ measuring the surface area, then we see Φ_* has a geometric interpretation as a kind of isoperimetric quantity. We can quantify this argument with the following theorem [67].

THEOREM 2.22 (CHEEGER'S INEQUALITY). *Let $\mathcal{M}$ be the transition kernel for an irreducible, aperiodic Markov chain* *with state space* Γ*. Assume the Markov chain is reversible with respect to* π*. Denote the spectral gap of* $\mathcal{M}$ *by* γ_* *and* Φ_**, then*

$$\frac{\Phi_*^2}{2} \leqslant \gamma_* \leqslant 2\Phi_*. \tag{2.4.8}$$

In high-dimensional spaces, satisfying Equations like (2.1.6) for how volumes grow for balls of increasing radius, bottlenecks are prevented as they would disrupt the growth. For example, for the hypercube we already mentioned that balls have the minimum boundary size for a fixed volume due to Harper's theorem [52]. So for the single point mutation process on the hypercube with $\kappa = 2$, we have

$$\Phi_* = \min_{r : r \leqslant n/2} \frac{\frac{1}{2^n}\varepsilon\frac{n-r}{n}\binom{n}{r}}{\frac{1}{2^n}\sum_{k=0}^{r}\binom{n}{k}} = \frac{\varepsilon}{n}. \tag{2.4.9}$$

Thus, applying Theorems 2.21 and 2.22, we find $t_{\mathrm{mix}} \leqslant \mathcal{O}\left(n^3/\varepsilon\right)$. This bound is much worse than our bound from Lemma 2.11, however, it still implies rapid mixing!

Theorem 2.22 is one way to quantify this idea that a lack of bottlenecks accelerates mixing. Now we introduce another approach that uses a quantity called the congestion ratio. To state this bound, we first introduce a more general technique called the path method. The path method is a way to compare the mixing times of two chains on the same state space, but as a special case of this method we obtain a bound on the mixing time of a single chain in terms of a geometric quantity.

Again with $\mathcal{M}$ the transition kernel for an irreducible, aperiodic, reversible Markov chain with state space Γ, define $E := \{(\alpha, \beta) : \mathcal{M}(\alpha, \beta) > 0\}$. Then an E-path from α to β is defined as a sequence of states from Γ, denoted $\phi_{\alpha\beta} = (\alpha_0, \ldots, \alpha_k)$ such that $\alpha = \alpha_0$, $\beta = \alpha_k$, and $(\alpha_i, \alpha_{i+1}) \in E$ for all i. The length of the E-path ϕ is k.

Also let $\tilde{\mathcal{M}}$ be the transition kernel for an irreducible, aperiodic, reversible Markov chain with state space Γ. Define $\tilde{E} := \{(\alpha, \beta) : \tilde{\mathcal{M}}(\alpha, \beta) > 0\}$. Then for each $(\alpha, \beta) \in \tilde{E}$, choose and fix some E-path from α to β (there must exist one by irreducibility) and denote it with $\phi_{\alpha\beta}$. Then define the *congestion ratio*

$$B := \max_{(\alpha,\beta)\in E}\left(\frac{1}{\mathcal{Q}(\alpha, \beta)} \sum_{\tilde{\alpha},\tilde{\beta} : (\alpha,\beta)\in\phi_{\tilde{\alpha}\tilde{\beta}}} \tilde{\mathcal{Q}}(\tilde{\alpha}, \tilde{\beta})|\phi_{\tilde{\alpha}\tilde{\beta}}|\right), \tag{2.4.10}$$

where the sum is over all pairs $\tilde{\alpha}, \tilde{\beta}$ such that the edge (α, β) occurs in the E-path from $\tilde{\alpha}$ to $\tilde{\beta}$. Roughly speaking, by assigning an E-path to each pair in $\tilde{E}$, we are trying to replicate the flow of probabilities in the chain $\tilde{\mathcal{M}}$ with the first chain $\mathcal{M}$. The congestion ratio is then used to measure how the flows defined by the E-paths depend on each edge (α, β). Now, we have the following theorem [67].

THEOREM 2.23 (THE COMPARISON THEOREM). *Let $\mathcal{M}$ and $\tilde{\mathcal{M}}$ be the transition kernels for irreducible, aperiodic Markov chains with state space Γ. Assume both Markov chains are reversible with respect to π and $\tilde{\pi}$ respectively. Denote the spectral gaps of $\mathcal{M}$ and $\tilde{\mathcal{M}}$ by γ_* and $\tilde{\gamma}_*$ respectively. Suppose that B is the congestion ratio defined in* (2.4.10) *for some fixed choice of E-paths. Then*

$$\tilde{\gamma}_* \leqslant \left(\max_{\alpha \in \Gamma} \frac{\pi(\alpha)}{\tilde{\pi}(\alpha)} \right) B \gamma_*. \tag{2.4.11}$$

From Theorem 2.23, we can immediately obtain the geometric bound we promised above for a single chain.

COROLLARY 2.24. *Let $\mathcal{M}$ be the transition kernel for an irreducible, aperiodic Markov chain with state space Γ. Assume the Markov chain is reversible with respect to π. Denote the spectral gap of $\mathcal{M}$ by γ. For each pair $\alpha, \beta \in \Gamma$ fix some E-path $\phi_{\alpha\beta}$ and define*

$$B := \max_{(\alpha', \beta') \in E} \frac{1}{\mathcal{Q}(\alpha', \beta')} \sum_{\alpha, \beta : (\alpha', \beta') \in \Gamma_{\alpha, \beta}} \pi(\alpha) \pi(\beta) |\phi_{\alpha\beta}|, \tag{2.4.12}$$

then

$$\gamma_* \geqslant 1/B. \tag{2.4.13}$$

PROOF OF COROLLARY 2.24. The proof follows quickly by letting $\mathcal{M}(\alpha, \beta) = \pi(\beta)$ and applying Theorem 2.23. Then obviously, $\mathcal{M}$ is an irreducible, aperiodic Markov chain that is reversible with respect to π. Moreover, the eigenvalues of $\mathcal{M}$ are easily calculated as $\lambda_1 = 1$ and $\lambda_i = 0$ for $i > 1$. ∎

In Theorem 2.23 we specified a fixed choice of E-paths. The choice of these E-paths can greatly affect the quality of the bound obtained. One useful technique for obtaining good bounds when the path choice is unclear is to average over many choices for the paths. Specifically, let $\nu_{\alpha\beta}$ be a measure on the set of E-paths from α to β—this measure effectively describes how to sample a random path from α to β and leads to us averaging over all of them. Now, for each $(\alpha, \beta) \in \tilde{E}$, fix some distribution $\nu_{\alpha\beta}$ on the set of E-path from α to β, then the congestion ratio is give by

$$B := \max_{(\alpha, \beta) \in E} \left(\frac{1}{\mathcal{Q}(\alpha, \beta)} \sum_{(\tilde{\alpha}, \tilde{\beta}) \in \tilde{E}} \tilde{\mathcal{Q}}(\tilde{\alpha}, \tilde{\beta}) \sum_{\phi_{\tilde{\alpha}\tilde{\beta}} : (\alpha, \beta) \in \phi_{\tilde{\alpha}\tilde{\beta}}} \nu_{\tilde{\alpha}\tilde{\beta}} \left(\phi_{\tilde{\alpha}\tilde{\beta}} \right) |\phi_{\tilde{\alpha}\tilde{\beta}}| \right). \tag{2.4.14}$$

With this new definition of the congestion ratio, we obtain new versions of Theorem 2.23 and Corollary 2.24 by replacing the congestion ratio with our new congestion ratio defined in (2.4.14).

At the beginning of this section, we discussed how short paths between all pairs of points is a property associated with high-dimensional spaces and the specific examples of genotype spaces we have given. Equation (2.4.12) for the congestion ratio explains partly why this is a significant property for us: many short paths between genotypes leads to rapid mixing. Additionally, in high-dimensional spaces there are often many short paths from one point to another. By choosing these short paths intelligently or randomizing over them, one is further able to avoid bottlenecks and prove rapid mixing.

For example, on the hypercube there is a path of distance at most n between each pair of points. In fact, for each pair of points $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ such that $\mathcal{D}(\boldsymbol{\alpha}, \boldsymbol{\beta}) = k$, consider the following path ϕ between them: starting from the left-hand side, change any coordinate one at a time that differs in $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. Clearly this path has length at most n. Then each edge $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ in the hypercube is contained in at most κ^{n+1} paths, by summing over part of the path before the edge $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ and the part after the edge $(\boldsymbol{\alpha}, \boldsymbol{\beta})$. Thus, we can bound the quantity (2.4.12) by

$$B \leqslant \max_{(\alpha', \beta') \in E} \frac{\kappa n}{\varepsilon} \kappa^n \sum_{\alpha, \beta : (\alpha', \beta') \in \Gamma_{\alpha, \beta}} \frac{1}{\kappa^{2n}} n \leqslant \kappa^2 n^2. \tag{2.4.15}$$

So applying Corollary 2.24 and Theorem 2.21, we find the mixing time of the single point mutation process on the hypercube is less than

$$\mathcal{O}\left(\frac{n^3}{\varepsilon}\right). \tag{2.4.16}$$

Again, worse than Lemma 2.11, but still rapid mixing.

The above example illustrates how Corollary 2.24 can be used to bound mixing time for spaces that are high-dimensional in that they have many short paths between points. Now, we have given an example of using Theorem 2.23 to bound the mixing times of chains that are related to some already understood chain. Recall the random transposition mutation process defined in Definition 2.17. In this process, we sampled pairs $i, j \in [\![n]\!]$ uniformly and independently. We could instead restrict which pairs are allowed and then sample uniformly from this set [87]. It is useful to think of this set of allowable mutations forming the edge set E of a graph G with vertex set $[\![n]\!]$. Then the process samples edges (i, j) of G uniformly at random and applies the permutation σ_{ij} to the process. Note that G can have self loops and in fact we assume each vertex has a self loop to avoid periodicity. We also assume that G is connected to ensure irreducibility. Clearly, the model we previously considered corresponds to the complete graph. One other interesting example is given by the star graph, where permutations σ_{1i} are selected randomly, that is, the first gene on the chromosome is shuffled into a random location. A second intersecting example comes from the cycle graph, where permutations $\sigma_{i(i+1)}$ are selected randomly, that is, genes are only swapped locally. Both examples are interesting biologically.

Using Theorem 2.23, we can bound all processes of this type. Note that any transposition σ_{ij} can be replicated on the restricted chain as follows: find an E-path from i to j, denoted ϕ_{ij}, then for each edge e in the path apply the transposition σ_e; then apply the same transpositions (except the last) in reverse order. Thus, the maximum path length is $2D$, where D is the diameter of G. Consider B from equation (2.4.10)

$$\max_{(i,j)\in E} \frac{1}{\frac{1}{n!}\frac{\varepsilon}{|E|}} \sum_{i',j':(i,j)\in\phi_{i'j'}} \frac{\varepsilon}{n!n^2} 2D = \frac{2D|E|}{n^2} \max_{(i,j)\in E} \sum_{i',j':(i,j)\in\phi_{i'j'}} 1. \tag{2.4.17}$$

The bound obtained in (2.4.17) obviously depends on the structure of the graph G. However, we can certainly say that $D \leqslant n$, $|E| \leqslant n^2$, and $\sum_{i',j':(i,j)\in\phi_{i'j'}} 1 \leqslant n^2$, thus

$$B \leqslant 2n^3. \tag{2.4.18}$$

Thus, Theorem 2.21 implies that the mixing time for the random transposition process on G is bounded by

$$4\frac{n^4}{\varepsilon}\log n + \mathcal{O}(n^4). \tag{2.4.19}$$

Again, while the bound is not optimal, it still implies rapid mixing.

Remark 2.25. Some Markov chains are highly structured and exhibit interesting symmetries that can be exploited to obtain optimal bounds on mixing times and exact expressions for the generating functions of hitting times. This structure takes the form that the state space of the Markov chain can be interpreted as a group with some rule for composing states, then the transition kernel is equivalent to sampling elements of the group with some fixed distribution and composing that element with the current state. In this setting it is possible to use techniques from representation theory to great effect. We mention this here because both the hypercube and symmetric group have this type of structure [63]. We do not pursue these techniques here, because the arguments require the introduction of a lot of mathematic machinery and we are not concerned with optimal bounds. Moreover, the group structure is very fragile and does not survive perturbation.

2.5. Adding disorder

So far the genotype spaces we have considered have been highly structured and symmetric (see Remark 2.25). It is unrealistic to expect to see exactly these structures in true biological genotype spaces, so we do not want our observations to break down when these spaces are perturbed and are no longer symmetric. If our results are robust to perturbation, it increases their applicability to real systems and encourages us that our observations should apply.

Our main focus in this section is to discuss and analyze different methods for perturbing genotype spaces. The main restriction is that we want to maintain reversibility in the perturbed chain. We require this so that the perturbed chain is tractable to analysis and for all the reasons outlined in Section 2.3. We provide several different types of perturbations that maintain reversibility.

Even though the perturbations maintain reversibility other key properties of the chain might change. So the main task of this section is to identify how the perturbations change the stationary distribution and control their effect on the mixing time. Often the effect on the stationary distribution can be seen directly from the change in the reversibility condition (2.3.1). To control the mixing time, we have two primary approaches: 1) we can analyze the mixing time of the perturbed chain directly or 2) we can use Theorem 2.23.

Some of the perturbations we consider are deterministic and we simply make assumptions on their magnitude. In other cases, we consider random perturbations to get a sense of the typical effects of noise. Random perturbations mean that we are dealing with two different sources of randomness; one from the perturbation (or environment) and one from the stochastic process itself (the mutation process). This setup is common in probability theory and there is a wide and rich literature on such problems [88–90].

The notation in this section is as follows: let $\tilde{\mathcal{M}}$ be a reversible mutation process on the genotype space Γ with stationary distribution $\tilde{\pi}$ and spectral gap $\tilde{\gamma}$, then we denote the mutation process obtained by perturbing $\tilde{\mathcal{M}}$ by $\mathcal{M}$ and its stationary distribution by π and spectral gap by γ.

2.5.1. Simple symmetric rescalings. A very simple type of perturbation that maintains reversibility is to symmetrically rescale the transition probabilities, that is, define

$$\mathcal{M}(\alpha,\beta) := p(\alpha,\beta)\tilde{\mathcal{M}}(\alpha,\beta) \tag{2.5.1}$$

such that $p(\alpha,\beta) = p(\beta,\alpha) > 0$ for all $\alpha,\beta \in \Gamma$. Obviously, the mutation kernel $\mathcal{M}$ still satisfies condition (2.3.1). However, after rescaling we cannot guarantee that $\sum_\beta \mathcal{M}(\alpha,\beta) = 1$ for all $\alpha \in \Gamma$ or that $\tilde{\mathcal{M}}(\alpha,\beta)p(\alpha,\beta) \in [0,1]$. There are two possibilities to fix this: (1) normalize each row by dividing each by a constant,

$$\mathcal{M}(\alpha,\beta) := \frac{p(\alpha,\beta)\tilde{\mathcal{M}}(\alpha,\beta)}{\sum_{\beta'} p(\alpha,\beta')\tilde{\mathcal{M}}(\alpha,\beta')}, \tag{2.5.2}$$

so that $\mathcal{M}(\alpha,\cdot)$ is automatically a well-defined probability distribution on Γ; (2) assume $p(\alpha,\beta) < 1/\tilde{\mathcal{M}}(\alpha,\beta)$ and $\sum_\beta^{(\alpha)} p(\alpha,\beta)\tilde{\mathcal{M}}(\alpha,\beta) \in (0,1]$, which is possible when $p(\alpha,\beta)$ is close enough to 1 for all $\alpha,\beta \in \Gamma$. Then define the off-diagonal terms as in (2.5.1) and define the diagonal terms as

$$\mathcal{M}(\alpha,\alpha) := 1 - \sum_{\beta\in\Gamma}^{(\alpha)} p(\alpha,\beta)\tilde{\mathcal{M}}(\alpha,\beta). \tag{2.5.3}$$

Note that option (1) changes the stationary distribution to

$$\pi(\alpha) = \frac{\tilde{\pi}(\alpha)\sum_\beta p(\alpha,\beta)\tilde{\mathcal{M}}(\alpha,\beta)}{D_p}, \tag{2.5.4}$$

where $D_p := \sum_{\alpha,\beta} \tilde{\pi}(\alpha)p(\alpha,\beta)\tilde{\mathcal{M}}(\alpha,\beta)$. Whereas option (2) retains the same stationary distribution, $\pi = \tilde{\pi}$. The difference between π and $\tilde{\pi}$ in (2.5.4) is easily controlled:

$$\begin{aligned}
|\tilde{\pi}(\alpha) - \pi(\alpha)| &= \pi(\alpha)\left|\sum_\beta \tilde{\mathcal{M}}(\alpha,\beta)\left(1 - \frac{p(\alpha,\beta)}{D_p}\right)\right| \\
&\leqslant \sqrt{\sum_\beta \left|\tilde{\mathcal{M}}(\alpha,\beta)\right|^2 \sum_\beta \left|1 - \frac{p(\alpha,\beta)}{D_p}\right|^2} \\
&\leqslant \sqrt{\sum_\beta \tilde{\mathcal{M}}(\alpha,\beta) \sum_\beta \left|1 - \frac{p(\alpha,\beta)}{D_p}\right|^2} \\
&\leqslant \sqrt{\sum_\beta \left|1 - \frac{p(\alpha,\beta)}{D_p}\right|^2}
\end{aligned} \tag{2.5.5}$$

by the Cauchy-Schwarz inequality. Note that this bound can be improved when $\tilde{\mathcal{M}}(\alpha,\beta)$ is close to constant.

To control the effect on the mixing times, we use Theorem 2.23. Note that this is overkill in this situation—a simple comparison between the Markov chains using Dirichlet forms would suffice (see for example Lemma 13.22 in [67]). Since the perturbation does not alter which transitions are possible, we define paths as $\phi_{\alpha,\beta} = (\alpha,\beta)$. Thus, the congestion ratio (2.4.10) is

$$\max_{\alpha,\beta:\mathcal{M}(\alpha,\beta)>0} \frac{\tilde{\pi}(\alpha)\tilde{\mathcal{M}}(\alpha,\beta)}{\pi(\alpha)\mathcal{M}(\alpha,\beta)}. \tag{2.5.6}$$

For option (1), we see

$$\max_{\alpha,\beta:\mathcal{M}(\alpha,\beta)>0} \frac{D_p}{\sum_\beta p(\alpha,\beta)\tilde{\mathcal{M}}(\alpha,\beta)} \frac{\sum_{\beta'} p(\alpha,\beta')\tilde{\mathcal{M}}(\alpha,\beta')}{p(\alpha,\beta)} = \max_{\alpha,\beta:\mathcal{M}(\alpha,\beta)>0} \frac{D_p}{p(\alpha,\beta)}. \tag{2.5.7}$$

In particular, if $\max_{\alpha,\beta}|1-p(\alpha,\beta)| \leqslant \varepsilon$, then

$$|1-D_p| \leqslant \sum_\alpha \pi(a) \sum_\beta \mathcal{M}(\alpha,\beta)\,|1-p(\alpha,\beta)| \leqslant \varepsilon, \tag{2.5.8}$$

so (2.5.7) is bounded by $1+\mathcal{O}(\varepsilon)$. In which case, $\gamma \geqslant (1+\mathcal{O}(\varepsilon))\tilde{\gamma}$.

A very similar argument can be used to control the mixing time for option (2).

2.5.2. Simple nonsymmetric rescalings. This perturbation rescales columns of the mutation kernel and redefines the diagonal elements to ensure $\mathcal{M}(\alpha)$ is a well-defined probability distribution for all $\alpha \in \Gamma$. Precisely, suppose $p(\beta) > 0$ and $\sum_\alpha \tilde{\mathcal{M}}(\alpha,\beta)p(\beta) \leqslant 1$, then define

$$\mathcal{M}(\alpha,\beta) := \tilde{\mathcal{M}}(\alpha,\beta)p(\beta) \tag{2.5.9}$$

and

$$\mathcal{M}(\alpha,\alpha) := 1 - \sum_\beta \tilde{\mathcal{M}}(\alpha,\beta)p(\beta) \tag{2.5.10}$$

for all $\alpha,\beta \in \Gamma$. Then condition (2.3.1) is satisfied with

$$\pi(\alpha) = \frac{\tilde{\pi}(\alpha)p(\alpha)}{\sum_\beta \tilde{\pi}(\beta)p(\beta)}, \tag{2.5.11}$$

since

$$\pi(\alpha)\mathcal{M}(\alpha,\beta) = \frac{\tilde{\pi}(\alpha)p(\alpha)}{\sum_{\beta'} p(\beta')\tilde{\pi}(\beta')}\tilde{\mathcal{M}}(\alpha,\beta)p(\beta) = \pi(\beta)\mathcal{M}(\beta,\alpha). \tag{2.5.12}$$

Now note that, similarly to (2.5.5), we have

$$\left|1 - \sum_{\beta} \tilde{\pi}(\beta)p(\beta)\right| = \left|\sum_{\beta} \tilde{\pi}(\beta)\left(1 - p(\beta)\right)\right| \leqslant \sqrt{\sum_{\beta} |1 - p(\beta)|^2}. \tag{2.5.13}$$

We can control the distance between π and $\tilde{\pi}$ similarly to before:

$$|\tilde{\pi}(\alpha) - \pi(\alpha)| = |1 - p(\alpha)| + \mathcal{O}\left(\sqrt{\sum_{\beta} |1 - p(\beta)|^2}\right). \tag{2.5.14}$$

For the spectral gap, we again use the Comparison Theorem and note the congestion ratio is

$$\max_{\alpha,\beta:\mathcal{M}(\alpha,\beta)>0} \frac{p(\alpha)p(\beta)}{\sum_{\beta'} \pi(\beta')p(\beta')}, \tag{2.5.15}$$

so if we assume $\max_{\alpha}|1 - p(\alpha)| \leqslant \varepsilon$, then $\gamma \geqslant (1 + \mathcal{O}(\varepsilon))\tilde{\gamma}$.

2.5.3. Convex combinations of chains. Suppose that $\mathcal{M}$ and $\tilde{\mathcal{M}}$ are both reversible with respect to π, then so is any convex combination of the two:

$$\mathcal{M}_p(\alpha,\beta) := p\mathcal{M}(\alpha,\beta) + (1-p)\tilde{\mathcal{M}}(\alpha,\beta) \tag{2.5.16}$$

for $p \in [0,1]$. Obviously, $(\Gamma, \mathcal{M}_p)$ is a well-defined mutation process with stationary distribution π by Equation (2.5.16). For the mixing time, we again use the congestion ratio (2.4.10) and compare to both chains $\mathcal{M}$ and $\tilde{\mathcal{M}}$. Note

$$\frac{\mathcal{M}(\alpha,\beta)}{p\mathcal{M}(\alpha,\beta) + (1-p)\tilde{\mathcal{M}}(\alpha,\beta)} \leqslant \frac{1}{p} \tag{2.5.17}$$

and similarly

$$\frac{\tilde{\mathcal{M}}(\alpha,\beta)}{p\mathcal{M}(\alpha,\beta) + (1-p)\tilde{\mathcal{M}}(\alpha,\beta)} \leqslant \frac{1}{1-p}, \tag{2.5.18}$$

thus $\gamma_p \geqslant \max\{p\gamma, (1-p)\tilde{\gamma}\}$.

2.5.4. Bond percolation. In Section 2.3, we observed that random walks on weighted graphs are always reversible.

Starting with a random walk on a weighted graph there is a natural way to perturb the mutation process—we simply perturb the weights of the graph. Of particular interest is the case where edges are deleted from the graph (or equivalently their weights are set to 0). Biologically, this means that mutations that were previously viable are no longer an option. The edges can be deleted in many different ways, but in this subsection we focus on the case where edges are deleted independently with probability $1 - p$. Importantly, the edges in the original graph are undirected, so when a mutation becomes inviable in one direction, it automatically does so in the other direction.

We choose to delete edges independently at random for several reasons. First, without a specific model in mind, choosing the simplest type of randomness is often reasonable [91]. Having said that, the statistical properties of the genotype spaces obtained with this type of randomness matches well with many observations from biological experiments [11,38]. Second, this type of randomness breaks the symmetry of the genotype spaces we have considered, which was our stated goal at the start of Section 2.5. Third, this choice of randomness is amenable to analytical analysis and puts us into contact with the extensive work on percolation theory.

Percolation adds an additional source of randomness to the mutation process. Obviously, one can ask questions that concern only the way that percolation changes the geometry of the graph or lattice. However, if we ask question about a mutation process on these random genotype spaces, we are immediately in the territory of random processes in random environments. Many dynamical question about the mutation process depend on the geometry of the genotype space (see Section 2.4), so we often have to study geometrical question about the random space.

The first technical issue with deleting edges is that doing so can disconnect the graph. Obviously, in such cases the mutation process is no longer irreducible. Thus, connectivity is an important concern for us. The probability that a graph is disconnected by random, independent edge deletions with probability $1-p$ was first studied on the complete graph by Erdős and Rényi [92], who found a sharp transition: if $p > (1+c)\log(n)/n$ for some small constant $c > 0$, the graph is connected with probability $1-o(1)$ as $n \to \infty$; if $p < (1-c)\log(n)/n$ for some small constant $c > 0$, the graph is disconnected with probability $1 - o(1)$ as $n \to \infty$. Thus, $\log(n)/n$ is a sharp threshold for connectedness, but more detailed information about the graph's geometry is known below this threshold. If $np \to c$ for some constant $c > 1$, then the graph has a unique giant component that contains $\mathcal{O}(n)$ vertices, whereas all other components contain at most $\mathcal{O}(\log n)$ vertices. If $np \to 1$, then the graph has some component that contains $\mathcal{O}(n^{2/3})$ vertices. Finally, if $np \to c$ for some constant $c < 1$, then the graph has no component that contains more than $\mathcal{O}(\log n)$ vertices. The existence of a giant component for some p that are below the sharp threshold for connectedness suggests that the dynamics of a random walk still warrant study if we simply restricted to this component.

The stationary distribution takes an especially simple form for random walks on graphs (see (2.3.5)):

$$\pi(\alpha) = \frac{d_\alpha}{2|E|} = \frac{d_\alpha}{\sum_\beta d_\beta}, \tag{2.5.19}$$

where d_α is the degree of vertex α and $|E|$ is the total number of edges in the graph. Since edges are deleted independently at random, the quantities d_α are a sum of independent random variables—this means we can get good control over them with concentration inequalities (see Appendix B), so long as the degrees are large before percolation. Obviously when considering the complete graph, each vertex has degree n before percolation (we include self-loops). This implies that after percolation each vertex has degree pn in expectation. Moreover, by a concentration argument, we can concluded that each vertex has degree $pn + \mathcal{O}(\sqrt{n})$ with very high probability for large p. Therefore, a simple union bound implies that all vertices have degree $pn + \mathcal{O}(\sqrt{n})$ with very high probability simultaneously. Therefore,

$$\pi(\alpha) = \frac{1}{n} + \mathcal{O}\left(\frac{1}{n^{3/2}}\right) = \tilde{\pi}(\alpha) + \mathcal{O}\left(\frac{1}{n^{3/2}}\right). \tag{2.5.20}$$

The mixing time of a random walk on the Erdős-Rényi random graph has also been studied. For comparison, we note that a random walk on the complete graph reaches it stationary distribution after a single step, and thus its mixing time is $\mathcal{O}(1)$. Many results on the Erdős-Rényi random graph model for large p are also demonstrated simultaneously for the random regular graph model, where a graph is sampled uniformly from all those that have constant degree d, since the degrees in the Erdős-Rényi model concentrate around pn. The upper bound lemma [66] and a combinatorial argument are used to show the mixing time is $\mathcal{O}(1)$ with high probability when the degrees are a fixed power of n in [93]. The results in [94] improve on the above combinatorial argument, and extends the mixing time result to graphs with degree that are $\mathcal{O}(\log n)^c$ for $c > 2$. For p below the sharp threshold, the graph is disconnected with high probability, so a random walk is no longer irreducible. However, if the walk is restricted to the unique giant component of the graph, [95] and [96] find using Cheeger type bounds that asymptotically almost surely the mixing time is $\Theta((\log n/d)^2)$ when $d = \mathcal{O}(\sqrt{\log n})$, and $\Theta(\log n/\log d)$ when $d \gg \sqrt{\log n}$. Interestingly when p is in the first regime, there is local structure of the graph that slows mixing, whereas when p is in the second regime the diameter of the giant component is the largest impediment to mixing. In [97], the connection probability is lowered all the way to the critical value $p = 1/n$. The authors find that a random walk on the largest connected component (which has diameter $\mathcal{O}(n^{1/3})$) has mixing time $\mathcal{O}(n)$ and so the walk no longer mixes rapidly. Finally, we note that [98] finds the mixing time of a random walk on a random regular graph displays a cutoff at $(d/(d-2)\log_{d-1} n$ with window order $\sqrt{\log n}$ and that when $d = n^o(1)$, the mixing time is $\mathcal{O}(1)$.

Turning our focus from the complete graph to the hypercube, we can ask similar questions to those we have discussed for the complete graph. Note that most results have studied case $\kappa = 2$, but results can easily be extended to general κ case. The hypercube has κ^n points and each point has degree $(\kappa-1)n$. We then apply a bond percolation where each edge is kept with probability p independently. Take a single point, then the probability that this point is isolated (that is, has no neighbors) is exactly

$$(1-p)^{(\kappa-1)n}. \tag{2.5.21}$$

Therefore, we can lower bound the probability that no point is isolated by

$$1-\kappa^n(1-p)^{(\kappa-1)n} = 1-\left(\kappa(1-p)^{\kappa-1}\right)^n. \tag{2.5.22}$$

Note that if

$$p > p_{\mathrm{c}} := 1-\kappa^{-1/(\kappa-1)}, \tag{2.5.23}$$

then with very high probability there are no isolated vertices. In fact, p_{c} is the critical probability for percolation above which the graph is connected with high probability [99].

We now state results for $\kappa = 2$. For p below this critical threshold there can still be a giant component. In [100], the authors find that for $p = c/n$ and $c > 1$, graph contains a component of size $\mathcal{O}(2^n)$ and that the second largest component is size $o(2^n)$. It is possible to find even more detailed information about the size of the connected subgraphs, and [101] studies this especially below the threshold $p = 1/n$. For low p most arguments use an approximation to a Poissonian branching process (see the proof of Theorem 3.6 for another example of such an approximation). For a survey of percolation results on the hypercube see [102] and for high-dimensional spaces in general see [103].

Formally, we arrive at the following mutation process.

Definition 2.26 (Single point mutation process on the bond disordered hypercube). *This process is a random Markov chain* $(\Gamma_n^{(p)}, \mathcal{M}^{(p)})$ *with state space* $\Gamma_n = [\![\kappa]\!]^n$ *and each* $\mathcal{M}^{(p)}(\boldsymbol{\alpha}, \boldsymbol{\beta})$ *a random variable defined as* *follows: for*

$$p \in \Big(1-\kappa^{-1/(\kappa-1)}, 1\Big], \tag{2.5.24}$$

and for $e \in E$ let x_e be i.i.d. Bern(p), where E is the edge-set of the hypercube Γ_n. Then

$$\mathcal{M}^{(p)}(\boldsymbol{\alpha},\boldsymbol{\beta}) := \begin{cases} \frac{\varepsilon x_e}{\kappa n} & \textit{if } e=(\boldsymbol{\alpha},\boldsymbol{\beta}) \\ 1-\sum_{\tilde{\boldsymbol{\beta}}:(\boldsymbol{\alpha},\tilde{\boldsymbol{\beta}})\in E}\mathcal{M}^{(p)}(\boldsymbol{\alpha},\tilde{\boldsymbol{\beta}}) & \textit{if } \boldsymbol{\alpha}=\boldsymbol{\beta} \\ 0 & \textit{otherwise} \end{cases}. \tag{2.5.25}$$

The interpretation of Definition 2.26 is that mutations are chosen according to Definition 2.4, but they only occur if they are viable according to the randomness. We have seen that with probability $1-o(1)$, the single point mutation process on the bond disordered hypercube is irreducible. It is also aperiodic, since $\varepsilon > 0$. Reversibility with respect to the uniform distribution is immediate from the symmetry of $\mathcal{M}^{(p)}$. The way we define $\mathcal{M}^{(p)}$ in Definition 2.26 makes the following analysis easier, because the stationary distribution remains uniform, but to see the similar results for a true lazy random walk on the bond percolated hypercube see Appendix C.

The mixing times of random walks on structured spaces after percolation has been less studied than the equivalent question for the Erdős-Rényi and random regular graph models. In [104], the largest cluster of a super critical bond percolation on the finite lattice $[\![-n,n]\!]^d$ (with $d>2$ and fixed) is studied. Using a generalized Cheeger type bound (called the isoperimetric profile or conductance profile and developed in [105]), they find the mixing time of a random walk is $\Theta(n^2)$—this is within a constant factor of the mixing time without percolation. However, for the hypercube, we wish to fix $n=1$ and take $d\to\infty$.

Now, we show that the mixing time of the single point mutation process on the bond disordered hypercube can be controlled for $p>p_{\mathrm{c}}$, that is, above the critical threshold for connectedness. We prove this using the Comparison Theorem 2.23. Since we are restricting $p>p_{\mathrm{c}}$, we know that with high probability the graph will be connected. This means that the Markov chain after percolation has the same state space and that it remains irreducible—both of which are necessary for Theorem 2.23 to apply. Moreover, we know the stationary distribution is uniform.

Theorem 2.27. *Let $(\Gamma_n^{(p)},\mathcal{M}^{(p)})$ be the single point mutation process on the bond disordered hypercube and denote* *its mixing time by* t_{mix}*. Suppose*

$$p > p_{c,3} := \sqrt[3]{1 - (1/\kappa)^{1/(\kappa-1)}} = \sqrt[3]{p_c} \geqslant p_c, \tag{2.5.26}$$

then the process mixes rapidly—specifically,

$$t_{mix} = \mathcal{O}\left(\frac{n^2 \log n}{\varepsilon}\right) \tag{2.5.27}$$

with probability $1 - o(1)$*.*

Proof. Our goal is to show that after deleting edges from the hypercube, there are still short (of length at most 3) paths between all points that were previously neighbors before deletions. Let E is the edge-set of the hypercube. Suppose for some edge in E, denoted $e = (\boldsymbol{\alpha}, \boldsymbol{\beta})$, that $x_e = 0$. Suppose also that the coordinate that $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ differ at is i with $\boldsymbol{\beta}(i) = k$; for notation, we write $\boldsymbol{\alpha}_k^i$ to signify $\boldsymbol{\alpha}$ has been changed in coordinate i to k, and thus $\boldsymbol{\beta} = \boldsymbol{\alpha}_k^i$ for some $k \in [\![\kappa]\!]$. Then a path of length 3 from $\boldsymbol{\alpha}$ to $\boldsymbol{\beta}$ in Γ would have the form

$$\left(\boldsymbol{\alpha}, \boldsymbol{\alpha}_{k'}^j, \left(\boldsymbol{\alpha}_{k'}^j\right)_k^i, \boldsymbol{\beta}\right) \tag{2.5.28}$$

for some $j \neq i$ and $k' \neq \boldsymbol{\alpha}(j)$, since $\boldsymbol{\beta} = \boldsymbol{\alpha}_k^i$. Such paths pick some other coordinate j not equal to i, change it, then change coordinate i to match $\boldsymbol{\beta}$, and finally change coordinate j back to its previous value. Note that once any edge in the path is determined, the path of this form is completely determined.

For any pair of $\boldsymbol{\alpha}, \boldsymbol{\beta}$ such that $\mathcal{D}(\boldsymbol{\alpha}, \boldsymbol{\beta}) = 1$, there are exactly $(\kappa - 1)(n - 1)$ paths of this form between them. Moreover, each of these paths shares no edges and thus the events

$$x_{\left(\boldsymbol{\alpha}, \boldsymbol{\alpha}_{k'}^j\right)} x_{\left(\boldsymbol{\alpha}_{k'}^j, \left(\boldsymbol{\alpha}_{k'}^j\right)_k^i\right)} x_{\left(\left(\boldsymbol{\alpha}_{k'}^j\right)_k^i, \boldsymbol{\beta}\right)} = 1 \tag{2.5.29}$$

for each path are independent for fixed $\boldsymbol{\alpha}, \boldsymbol{\beta}$. For fixed $\boldsymbol{\alpha}, \boldsymbol{\beta}$, the probability that a single path exists is p^3, thus the probability that at no path exists is

$$(1 - p^3)^{(\kappa-1)(n-1)}. \tag{2.5.30}$$

Therefore, the probability that all pairs $\boldsymbol{\alpha}, \boldsymbol{\beta}$ such that $(\boldsymbol{\alpha}, \boldsymbol{\beta}) \in E$ have at least one path of length 3 between them is lower bounded by

$$1 - n(\kappa - 1)\kappa^n (1 - p^3)^{(\kappa-1)(n-1)}, \tag{2.5.31}$$

as there are $n(\kappa - 1)\kappa^n$ such pairs $\boldsymbol{\alpha}, \boldsymbol{\beta}$. Note that the probability (2.5.31) is very high given that

$$p_{c,3} := \sqrt[3]{1 - (1/\kappa)^{1/(\kappa-1)}} < p. \tag{2.5.32}$$

So for $p > p_{c,3}$, there is some path with probability $1 - o(1)$ between all pairs $\boldsymbol{\alpha}, \boldsymbol{\beta}$ such that $(\boldsymbol{\alpha}, \boldsymbol{\beta}) \in E$. Let $\mathcal{E}$ denote the event that there is a path of length at most three between all pairs $\boldsymbol{\alpha}, \boldsymbol{\beta}$ such that $(\boldsymbol{\alpha}, \boldsymbol{\beta}) \in E$. Above we have argued that $\mathbb{P}\{\mathcal{E}\} \geqslant 1 - o(1)$.

Now, we use the Comparison Theorem 2.23. We construct paths as follows conditional on the event $\mathcal{E}$: For any pair $\boldsymbol{\alpha}, \boldsymbol{\beta}$ such that $(\boldsymbol{\alpha}, \boldsymbol{\beta}) \in E$ where $x_{(\boldsymbol{\alpha},\boldsymbol{\beta})} = 1$, we set $\phi_{\boldsymbol{\alpha},\boldsymbol{\beta}} = (\boldsymbol{\alpha}, \boldsymbol{\beta})$. For any pair $\boldsymbol{\alpha}, \boldsymbol{\beta}$ such that $(\boldsymbol{\alpha}, \boldsymbol{\beta}) \in E$ where $x_{(\boldsymbol{\alpha},\boldsymbol{\beta})} = 0$, we choose $\phi_{\boldsymbol{\alpha},\boldsymbol{\beta}}$ arbitrarily from paths of the form (2.5.28), of which one is guaranteed to exist on the event $\mathcal{E}$.

Consider an arbitrary edge e in E, there are $3(\kappa - 1)(n - 1)$ paths of the form (2.5.28) that pass through e. To see this note that e can either form the beginning, middle, or end of a path, and after e is fixed there are exactly $(\kappa - 1)(n - 1)$ paths of this form. Alternatively, we may use the symmetry of the hypercube to note that the answer should not depended on the choice of e. Then a path of the form (2.5.28) is determined by its end points, of which there are $(\kappa - 1)n\kappa^n$ possible choices, and then there are $(\kappa - 1)(n - 1)$ paths for each pair of end points. Moreover, each path contains 3 edges, so the each edge must be included in

$$\frac{3(\kappa - 1)n(n - 1)\kappa^n}{(\kappa - 1)n\kappa^n} = 3(n - 1)(\kappa - 1) \tag{2.5.33}$$

paths.

Let $E' := \{(\boldsymbol{\alpha}, \boldsymbol{\beta}) : x_{(\boldsymbol{\alpha},\boldsymbol{\beta})} = 1\}$. Now we can bound B from Equation (2.4.10) as follows:

$$\max_{(\alpha,\beta)\in E'} \left(\frac{\kappa^n \kappa n}{\varepsilon} \sum_{\tilde{\alpha},\tilde{\beta}:(\alpha,\beta)\in\phi_{\tilde{\alpha}\tilde{\beta}}} \frac{\varepsilon}{\kappa^n \kappa n} \cdot 3 \right) \leqslant 9(\kappa - 1)(n - 1). \tag{2.5.34}$$

Then applying Theorem 2.23, Lemma 2.11, and Theorem 2.21 completes the proof. ■

Remark 2.28. While we have only studied the mixing properties of the hypercube after bond percolation in the super-critical regime, there are conjectures about these properties down to $p > 1/n$ for $\kappa = 2$. If the random walk on the percolated hypercube is restricted to the unique giant component (which exists so long as $p > 1/n$), the mixing time is conjectured to be polynomial in n with high probability [106]. Perhaps the mixing time should even be $\mathcal{O}(n^2)$, as this is what one expects from the Erdős-Rényi case.

Note that the argument in the proof of Theorem 2.27 can be generalized to bond percolation in other high-dimensional spaces. The main requirement is that there are many disjoint short paths between all neighbors in the original space.

2.5.5. Site percolation. In Subsection 2.5.4 we added disorder to the hypercube by making some mutations inviable by symmetrically deleting edges from the hypercube. We saw that the changes this caused in the stationary distribution and mixing time could be controlled for moderate levels of disorder. In this subsection, we consider a different type of disorder with a different biological interpretation. Again, we start with a random walk on a weighted graph, but in bond percolation instead of deleting edges independently with probability $1 - p$, we delete vertices of the graph. When a vertex is deleted all edges incident to this vertex are also deleted. This setting has an interesting biological interpretation that we discuss more in Section 2.6, but for now we imagine deleted vertices represent genotypes that are inviable. While there is experiment and theoretical work studying how genotypes might be segregated in this way [107–111], we lack a complete understanding of the structure of this distinction. However, some of the statistical information, we do have about the geometry of the subset of viable genotypes in genotype space can be replicated with this simple random model, but more complex models are required to match all the experimental data [107]. Thus, we might hope that even this simple case can inform us about evolution in real genotype spaces.

This idea of separating genotypes into viable and inviable subsets goes back at least to [112]. Choosing the viable and inviable subsets uniformly and independently at random was considered in [113] for $\kappa = 2$. As we did in Subsection 2.5.4, [113] uses results from percolation theory to get geometric information about the subset of viable genotypes. Similarly to our calculation in Equation (2.5.23), one can show that the site percolated hypercube is connected with high probability if $p > 1/2$. While the site percolated hypercube is disconnected with high probability when $p < 1/2$, there is still a unique giant component of size $2^n p$ for $p > 1/n$. The second largest component has size at most $\mathcal{O}(n)$. For $p < 1/n$, there are many components of size at most $\mathcal{O}(n)$. Interestingly, [113] also finds a distinction in the number of evolutionary paths between genotypes for p above and below $1/n$. As we discussed in Section 2.4, the existence of many paths between pairs of points in a space is a feature of high-dimensional spaces and can lead to rapid mixing (see Theorem 2.23). All these results are summarized in the review article [114], although there is a focus on how this may be a mechanism for speciation.

In fact, motivated by spin glasses, the relaxation time (which is related to the mixing time) of a random walk on a site percolated hypercube has been studied both experimentally and theoretically [115]. It is found that the relaxation time is exponential when $p > 1/2$ and that the behavior shifts to a stretched exponential with exponent $1/3$ when $1/n < p < 1/2$. Both regimes suggest that the mixing time of a random walk should be fast (although faster when $p > 1/2$) because of the high-dimensionality of space.

While these results are suggestive, this model of disorder presents additional technical challenges as the state space of the mutation process changes after percolation. Many of the techniques for Markov chains that control the effects of perturbation, assume the perturbed chain has the same state space (see Theorem 2.23 for example).

Note that the stationary distribution of a random walk on the site percolated hypercube can be controlled in the same way as we did in Equation (C.0.3).

2.5.6. Other subgraphs of the hypercube. In this subsection, we highlight some very interesting work that studies the mixing time of a random walk on a subset of the hypercube. In the previous subsection, we did exactly that, but for a random subset of the hypercube. Our biological motivation for this was to point out that some genotypes are inviable and that choosing the viable genotypes uniformly and independently at random matched the statistical geometric properties from experiments. However, we could instead have used some deterministic criteria to choose these viable genotypes. This would then specify the subset of the hypercube our random walk would walk on. For example, [116] use a model for how RNA sequences fold to produce secondary structure, then an RNA sequence is considered viable if it has a specific secondary structure. Note that this criterion is deterministic.

It is not clear in general what types of criteria are biological realistic, but [117] gives an interesting example from computer science. The knapsack problem is a classical problem in combinatorial optimization [118]. Imagine you have a knapsack whose capacity is limited by a maximum weight W. Then, given a list of n items each with a weight w_i and value v_i, find the combination of items that can fit in the knapsack that has maximum value. Mathematically, the maximal subset $M \subseteq [\![n]\!]$ is defined

$$M := \operatorname{argmax}_{S\subseteq[\![n]\!]} \left\{ \sum_{i\in S} v_i : \sum_{i\in S} w_i \leqslant W \right\}. \tag{2.5.35}$$

Obviously, any subset S can be represented by the point $(\mathbf{1}(1 \in S), \ldots, \mathbf{1}(n \in S))$ on the hypercube, and we can consider the subset of the hypercube containing all points that do not violate the weight constraint.

While the knapsack problem is computationally difficult, an efficient randomized algorithm is found using a simple Metropolis chain [67]. Metropolis chains take a current solution as their current state, then randomly sample possible neighboring solutions that add or remove one item from the knapsack, and then they transition to this neighboring state if it does not violate the weight constraint. Because any solution can be transformed into any other solution by adding and removing one item at a time, and because the Metropolis chain is symmetric, the stationary distribution is uniform on the set of possible solutions to the knapsack problem. Thus, we can sample approximately uniformly for the set of possible solutions to the knapsack problem. The efficiency of the randomized algorithm is then linked to how quickly the chain mixes. This motivates Morris and Sinclair's study of the mixing time [117]. Significantly, the mixing time is found to be $\mathcal{O}(n^{9/2+\varepsilon})$ for any $\varepsilon > 0$, which means the chain mixes rapidly.

This bound on the mixing time extends to any hyperplane on the hypercube. So any biological criteria that can be described by a hyperplane on the hypercube produce a set of viable genotypes that have rapid mixing. For example, a hyperplane is formed if we imagine the coordinates of the hypercube representing the presents of a gene, each of which increase fitness additively, and specify that a genotype is viable if its fitness exceeds a given value.

2.6. Fitness

The purpose of this section is to discuss how to assign a nonnegative real number, called fitness, to each genotype in a genotype space [119]. This idea goes back at least to Wright [120, 121]. We pay special attention here to the case where fitness take values in $\{0, 1\}$ to distinguish between viable and inviable genotypes (or neutral regions in genotype space) [112].

We are impelled to introduce fitnesses, as these values are the avenue by which selection enters into models of evolutionary dynamics and so are necessary to develop the models in Chapters 3 and 5. Attempting to make sense of fitness biologically and philosophically is challenging, and we discuss some of these challenges here [23]. However, thinking about the notion of fitness in biological examples serves to motivate and form it as a parameter in our models. Indeed, mathematical models provide the clearest way to *define* a coherent notion of fitness.

In the introduction of this section, we vaguely mentioned that fitness describes the average rate at which a genotype is replicated in the population of genotypes that our evolutionary process is tracking. What determines this rate? So far we have only considered genotypes syntactically—labeling them and discussing how they relate to each other geometrically through mutation. To discern more about fitness, we must consider genotypes semantically. We must think carefully about the mechanisms that actually replicate the genotypes.

Consider the following simple experiment that we use to work through the idea of fitness as average rate of replication. Imagine a Petri dish full of bacteria and some media containing enough nutrients for the bacteria to survive. In this experiment, we tract the genotypes of all the bacteria in the dish. When a bacterium undergoes cell division it makes a copy of its DNA. Working backwards from this event, we see the mechanism for this copying is determined by the bacterium's phenotype, that is, all of its physiological, behavioral, and biochemical properties and traits. Think of how cell devision is initiated by a complex interaction of regulator networks, or consider the physical manner in which DNA stores the genotype and provides a specific mechanism for its replication via its double helical structure. Both are part of the phenotype. We might even *define* the phenotype as the mechanism that reproduces the genotype. This mechanism might do many other things that are seemingly unrelated to replication, but from the perspective of modeling evolution they are irrelevant. All we are concerned with is how the phenotype causes the reproduction of its genotype.

So the phenotype does the copying, but the genotype provides the heritable instructions for the phenotype. Returning to our example, the bacteria need energy to do work and they get this energy by liberating it from the media with chemical reactions. These chemical reactions are catalyzed by proteins that are translated, via RNA, from the bacterial DNA (as explained in the central dogma of molecular biology). Through a multitude of interactions of this type, the genotype influences its own replication, but fitness is only a function of genotype through the phenotype and not directly.

Despite the genotype providing instructions for the phenotype, it does not completely determine it. The phenotype is also influenced by the environment. Thus, organisms with identical genotypes can have different phenotypes, either due to variation in the environment or noise in how the genotype is translated into the phenotype. For example, the proteins that our bacteria produce may fold differently as the temperature of the media changes. Taking this to an extreme, if the temperature is high enough, no bacteria can survive let alone replicate. So fitness also depends on the environmental context. Does this mean we are assuming variables in the environment like temperature are constant over the timescales in our models? Or, as we are talking about *average* fitness, are we able to average over changes in the environment over time? However, the environment varies even on microscopic scales. The bacteria at the edge of a growing colony might replicate faster because of increased availability of nutrients and better access to space for division. Again, are we averaging over these spatial variations when we describe fitness? If we consider the composition of the population as part of the environment, then it is easy to imagine that this might also affect fitness. Say some bacteria have a mutation that means they can metabolize a waste product from the metabolic cycle of the wild-type bacteria. In this case, the abundance of the wild-type bacteria could increase the fitness of the mutated bacteria. Is this variation in frequency something that can be accounted for by averaging too? Specifically, the question is can the phenomena we want to describe be captured in a model where fitness does not depend on such variations or bundles them together in some complex way? The existence of models with changing environments [122], structured populations (see Chapter 6) [123,124], and frequency dependent selection [125] suggest that there are important phenomena that are not.

The problems we have described above relate to what fitness can and cannot incorporate, and how, as its purview expands, it becomes more obscure and less interpretable—almost to the point of meaninglessness or tautology. By all accounts, fitness seems undetermined by the genotype, but perhaps it can be salvaged as a relative rather than an absolute concept. In models that consider competition between two genotypes that differ in some small way and are otherwise equal (see Moran process in Section 3.1 and Wright-Fisher process in Section 3.2), fitness intuitively seems on firmer ground. In these models, we do not think of this difference in genotype as radically affecting the mechanism of replication, only its rate. We assume that this difference in rate is consistent over all the changes in environment that we outlined above, and because the change in genotype is small, our *ceteris paribus* assumption seems reasonable. However, for models that try to capture longer timescales, where evolution is exploring a vast genotype space and our aim is to talk about major evolutionary developments in the phenotype (see origin-fixation models in Chapter 5), the *ceteris paribus* assumption cannot be justified by the same argument.

These problems have lead some to suggest the notion of fitness must be revised [126, 127]. However, it is difficult to know how it can be replaced in our models of evolutionary dynamics. Even models that address how fitness depends on its environmental context still have some parameter that controls the rate of replication in some way.

The alternative to fitness, of modeling the whole mechanism of reproduction, seems unlikely to lead to tractable models. Perhaps fitness should be thought of as a kind of emergenic property—analogously to how atoms do not on their own have the properties of fluids, but how their behavior is well predicted in the aggregate in certain situations by equations that contain terms like viscosity. Ultimately, the role of fitness in these models will be justified by what they tell us about evolution.

Our discussion has made the assumptions that go into models of evolutionary dynamics more apparent and also lead us to some important constraints on fitness. First, we know that fitness depends on the genotype via the phenotype. Second, we saw that it is easier to understand fitness as a relatively term that compares differences in the rate of replication. In this way, we discuss the effect of a mutation on fitness. These two observations are especially important for neutral evolution. We can argue that if genotypic changes do to change the phenotype, those variants should have the same fitness. Many models that give some semantic interpretation to a genotype rely on this argument, which proves crucial to saying anything sensible about evolution on long timescales. We discuss some of these models now.

2.6.1. Models of fitness via phenotypic functionality. To start with a very simple example, consider a subsequence of the genome that is noncoding, that is, a sequence of DNA that is never transcribed into RNA (and thus not translated into protein). Important examples of such subsequences include pseudogenes. Since these subsequences are not transcribed, it is reasonable to suppose that most mutations in these subsequences to not lead to changes in fitness. In this way, we find a biological motivation for a completely neutral genotype space that contains all the potential configurations of this noncoding subsequence of DNA.

In Section 2.2, we already considered a second simple example in this vein of neutral genotype spaces. In this example, a genotype described the order of a list of genes on a chromosome. Again, it is reasonable to assume that reordering these genes (but maintaining their content) should have no or very little affect on fitness.

As a third example consider a gene that has a duplicate copy, so that the second copy is essentially redundant [14,128,129]. Again whatever the effect the first copy of the gene has on fitness can be maintain while the second copy is subject to mutation. Considering the second copy as the genotype is a further way to motive a neutral landscape.

These three examples are fairly trivial, but there are several examples that use more careful thought about how the genotype is translated into the phenotype. The main assumption in these examples is that fitness, which is difficult to measure or define, is related to some concrete, measurable function, like catalyzing a specified chemical reaction or binding some molecule. While there may not be a direct interpretable relationship between functionality and fitness, we can at least make a distinction between phenotypes that have that functionality to some degree and those that do not.

In our first example of this type, consider a gene that encodes for a sequence of amino acids that folds into a protein that performs a particular function in the cell. That protein's function (or phenotype) is determined by its physical and chemical properties. These properties are separated into three levels: primary, secondary, and tertiary, where the former determines the subsequent level. Primary structure consists of the sequence of amino acids. Secondary structure consists of the local structures of the folded protein, like alpha helices and beta sheets. Tertiary structure is the full three-dimensional shape of the folded protein.

So any mutations that maintain these properties should be neutral. Molecular biology tells us that once DNA is transcribed to RNA, it is translated into a string of amino acids three base pairs at a time. Each of these three base pairs is called a codon and there are $4^3 = 64$ possible combinations, but since there are only 20 amino acids, there are some codons that specify the same amino acids. In this way, we can argue that mutations that change one codon to a synonymous one should be neutral, as they maintain the primary structure of the protein and thus also the high level properties [130–133]. We can take this approach a step further, and ask what is the set of genotypes that translate into protein with a specified secondary structure? How abundant are the different secondary structures in genotype space? What is the geometry of this set?

Due to the massive size of sequence space, studying these questions empirically is challenging. However, experiments have shown that proteins that fold (rather than aggregating), show helicity, or are soluble are abundant in sequence space [111, 134, 135]. Miraculously, even ATP-binding proteins can be found by sampling from a random-sequence library [136]. This must indicate that the density of such proteins is high, as if the density were low, an experimentally infeasible number of random sample would be required to find one. Indeed, some experiments do not start with unbiased random proteins, but instead start with a functioning protein and use targeted mutagenesis to explore how functionality changes in neighboring proteins [137, 138].

Due to the obvious empirical challenges in dealing with such large spaces, finding good physical models for predicting protein secondary structure has been a priority. Simple models of protein folding have been used to study the geometric properties of the set of amino acids that fold into a specific structure [139]. Under these models, the set is found to spread throughout sequence space (not cluster together) with no obvious sequence homology between pairs in the set. However, despite being spread throughout sequence space, the subset forms a *neutral net*—that is, for ever pair of sequences there is a path of mutations between them that does not leave the subset. Moreover, none of these observations appear to depend on the specific folding structure under consideration.

More complex models of protein folding have shown great progress by exploiting multiple sequence alignment and modern machine learning techniques, with current models reaching over 90% accuracy in predicting secondary structure [140–144]. Determining the full tertiary structure of proteins from sequences is a notoriously difficult problem, but state of the art algorithms have shown promise [145–147]. The computational resources required for these models often make the deep exploration of sequence space that is required to determine geometrical properties difficult.

Similar work has been done on RNA, which instead of simply encoding for proteins, can have functional significance itself. Similarly, it can be argued that this functionality should be mediated by the secondary structure of the folded RNA. Shuster and Fontana's work on this topic is extensive [107,148–151]. They have produced mathematical models of RNA secondary structure, which are less computational intensive than those for proteins, and used them to study how sequences yielding specific secondary structures are distributed throughout sequence space. From these models, we can get information about the geometry and abundance of secondary structures in sequence space. Two significant findings are: (1) that the abundance of secondary structures shows a power-law decay: a few common structures are very abundant, and most structures are quite rare; (2) the geometry of the subset of sequences with a specific secondary structure is again described by a neutral net—each sequence in this subset has neighbors with the same secondary structure and set is connected together by single point mutations [152].

Metabolic reaction networks provide a third way to model the relationship between the genotype, phenotype, and fitness. The genotype for a metabolic reaction network is a binary sequence, where each bit specifies the presence or absence of a catalysis for a specific chemical reaction [11]. When a catalyst is present in the genome, we can say that reaction is available to the phenotype. We can imagine all the molecules that can be produced from some starting substrates, like glucose or fructose, by chaining together available reactions. Then, for each genotype, we can decide whether it produces all necessary biomolecules and is thus a viable genotype. In this way, we can separate the genotype space into two disjoint subsets of viable and inviable genotypes for each starting substrate. The viability of a metabolic reaction networks on a substrate is its phenotype. Wagner has explored this model with a random walk to neighboring, viable sequences and discovered many intriguing properties [109, 110, 153, 154]. They find that most reactions can be substituted, very different genotypes can be viable on the same substrate, and that the mutational neighborhood of most sequences contains many different phenotypes. The geometry of the subset of viable sequences is also fascinating: it has the same neutral net structure as described above for RNA, the mean and maximum distance of pairs of sequences from the space are close to those statistics for points from the whole space, and the viable subset is spread throughout sequence space—not clustered together.

Taken together, all of these examples of the relationship between genotype and phenotype inform our expectations of the function $\mathcal{F}$ and how we should introduce it into our models. In particular, we have highlighted some observations about the geometry of neutral regions of genotype space—or at least regions where fitness is only varying weakly. We now turn to some statistical models that try to replicate some of these observed properties without relying on complex ways of determining fitness or viability by modeling the phenotype [155].

2.6.2. Statistical models of fitness. In this subsection, we briefly highlight some random models for the function $\mathcal{F}$ that have been developed for $\Gamma = [\![\kappa]\!]^n$ [32, 119]. Often the purpose of randomizing the function $\mathcal{F}$, is that is allows us to discover properties of evolution that hold in general or are "typical." The NK model is probably the most well known, which provides a way of parameterizing the ruggedness and correlations in $\mathcal{F}$ [46]. Briefly, the fitness of a sequence $\boldsymbol{\alpha}$ is given by the sum of the fitness of each of its coordinates

$$\mathcal{F}(\boldsymbol{\alpha}) := \sum_{i=1}^{n} f_i(\boldsymbol{\alpha}(i); \boldsymbol{\alpha}(i_1), \ldots, \boldsymbol{\alpha}(i_K)), \tag{2.6.1}$$

where the values of the function f_i are sampled independently from some distribution. Note that f_i takes K other coordinates of $\boldsymbol{\alpha}$ as an arguments. By increasing K, the ruggedness of the landscape can be increased. The NK model has proved particularly important in the case of strong selection and large fitness differences between genotypes, which contrasts with the biological motivations we described in Subsection 2.6.1. In this regime, evolution takes on a more deterministic character called adaptive walks, where fitness increases at each step. So the number and basin of attraction of the local optima of $\mathcal{F}$ are a particularly important object of study.

In contrast, the theory of holey landscapes provides a model for neutral regions of fitness [113, 114]. In this model,

$$\mathcal{F}(\boldsymbol{\alpha}) \sim \mathrm{Bern}(p) \tag{2.6.2}$$

independently at random. Note this is exactly the method we describe to disorder a genotype space in Subsection

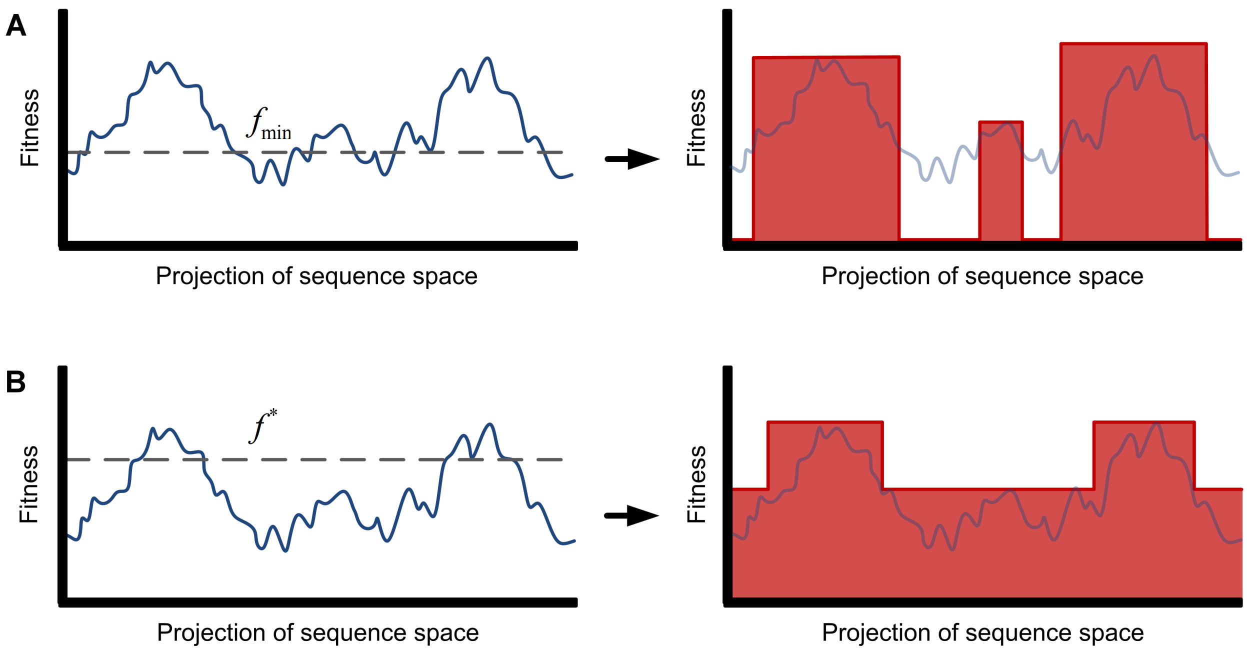

Figure 2.1: Mostly neutral landscapes by truncating fitness.

2.5.5. We can also interpret this landscape as a truncation of the NK model when $N = K$: let $\tilde{\mathcal{F}}$ be given by the NK model, then

$$\mathcal{F}(\boldsymbol{\alpha}) = \mathbf{1}\left(\tilde{\mathcal{F}}(\boldsymbol{\alpha}) \geqslant c\right) \tag{2.6.3}$$

for some $c > 0$. Despite its simplicity, this model can replicate many of the statistical and geometric properties we outlined above for fitness landscapes that derived from some kind of functionality. Specifically, we can reproduce connected, neutral nets. However, not all properties can be reproduce in this way [107]. In particular, fitnesses are uncorrelated in holey landscapes, which does not agree with our knowledge of biology. We can still hope to learn something from these landscapes, as often independence assumption are necessary for analysis and give the same results as models with weak sources of correlation due to some universality [156–158]. In Chapter 5, we use a related model to study evolution on long timescales. Note that dynamics on these holey landscape have been considered before in [159–161].

2.7. One-dimensional projections

While the hypercube has many nice properties, its high-dimensionality can make it difficult to analyze in some cases. Moreover, for some questions, it is not important to know the complete genotype sequence, but only a summary statistic of it. For example, say we are interested in how long it take to reach a particular genotype in the space in the single point mutation process (denoted by $\boldsymbol{\alpha}_t$). Mathematically, we want to get information about the random variable $T := \min\{t : \boldsymbol{\alpha}_t = \boldsymbol{\beta}\}$ such that $\boldsymbol{\alpha}_0 = \boldsymbol{\alpha}$ for $\boldsymbol{\alpha}, \boldsymbol{\beta} \in \Gamma$. Then we can consider the process $x_t := \mathcal{D}(\boldsymbol{\alpha}_t, \boldsymbol{\beta})$.

The process x_t is a projection of the process $\boldsymbol{\alpha}_t$ and, significantly, the process x_t is still a Markov chain. The transition probabilities of the process are given by

$$\mathbb{P}\{x_{t+1} = x_t + 1 | x_t\} = \frac{\varepsilon (n - x_t)(\kappa - 1)}{\kappa n}, \quad \mathbb{P}\{x_{t+1} = x_t - 1 | x_t\} = \frac{\varepsilon x_t}{\kappa n}, \tag{2.7.1}$$

and $\mathbb{P}\{x_{t+1} = x_t | x_t\} = 1 - \mathbb{P}\{x_{t+1} = x_t + 1 | x_t\} - \mathbb{P}\{x_{t+1} = x_t - 1 | x_t\}$. The reason the process x_t is easier to analyze is that it is a birth-death chain. Since the process is a birth-death chain, there are closed formulae for statistics like the $\mathbb{E}\tilde{T}$ (where $\tilde{T} := \min\{t : x_t = 0\}$ is simply T projected and defined for the process x_t) and the stationary distribution of the process (See Appendix A).

Using Theorem A.2, we can calculate $\mathbb{E}\tilde{T}$ conditioned on $x_0 = i$, which we denote by t_i:

$$\begin{aligned} t_i &= \sum_{j=1}^{i} \sum_{k=j-1}^{n-1} \frac{(\kappa-1)(n-j)}{j} \cdots \frac{(\kappa-1)(n-k)}{k} \frac{\kappa n}{\varepsilon(k+1)} \\ &= \sum_{j=1}^{i} \sum_{k=j-1}^{n-1} (\kappa-1)^{k+1-j} \frac{(n-j)!}{(n-k-1)!} \frac{(j-1)!}{(k+1)!} \frac{\kappa n}{\varepsilon}. \end{aligned} \tag{2.7.2}$$

It is instructive to find asymptotics for these times to show how large they are with respect to n. Using a simple bound on $\binom{n-1}{k}$, we see

$$t_i \geqslant t_1 = \sum_{k=0}^{n-1} (\kappa-1)^k \binom{n-1}{k} \frac{1}{k+1} \frac{\kappa n}{\varepsilon} \geqslant \frac{n}{\varepsilon} \left(\frac{(n-1)(\kappa-1)}{\tilde{k}} \right)^{\tilde{k}} \tag{2.7.3}$$

for any $\tilde{k} \in [\![0, n-1]\!]$. Choosing $\tilde{k} = (\kappa-1)(n-1)/C$ for some $C > 1$ in the above expression and we find

$$t_i \geqslant \frac{1}{\varepsilon} C^{\frac{\kappa-1}{C}(n-1)}. \tag{2.7.4}$$

In particular, we see that time time to reach 0 is exponential in n. This is a prelude to many similar results we find in Chapter 5— finding particular genotypes in genotype space can take a very long time.

REMARK 2.29. The idea of projecting a random walk on the hypercube in this way goes back to Ehrenfest and is a common technique to analyze Markov chains [67, 162]. We are interested in it here because it motivates genotype spaces with a different sort of geometry. It also proves crucial in our analysis of the regeneration process in Section 5.4.

With this in mind, we can define this as a genotype space in its own right: Let $\Gamma := [\![0, n]\!]$ and define

$$\mathcal{M}(\alpha, \beta) := \begin{cases} \frac{\varepsilon(n-x_t)(\kappa-1)}{\kappa n} & \text{if } \beta = \alpha + 1 \\ \frac{\varepsilon x_t}{\kappa n} & \text{if } \beta = \alpha - 1 \\ \frac{\kappa n(1-\varepsilon)-(\kappa-1)n\varepsilon}{\kappa n} & \text{if } \beta = \alpha \\ 0 & \text{otherwise} \end{cases}. \tag{2.7.5}$$

While we have obtained this genotype space as a projection of the familiar hypercube, it is possible to take a more abstract perspective on the space.

DEFINITION 2.30. *This process is a Markov chain $(\Gamma, \mathcal{M})$ with state space $\Gamma := [\![0, n]\!]$ or $\Gamma = \mathbb{N}$ and transition kernel $\mathcal{M}$ defined by*

$$\mathcal{M}(\alpha, \beta) := \begin{cases} m_\alpha^+ & \textit{if } \beta = \alpha + 1 \\ m_\alpha^- & \textit{if } \beta = \alpha - 1 \\ 1 - m_\alpha^+ - m_\alpha^- & \textit{if } \beta = \alpha \\ 0 & \textit{otherwise} \end{cases} \tag{2.7.6}$$

for $m_\alpha^\pm \in [0, 1]$ for all $\alpha \in \Gamma$.

By choosing different values for the probabilities $m_i^\pm$, we can obtain many different interpretations of this genotype space. One interesting example is to consider $\alpha \in \Gamma$ counting the number of times a base pair is repeated. For example, $\cdots$ACGGGGGGCTTA$\cdots$ would have $\alpha = 6$ and $\cdots$ACGGGGCTTA$\cdots$ would have $\alpha = 4$. Variations like this in the genome of cancerous cells are import in determining the genealogy of tumors, as cancerous cells often loose the ability to copy sequences of repeated base pairs with high fidelity.

Now we turn to another interesting space that arises by projection. A common intuition for the transition probabilities (2.7.1), is to think of a particle performing a random walk in one dimension with a Gaussian potential centered at $x = n(\kappa-1)/\kappa$. We saw this in Theorem 2.7, but intuitively the transition probabilities have the following property: If $x_t > n(\kappa-1)/\kappa$, then it is more likely that x_t will decrease in value. If $x_t < n(\kappa-1)/\kappa$, then it is more likely that x_t will increase in value. The point $x_t = n(\kappa-1)/\kappa$, is where the transition probabilities are balanced, since

$$\frac{\varepsilon\,(n-x_t)\,(\kappa-1)}{\kappa n} = \frac{\varepsilon x_t}{\kappa n} \tag{2.7.7}$$

if and only if $x_t = n(\kappa-1)/\kappa$. This means that mutation pushes the process x_t away from 0 very strongly, which as we saw in (2.7.4), leads to $\mathbb{E}\tilde{T}$ growing exponentially as n gets larger. What are possible mechanisms that might reduce this hitting time? In Subsection 5.3.1, we consider how fitness affects this time, but here we want to focus on mutation alone.

Imagine there is some mutation that effectively resets the process $\boldsymbol{\alpha}_t$ to its initial condition $\boldsymbol{\alpha}_0$ and that the distance of $\boldsymbol{\alpha}_0$ from our target $\boldsymbol{\beta}$ is small relative to n—say $d = \mathcal{O}(1)$. We call this type of mutation *regeneration* and assume it happens with some constant probability regardless of the genotype. If we again project the process by $x_t := \mathcal{D}(\boldsymbol{\alpha}_t, \boldsymbol{\beta})$, then this setup leads to the following genotype space.

Definition 2.31. *This process is a Markov chain* $(\Gamma, \mathcal{M})$ *with state space* $\Gamma := [\![0, n]\!]$ *and transition kernel* $\mathcal{M}$ *defined by*

$$(1-\varepsilon_{\mathrm{R}})\tilde{\mathcal{M}}(\alpha,\beta) + \varepsilon_{\mathrm{R}}\mathbf{1}\,(\beta = d)\,, \tag{2.7.8}$$

where $\tilde{\mathcal{M}}$ *is given by* (2.7.5)*. The parameter* $\varepsilon_{\mathrm{R}} \in (0,1)$ *is the probability of regeneration to the genome* $d \in \Gamma$.

We can get a crude upper bound on the hitting time of 0 in this modified process. First, we calculate the probability that we reach state 0 before state $d+1$ when starting in state d–denoted by p. We assume that $\varepsilon_{\mathrm{R}} \ll \varepsilon$, then note that

$$\frac{\mathbb{P}\left\{x_{t+1} = x_t + 1 | x_t\right\}}{\mathbb{P}\left\{x_{t+1} = x_t - 1 | x_t\right\}} = \mathcal{O}\left(\frac{(\kappa-1)(n-x)}{x}\right) \tag{2.7.9}$$

since $d = \mathcal{O}(1)$. Thus,

$$p = \mathcal{O}\left(\frac{(\kappa-1)^{-d}\binom{n-1}{d}^{-1}}{1+\sum_{j=1}^{d}(\kappa-1)^{-j}\binom{n-1}{j}^{-1}}\right) = \mathcal{O}\left(d!(\kappa-1)^{-d}n^{-d}\right). \tag{2.7.10}$$

Since it takes $1/p$ tries to get from d to 0 in expectation and the the time between tries (or regeneration mutations) is $1/\varepsilon_{\mathrm{R}}$, the total expected time to reach state 0 is given by

$$\frac{(\kappa-1)^d n^d}{d!\varepsilon_{\mathrm{R}}}. \tag{2.7.11}$$

Note that this is much less than (2.7.4).

Taking a more abstract perspective on the above process, we can considered mutations that only move toward 0 except for a regeneration mutation. This process is easier to analyze and still has interesting properties.

Definition 2.32 (Mutation process on the line with regeneration). *This process is a Markov chain* $(\Gamma, \mathcal{M})$ *with state space* $\Gamma := [\![0, n]\!]$ *or* $\Gamma = \mathbb{N}$ *and transition kernel* $\mathcal{M}$ *defined by*

$$\mathcal{M}(\alpha, \beta) := \begin{cases} m_\alpha^+ & \textit{if } \beta = \alpha + 1 \\ 1 - m_\alpha^+ - \varepsilon_\alpha^{\mathrm{R}} & \textit{if } \beta = \alpha \\ \varepsilon_\alpha^{\mathrm{R}} & \textit{if } \beta = 0 \\ 0 & \textit{otherwise} \end{cases}. \tag{2.7.12}$$

To motivate Definition 2.32, suppose that there is some new functionality a cell can develop, but to do this it must make a number of neutral mutations n to reach this functionality. Suppose that $\alpha \in \Gamma$ records the number of mutations a particular cell has, so that $\alpha = 0$ is the wild-type and $\alpha = n$ is a cell with the new functionality. If we assume the mutations must be developed in some specified order, we would set

$$m_\alpha^+ = \frac{u}{n} \tag{2.7.13}$$

for some constant $u \in [0, 1]$. If the mutations can develop in any order, we have

$$m_\alpha^+ = u\frac{n-\alpha}{n}. \tag{2.7.14}$$

These steps are the forward mutations in the search process toward the new function. The search is lost at rate v: each $\alpha = k$ cell (with $k > 0$) mutates back to an $\alpha = 0$ cell with rate $\varepsilon_\alpha^{\mathrm{R}} = v$. The mutation rate v can represent the rate of deletion events, nonsense mutations, or any missense mutation that leads away from the target (because then the search is essentially lost). It is natural to assume that $v > u$, meaning that at each step, it is more likely that the search is lost than that a mutation is made in the direction of the target. This mutation scheme is known as the "regeneration process" and was introduced in [2, 10].

2.8. Continuous genotype spaces

Almost all of the examples of mutation processes we have considered so far have been finite—with $\Gamma = \mathbb{N}$ in Definition 2.30 being the only exception—and all have been discrete countable spaces. For completeness, in this section we give an example of an uncountable, continuous genotype space.

When considering continuous genotype spaces, there is a slight technical change in the way we denote the mutation kernel $\mathcal{M}$. Previously, we could use $\mathcal{M}(\alpha, \cdot)$ to represent the p.m.f. of a distribution on Γ, but in continuous spaces this might be an undefined. So in continuous genotype spaces, we specify

$$\mathcal{M}(\alpha, \mathcal{A}) \tag{2.8.1}$$

for $\mathcal{A} \in \sigma(\Gamma)$, where $\sigma(\Gamma)$ is a sigma algebra over Γ. Here we always have $\Gamma = \mathbb{R}$, and so it suffices to specify $\mathcal{M}(\alpha, [\beta_0, \beta_1])$ for all intervals $[\beta_0, \beta_1] \subset \mathbb{R}$.

Let the type of each individual be a positive real number $\alpha \in \mathbb{R}$ (or perhaps a vector in $\mathbb{R}^n$ for many features). Suppose that α records the value of some morphological feature or perhaps the expression level of a gene. A simple model of mutation in this case is multiplication by an independent random variable.

Definition 2.33 (Independent multiplicative mutation process on $\mathbb{R}^+$). *This process is a Markov chain* $(\Gamma, \mathcal{M})$ *with state space* $\Gamma := \mathbb{R}^+$ *and transition kernel* $\mathcal{M}$ *defined by*

$$\mathcal{M}(\alpha, [\beta_0, \beta_1]) := M(\beta_1/\alpha) - M(\beta_0/\alpha) \tag{2.8.2}$$

where $[\beta_0, \beta_1]$ *is an interval in* $\mathbb{R}^+$ *and* M *is the c.d.f. of some distribution on* $\mathbb{R}^+$.

A similar definition can be made for mutations that additively change the genotype.

Definition 2.34 (Independent additive mutation process on $\mathbb{R}$). *This process is a Markov chain* $(\Gamma, \mathcal{M})$ *with state space* $\Gamma := \mathbb{R}$ *and transition kernel* $\mathcal{M}$ *defined by*

$$\mathcal{M}\left(\alpha, [\beta_0, \beta_1]\right) := M(\beta_1 - \alpha) - M(\beta_0 - \alpha) \tag{2.8.3}$$

where $[\beta_0, \beta_1]$ *is an interval in* $\mathbb{R}$ *and* M *is the c.d.f. of some distribution on* $\mathbb{R}$.

3

Evolutionary dynamics

In this chapter, we introduce a number of classic stochastic models of evolutionary dynamics. These examples serve to motivate the typical problems we study in evolutionary dynamics, like determining the likelihood and expected time taken for a mutant to fix in a population. Later in the chapter, we define a very general stochastic model of evolutionary dynamics, that has as a special case all of the previous models.

In evolution, the population evolves, and so it forms the heart of all the models in this chapter. These models differ in exactly how the individuals in a population reproduce, die, and interact, but they all have some fundamental statistics that we focus on calculating. More complex models try to capture the effect of other variables, such as population structure, on these statistics. The models of this chapter describe microscopic evolutionary change. The events that concern us here all occur in relatively few generations compared to the totality of evolutionary history. Even statistics that are taken when the process has reached stationarity or equilibrium concern short timescales, as in most models stationarity is reached quickly.

As we mentioned in the introduction there are two places for stochasticity to enter into evolution: mutation and the way that reproduction and death evaluate and depend on fitness. Incorporating this randomness into evolutionary models is essential if we wish to capture some phenomena that are fundamental to evolution [1]. So we ommit here deterministic models, such as the replicator or quasi-species ODEs.

3.1. Moran process

The basic Moran process is perhaps the simplest stochastic model of evolution [13, 163, 164]. We consider a genotype space $\Gamma = \{\alpha, \beta\}$ and associate a fitness 1 to α (without loss of generality) and a fitness f to β, that is, $\mathcal{F}(\alpha) = 1$ and $\mathcal{F}(\beta) = f$. We consider a population of N individuals, each of which is either type α or β. This might indicate the absence or presence of a particular mutation, for example. Due to this interpretation, and because we often introduce a small number of type β genotypes into a mostly type α population, α is often referred to as the wild-type and β as the mutant. The state of the process at time t can be represented as a vector

$$\mathbf{x}(t) \in \{\alpha, \beta\}^N . \tag{3.1.1}$$

Often the initial population consists mainly of the wild-type with a small number of mutants. Specifically, we can choose the initial state of the process $\mathbf{x}(0)$ by simply placing a mutant at a uniformly random location.

The process is a Markov chain, where the next state is obtained from the previous one: First, randomly choose an individual proportional to its fitness. This individual now reproduces and randomly replaces another individual. Exactly how the individual to be replaced is randomly chosen can change (see Section 3.3), but here we focus on a well-mixed population and thus choose the individual uniformly at random.

Remark 3.1. Note that for this process, it is not actually necessary to keep track of the type of each individual. By symmetry, we can just project to the number of mutants $x_t := \sum_i \mathbf{1}\left(\mathbf{x}_t(i) = \beta\right)$. The projected process is still a Markov chain because the transition probabilities depend only on the number of mutants and not other details of the population's composition.

We might denote this event, some individual j getting replaced by some individual i, as $r(j) = i$, so that

$$\mathbb{P}\left\{r(j) = i | \mathbf{x}_t\right\} = \frac{\mathcal{F}(\mathbf{x}(i))}{\sum_{k=1}^{N} \mathcal{F}(\mathbf{x}(k))} \frac{1}{N} = \frac{1}{N} \frac{1 + (f-1)\mathbf{1}\left(\mathbf{x}_t(i) = \beta\right)}{f x_t + N - x_t} . \tag{3.1.2}$$

From here we can immediately write the transition probabilities for the process x_t:

$$\mathbb{P}\left\{x_{t+1} = x_t + \delta | x_t\right\} = \begin{cases} \frac{f x_t}{f x_t + N - x_t} \frac{N - x_t}{N} & \text{if } \delta = 1 \\ \frac{N - x_t}{f x_t + N - x_t} \frac{x_t}{N} & \text{if } \delta = -1 \\ \frac{f x_t}{f x_t + N - x_t} \frac{x_t}{N} + \frac{N - x_t}{f x_t + N - x_t} \frac{N - x_t}{N} & \text{if } \delta = 0 \\ 0 & \text{otherwise} \end{cases}. \tag{3.1.3}$$

So x_t is a simple birth-death process that makes jumps of size at most 1. Note that

$$\frac{\mathbb{P}\left\{x_{t+1} = x_t + 1 | x_t\right\}}{\mathbb{P}\left\{x_{t+1} = x_t - 1 | x_t\right\}} = f. \tag{3.1.4}$$

This ratio's independence of the current state x_t, makes analysis of the Moran process very straightforward.

3.1.1. Longrun statistics. In the long run, the process has only two possible outcomes as exactly two of the states, 0 and N, are absorbing. Either the mutants fix and the wild-type dies out or the reverse. We call the probability of the first eventuality *the fixation probability.* We can describe this event mathematically using hitting times. Define $T = \min\{t : x_t = 0 \text{ or } x_t = N\}$, then the fixation probability is defined by

$$\rho := \mathbb{P}\left\{X_T = N\right\} = \mathbb{E} X_T / N. \tag{3.1.5}$$

In the case of the Moran process, ρ has a closed form solution. There are many ways to calculate ρ as the process is very similar to the gambler's ruin problem. The fundamental difference between the two processes is that the rate at which x_t changes value is not constant in the Moran process. As x_t gets closer to 0 or N the rate of change slows down, whereas it remains constant in the gambler's ruin process. However, the varying rate of change does not affect absorption probabilities. Probably the cleanest way to find the fixation probability is to note that when $f \neq 1$, f^{-x_t} is a martingale with respect to x_t (when $f = 1$, simply use x_t itself).

Lemma 3.2. *Define $y_t := f^{-x_t}$ (use $y_t := x_t$ when $f = 1$) is a martingale with respect to x_t.*

Proof. Just calculate and use (3.1.4) to see

$$\mathbb{E}\left[f^{-x_{t+1}} | x_t = i\right] = f^{-i-1} + f^{-i+1} P_i^- + f^{-i}\left(1 - P_i^+ - P_i^-\right) = f^{-i} = y_t, \tag{3.1.6}$$

where $P_i^{\pm}$ denotes $\mathbb{P}\{x_{t+1} = x_t \pm 1 | x_t\}$. Note that f^{-x_t} is a martingale whenever $P_i^+/P_i^- = f$ for all i. ■

With this martingale in hand, we can quickly find ρ.

THEOREM 3.3. *The fixation probability of the Moran process is given by*

$$\rho = \frac{1 - f^{-1}}{1 - f^{-N}}. \tag{3.1.7}$$

PROOF. Let T be the absorption time, then apply optional stopping to the above martingale:

$$r^{-1} = \mathbb{E}y_0 = \mathbb{E}y_T = r^{-N}\rho + r^{-0}(1 - \rho), \tag{3.1.8}$$

which rearranges to

$$\rho = \frac{1 - f^{-1}}{1 - f^{-N}}. \tag{3.1.9}$$

Note that optional stopping is stronger than needed here as the state space of the process is finite—all that is required is that the expectation of y_t is constant in time. ■

Another fundamental statistic of the process is the *expected absorption time*, which is defined as

$$t_i := \mathbb{E}\left[T | x_0 = i\right], \tag{3.1.10}$$

where we have conditioned on the initial value of the process. While Equation (3.1.10) has units of time steps, it is common to state results about absorption times in terms of number of generations, where one generation is the number of time steps required for N reproduction (which is exactly N time steps in the Moran process). For the Moran process, there is a closed-form solution for t_i. We state the result below and prove it in the special case $f = 1$. Note that absorption takes a constant number of generations when $f < 1$ and a logarithmic number when $f \geqslant 1$.

THEOREM 3.4. *The expected absorption time of the Moran process is given by*

$$t_1 = \frac{N}{\rho} \sum_{k=1}^{N-1} \sum_{l=1}^{k} \frac{fl + N - l}{l(N - l)} f^{l-k-1}. \tag{3.1.11}$$

When $f < 1$,

$$t_1 = \mathcal{O}(N), \tag{3.1.12}$$

and when $f \geqslant 1$,

$$t_1 = \mathcal{O}(N \log N). \tag{3.1.13}$$

Proof. For the general case, we have the recurrence equations

$$(1 - \mathbb{P}\{x_{t+1} = x_t | x_t\})\, t_i = 1 + \mathbb{P}\{x_{t+1} = x_t + 1 | x_t\}\, t_{i+1} + \mathbb{P}\{x_{t+1} = x_t - 1 | x_t\}\, t_{i-1}, \tag{3.1.14}$$

using the definition in (3.1.3). These equations are straightforward but messy to solve. See [165] for details.

For the case $f = 1$, it is easy to see that $y_t := t + g(x_t)$ is a martingale with respect to x_t, where

$$g(x) := N \sum_{j=1}^{x} \frac{N-x}{N-j} + N \sum_{j=x+1}^{N-1} \frac{x}{j}. \tag{3.1.15}$$

Hence, by Doob's optional stopping theorem, we see

$$N \sum_{j=1}^{i} \frac{N-x}{N-j} + N \sum_{j=i+1}^{N-1} \frac{x}{j} = f(i) + 0 = \mathbb{E}y_0 = \mathbb{E}y_T = 0 + \mathbb{E}T = t_i. \tag{3.1.16}$$

Note in particular that when $i = 1$, the left-hand side of Equation (3.1.16) is N times a finite harmonic sum. Thus, $t_1 = \mathcal{O}(N \log N)$.

A similarly easy argument can be use to prove (3.1.13) for large f. For $f \gg N$, it is very unlikely that the number of mutants ever decreases, thus we can approximately decompose T as a sum of geometric random variables. Define $T_i := \min\{t : X_t = i\}$, then

$$\mathbb{E}T = \mathbb{E}T_N \approx \sum_{i=1}^{N-1} \mathbb{E}[T_{i+1} - T_i] \approx \sum_{i=1}^{N-1} \frac{N}{i} = \mathcal{O}(N \log N), \tag{3.1.17}$$

since $T_{i+1} - T_i \sim \mathrm{Geo}((N-i)/N)$ because a mutant will always be selected for reproduction but only replaces a wild-type with probability $(N-i)/N$..

It is interesting to note that we get the same result for $f = 1$ as $f \to \infty$, but for completely different reasons. When $f = 1$, typically the process is absorbed at 0, but this can require x_t to change its value $\log N$ times. Moreover, initially x_t changes value roughly every N steps, leading to the expected absorption time of $N \log N$. When $f \to \infty$, the process is always absorbed at N. This requires x_t to change value N times. In this case, x_t changes value roughly every $N/(N - x_t)$ steps. This leads to the absorption time $N \log N$.

To find (3.1.13) in general for $f > 1$, we can analyze (3.1.11) directly:

$$\begin{aligned} t_1 &= \mathcal{O}\left(\sum_{k=1}^{N-1}\sum_{l=1}^{k} \frac{fl + N - l}{fl}\frac{N}{N-l} f^{l-k}\right) \\ &= \mathcal{O}\left(\frac{f-1}{f}\sum_{k=1}^{N-1}\frac{N}{N-k} + \frac{1}{f}\sum_{k=1}^{N-1}\frac{N}{k}\frac{N}{N-k}\right) \\ &= \mathcal{O}\left(N \log N\right) \end{aligned} \tag{3.1.18}$$

since $\rho = \mathcal{O}(1)$.

When $f < 1$, we analyze (3.1.11) directly again:

$$\begin{aligned} t_1 &= \mathcal{O}\left(f^N \sum_{k=1}^{N-1}\sum_{l=1}^{k} \frac{fl + N - l}{fl}\frac{N}{N-l} f^{l-k}\right) \\ &= \mathcal{O}\left(f^N N \sum_{k=1}^{N-1} f^{-k}\right) \\ &= \mathcal{O}\left(N\right), \end{aligned} \tag{3.1.19}$$

since $\rho = \mathcal{O}\left(f^N\right)$. ■

The expected absorption time can be decomposed as

$$\mathbb{E}T = (1-\rho)\,\mathbb{E}\left[T | X_T = 0\right] + \rho\,\mathbb{E}\left[T | X_T = N\right]. \tag{3.1.20}$$

The statistics $\mathbb{E}\left[T | X_T = 0\right]$ and $\mathbb{E}\left[T | X_T = N\right]$ are also of interest and are called the *expected conditional extinction time* and *expected conditional fixation time* respectively. In some cases, they can be studied in a similar way as in the proof of Theorem 3.4, since under the measures $\mathbb{P}\left\{\cdot | X_T = 0\right\}$ and $\mathbb{P}\left\{\cdot | X_T = N\right\}$ the process is still Markovian [13].

It is also possible to study the distribution of the random variable T. For any birth-death chain, the absorption time to any state can be decomposed into a sum of independent geometric random variables due to a result attributed to Keilson [166, 167] (also see [168] for a probabilistic proof). These exponential random variables are parameterized by the eigenvalues of the tridiagonal transition matrix.

3.1.2. Stationary statistics. Unlike many of the genotype spaces we consider in Chapter 2, we have so far implicitly assumed that there is no mutation between the types α and β. This means that the mutation process on Γ is irreducible and thus so is the Moran process (see Theorem 4.1). However, if we define

$$\begin{pmatrix} \mathcal{M}(\alpha,\alpha) & \mathcal{M}(\alpha,\beta) \\ \mathcal{M}(\beta,\alpha) & \mathcal{M}(\beta,\beta) \end{pmatrix} := \begin{pmatrix} 1-\varepsilon_\alpha & \varepsilon_\alpha \\ \varepsilon_\beta & 1-\varepsilon_\beta \end{pmatrix}, \tag{3.1.21}$$

the Moran process is an irreducible, aperiodic Markov chain on $[\![N]\!]$. It is easy to show that

$$\pi_\alpha = \frac{\varepsilon_\beta}{\varepsilon_{\alpha+\varepsilon_\beta}} \quad \text{and} \quad \pi_\beta = \frac{\varepsilon_\alpha}{\varepsilon_\alpha + \varepsilon_\beta}, \tag{3.1.22}$$

where π is the stationary distribution of the mutation process (3.1.21).

The entries of $\mathcal{M}$ specify the probability of each type being produced during reproduction, so with this genotype space, it is natural to define the following transition probabilities for the Moran process with mutation.

$$\mathbb{P}\{x_{t+1} = x_t + \delta | x_t\} = \begin{cases} f(1-\varepsilon_\beta)\frac{x_t}{fx_t+N-x_t}\frac{N-x_t}{N} + \varepsilon_\alpha\frac{N-x_t}{fx_t+N-x_t}\frac{N-x_t}{N} & \text{if } \delta = 1 \\ (1-\varepsilon_\alpha)\frac{x_t}{fx_t+N-x_t}\frac{N-x_t}{N} + f\varepsilon_\beta\frac{x_t}{fx_t+N-x_t}\frac{x_t}{N} & \text{if } \delta = -1 \\ f(1-\varepsilon_\beta)\frac{x_t}{fx_t+N-x_t}\frac{x_t}{N} + (1-\varepsilon_\alpha)\frac{N-x_t}{fx_t+N-x_t}\frac{N-x_t}{N} & \text{if } \delta = 0 \\ \quad +f\varepsilon_\beta\frac{x_t}{fx_t+N-x_t}\frac{N-x_t}{N} + \varepsilon_\alpha\frac{x_t}{fx_t+N-x_t}\frac{N-x_t}{N} & \\ 0 & \text{otherwise} \end{cases}. \tag{3.1.23}$$

Note, we can find the stationary distribution of the Moran process with mutation using the formula in Theorem A.1. However, the expression is quite complicated, as there is very little cancelation in the terms from Equation (3.1.23), so we avoid writing it down here.

Remark 3.5. The process we have considered above is sometimes referred to as the Moran process with birth-death updating, where "birth-death" is used to refer to the order in which these two events occur. Alternatively, the order can be reversed: First an individual is chosen uniformly at random from the population. Second, this chosen individual is replaced by the offspring of another individual who is randomly chosen from the population proportional to fitness. While in well-mixed populations this change has little effect, the dynamics can look quite different in more complex cases [169].

Another possible change is to change the rate at which individuals die (or are replaced). In this case, individuals are selected for replacement proportional to their death rates. In most cases, this change is minor and leads to similar dynamics to a standard Moran process with fitnesses $\tilde{f} = f/d$, where f is the old fitness and d is the old death rate.

The Moran process is often generalized to model other more complicated system (see [7] for an example). Such models are referred to as Moran-type models, where "Moran-type" simply means that a single reproduction and death event occurs at each time step and that the individuals dying and reproducing are selected randomly, proportional to some rate.

3.2. Wright-Fisher process

The Wright-Fisher process is another simple stochastic model of evolution on short timescales. The major difference being that in the Wright-Fisher process the whole population is replaced at each time step rather than a single individual [13, 170, 171]. As we did for the Moran process in Section 3.1, we can develop a similar understanding of many of the relevant statistics.

Again, we focus on the genotype space $\Gamma = \{\alpha, \beta\}$ and at first assume no mutation between types. At each time step, the whole population is replaced—we call this a new generation. The type of individual i is sampled form the previous generation proportional to fitness. As in the Moran process, it is not necessary to know the type of each individual, as projecting to simply the number of individuals of type β still leads to a Markovian process. We record the number of individuals of type β at time t with x_t. Thus, we can write the transition probabilities:

$$\mathbb{P}\left\{x_{t+1} = i | x_t\right\} = \binom{N}{i} \left(\frac{f x_t}{f x_t + N - x_t}\right)^i \left(\frac{N - x_t}{f x_t + N - x_t}\right)^{N-i}, \tag{3.2.1}$$

for $i \in [\![N]\!]$. Note that (3.2.1) is equivalent to

$$(x_{t+1} | x_t) \sim \mathrm{Bin}\left(N, \frac{f x_t}{f x_t + N - x_t}\right). \tag{3.2.2}$$

Thus, when $f = 1$, x_t is a martingale with respect to itself. So by the same argument as in the proof of Lemma 3.2, we find $\rho = 1/N$ when $f = 1$. Define the fixation probability conditional on starting with i individuals of type β as $\rho_i := \mathbb{P}\{x_T = N | x_0 = i\}$. When $f \neq 1$, no closed form solution is known for ρ, but it can be approximated with an error that decreases to 0 as N goes to infinity. There are two regimes: (1) the fitness of type β and that the initial number of type β are both order one; (2) the fitness of type β is order $1/N$ and the initial number of type β is order N. In the second regime, we write $\tilde{\rho}(i/N) = \rho_i$ and $\tilde{f} = N(f-1)$ and assume that $\tilde{\rho}$ has a continuously differentiable limit as $N \to \infty$.

THEOREM 3.6. *In the first regime, the fixation probability of the Wright-Fisher process ρ_1 is given by $1-\theta$ (with an error of size $\mathcal{O}(1/N)$), where θ is the minimum solution of the equation*

$$\theta = e^{f(\theta-1)}. \tag{3.2.3}$$

In the second regime, the fixation probability of the Wright-Fisher process is given by

$$\tilde{\rho}(x) = \frac{1-e^{-\tilde{f}x}}{1-e^{-\tilde{f}}} + \mathcal{O}(1/N). \tag{3.2.4}$$

PROOF (INFORMAL). Using the Markov property and conditioning on the next step of the process, we see

$$\rho_i = \mathbb{E}\rho_X, \tag{3.2.5}$$

where $X \sim \mathrm{Bin}\left(N, \frac{fi}{fi+N-i}\right)$. In the first regime, note that asymptotically $X \sim \mathrm{Pois}(fi)$. Moreover, let $\theta_i = 1-\rho_i$, and note that in the large population limit $\theta_i = \theta_1^i$ [13]. This automatically incorporates the initial condition $\theta_0 = 1$. Thus, (3.2.5) become

$$\theta_1 = \sum_{j=0}^{\infty} \theta_1^j \frac{f^j e^{-f}}{j!} + \mathcal{O}(1/N) = e^{f(\theta_1-1)} + \mathcal{O}(1/N). \tag{3.2.6}$$

In the second regime, we Taylor expand $\tilde{\rho}$ at $x = i/N$, and use (3.2.5) to see

$$\begin{aligned}
\tilde{\rho}(x) &= \tilde{\rho}(x) + \tilde{\rho}'(x)\mathbb{E}\left(\frac{X}{N} - x\right) + \tilde{\rho}''(x)\mathbb{E}\left(\frac{X}{N} - x\right)^2 \\
&= \tilde{\rho}(x) + \tilde{\rho}'(x)\left(\frac{\tilde{f}x(1-x)}{N} + \mathcal{O}\left(1/N^2\right)\right) + \tilde{\rho}''(x)\left(\frac{x(1-x)}{N} + \mathcal{O}(1/N^2)\right).
\end{aligned} \tag{3.2.7}$$

Thus, $\tilde{f}\tilde{\rho}'(x) + \tilde{\rho}''(x) = \mathcal{O}\left(1/N\right)$, and solving the differential equation yields

$$\tilde{\rho}(x) = \frac{1 - e^{-\tilde{f}x}}{1 - e^{-\tilde{f}}} + \mathcal{O}\left(1/N\right). \tag{3.2.8}$$

This technique is a precursor to diffusion theory [13]. ∎

The following Theorem expands the formula for ρ_1 from Theorem 3.6 in the weak selection limit. The expression we obtain proves useful in Chapter 5.

THEOREM 3.7. *Suppose $f = 1 + \phi/N$ for $\phi \geqslant 0$, then the fixation probability of the Wright-Fisher process ρ_1 is given by*

$$\rho_1 = \frac{\phi}{2N} + \mathcal{O}(N^{-2}). \tag{3.2.9}$$

PROOF. Taking logs of Equation (3.2.3), substituting $\theta = 1 - \rho$, and Taylor expanding, we see

$$\rho + \rho^2/2 + \mathcal{O}\left(\rho^3\right) = \rho + \rho\frac{\phi}{N} + \mathcal{O}\left(N^{-2}\right), \tag{3.2.10}$$

thus assuming ρ is small, we see

$$\rho \approx \frac{\phi}{2N}. \tag{3.2.11}$$

∎

Finally, we consider the absorption time for the Wright Fisher process.

THEOREM 3.8. *Suppose $f = 0$ and that $x = i/N$, then the absorption time T satisfies*

$$\mathbb{E}T = -N\left(x\log x + (1-x)\log(1-x)\right) + \mathcal{O}(1). \tag{3.2.12}$$

PROOF (INFORMAL). As we did for ρ previously, we write $\tilde{t}(i/N) := t_i$ and assume that $\tilde{t}$ has a continuously differentiable limit as $N \to \infty$. Let $X \sim \text{Bin}\left(N, \frac{i}{N}\right)$. Then, using an Equation similar to (3.1.14), we find

$$\begin{aligned}\tilde{t}(x) &= 1 + \tilde{t}(x) + \tilde{t}'(x)\mathbb{E}\left(\frac{X}{N} - x\right) + \tilde{t}''(x)\mathbb{E}\left(\frac{X}{N} - x\right)^2 \\ &= 1 + \tilde{t}(x) + \tilde{t}''(x)\left(\frac{x(1-x)}{N} + \mathcal{O}(1/N^2)\right). \end{aligned} \tag{3.2.13}$$

by Taylor expansion. Finally, solving the differential equation $\tilde{t}''(x) = -\frac{N}{x(1-x)} + \mathcal{O}(1/N^2)$ and using the boundary condition $\tilde{t} = \tilde{t} = 0$, we find

$$\tilde{t}(x) = -N\left(x \log x + (1-x)\log(1-x)\right) + \mathcal{O}(1). \tag{3.2.14}$$

■

Note that when $x = 1/N$, we have

$$\mathbb{E}T = \log N + (N-1)\log(1-1/N) = \log N + \mathcal{O}(1) \tag{3.2.15}$$

for the expression in Equation (3.2.12). The expression for the absorption time when $f \neq 1$ is more complicated (see [13]), but we still have $\mathbb{E}[T|x_0 = 1] = \mathcal{O}(\log N)$.

3.3. Evolutionary graph theory

So far we have focused on stochastic models that implicitly assume that the population structure is well-mixed—that is, all interaction between pairs of individuals are equal. In particular, we have assumed that when some individual reproduces, it is equally likely to replace any individual. In this section we introduce a model of evolutionary dynamics in structured populations that modifies this assumption.

The Moran process on graphs is the standard model for population structure in evolutionary dynamics. [1, 123] The process is defined for a directed, weighted graph, $G_N = ([\![N]\!], W_N)$, where W_N is a stochastic matrix of edge weights. The matrix W_N describes the population structure and the entry $W_N(i,j)$ is the probability that individual i replaces individual j when it reproduces. The process, denoted by $\mathbf{x}_t$, is Markovian with state space $\{\alpha, \beta\}^N$, where $\mathbf{x}_t$ is a vector describing the type of the individual located at each vertex at time t. The state of the process can also be recorded with $S_t := \{i : \mathbf{x}_t(i) = \beta\}$, that is the subset of individuals who are of type β at time t. Note that for general graphs, simply recording the number of individuals of type β no longer projects to a Markov chain.

At time 0 a mutant is placed at one of the vertices uniformly at random or formally,

$$\mathbb{P}[S_0 = S] = \begin{cases} N^{-1} & \text{if } |S| = 1 \\ 0 & \text{otherwise} \end{cases}. \tag{3.3.1}$$

Then at each subsequent time step exactly one vertex is chosen randomly, proportional to its fitness, for reproduction: so the probability of choosing a vertex of type α is $1/F_t$ and the probability of choosing a particular mutant vertex is f/F_t, where

$$F_t = N - |S_t| + f\,|S_t| \tag{3.3.2}$$

is the total fitness of the population at time t. An edge originating from the chosen vertex is then selected randomly with probability equal to its edge weight, which is well defined since W is stochastic, and the vertex at the destination of the edge takes on the type of the vertex at the origin of the edge. We can write the transition probabilities of the process precisely as

$$\mathbb{P}\left[S_{t+1} = S|S_t\right] = \begin{cases} fW(S_t, i)/F_t & \text{if } S = S_t \sqcup \{i\} \text{ for some } i \notin S \\ W(S_t^c, i)/F_t & \text{if } S = S_t \backslash \{i\} \text{ for some } i \in S \\ \left(fW(S_t, S_t) + W(S_t^c, S_t^{c)}\right)/F_t & \text{if } S = S_t \\ 0 & \text{otherwise} \end{cases} . \tag{3.3.3}$$

Typically, there are exactly two absorbing states, $\mathbf{x}_t = (\alpha, \ldots, \alpha)$ and $\mathbf{x}_t = (\beta, \ldots, \beta)$, corresponding to the wild-type fixing in the population and the mutants fixing in the population respectively. This is true whenever the graph has at most one root and is strongly connected; we assume this from now on. Thus, almost surely, one of these two absorbing states is reached in finite time T, where

$$T := \min\{t : \mathbf{x}_t = (\alpha, \ldots, \alpha) \text{ and } \mathbf{x}_t = (\beta, \ldots, \beta)\} = \min\{t : |S_t| = 0 \text{ or } |S_t| = N\} \tag{3.3.4}$$

is called the absorption time. The probability that the process reaches $(\beta, \ldots, \beta)$ and not $(\alpha, \ldots, \alpha)$ is called the *fixation probability* and for a graph G we denote its fixation probability for a mutant of fitness $f > 0$ by

$$\rho_G := \mathbb{P}\{\mathbf{x}_T = (\beta, \ldots, \beta)\}. \tag{3.3.5}$$

A simple argument shows that $\rho = 1/N$ for all graphs when $f = 1$: let X_i be a random variable indicating that the lineage of individual i eventually fixed in the population. Then note the events $\{X_i = 1\}$ are mutually exclusive, so

$$\sum_{i=1}^{N} X_i = 1, \tag{3.3.6}$$

as some lineage must fix. However, the mutant is neutral and starts at a uniformly random location in the population, so

$$\rho = \sum_{i=1}^{N} \frac{1}{N} \mathbb{P}\left\{S_T = [\![N]\!] | S_0 = \{i\}\right\} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{E} X_i = \frac{1}{N}. \tag{3.3.7}$$

So ρ agrees with our formula for fixation probability in the Moran and Wright-Fisher processes when $f = 1$. What about when $f \neq 1$?

A fundamental point of comparison is the fixation probability ρ_{M} for a *well-mixed* population structure, where the graph structure is given by $W(i,j) = 1/N$ for all $i, j \in [\![N]\!]$ and M stands for "Moran" or "mixed." We calculated ρ_{M} in Theorem 3.3. An easy way to illustrate evolutionary graph theory is the *cycle*. The cycle is a graph with weight matrix

$$\begin{pmatrix} 0 & 1/2 & 0 & \cdots & 0 & 1/2 \\ 1/2 & \ddots & \ddots & \ddots & \ddots & 0 \\ 0 & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ 0 & \ddots & \ddots & \ddots & \ddots & 1/2 \\ 1/2 & 0 & \cdots & 0 & 1/2 & 0 \end{pmatrix}. \tag{3.3.8}$$

From an initial state of a single type β, only states where the β types are clustered together can be reached. Thus, by the symmetry of the graph, we can simply project to the number of individuals of type β, denoted by x_t, and still have a well-defined Markov chain. Moreover, this Markov chain is a birth-death process on $[\![N]\!]$. As in the proof of Theorem 3.3, we can show that f^{-x_t} is a martingale by showing that the ratio of increasing over decreasing the number of mutants is constant. To increase the number of mutants, we must choose a mutant to reproduce; also, this individual must be on the boundary of the cluster and must replace in the right direction, otherwise it will replace another type mutant. Thus,

$$\frac{2f}{F_t} \frac{1}{2} \tag{3.3.9}$$

is the probability of x_t increasing. By similar reasoning, the probability of x_t decreasing is

$$\frac{f}{F_t}\frac{1}{2}, \tag{3.3.10}$$

and thus their ratio is f. This proves that $\rho = \rho_{\mathrm{M}}$ as the martingale argument in the derivation of ρ_{M} can now be applied.

Many other graphs have $\rho = \rho_{\mathrm{M}}$ and for many years only such examples were known [172]. This is because only symmetric population structures (where $W(i,j) = W(j,i)$) had been considered. Graphs with exactly the same fixation probability as ρ_{M} are classified by the isothermal theorem, which gives sufficient a condition for a general graph G to have the same fixation probability as ρ_{M} [123].

We introduce the following notation for sums of subsets of matrix entries:

$$W\left(S, S'\right) := \sum_{i\in S}\sum_{j\in S'} W(i,j), \tag{3.3.11}$$

for an $N \times N$ matrix W and subsets $S, S' \subseteq [\![N]\!]$. Throughout this section, we shall require that the matrix W is stochastic—that is, the row sums are all equal to 1:

$$\sum_{j=1}^{N} W(i,j) = 1, \tag{3.3.12}$$

for all $i \in [\![N]\!]$. Any graph with nonnegative edge weights can be normalized to produce a graph with a stochastic W, so long as each row has a nonzero entry, without changing the behavior of the process as defined above. Some authors prefer, particularly when the population structure is described by a simple, unweighted graph, to avoid explicitly requiring or normalizing the weight matrix to be stochastic, but this is still done implicitly in the definition of the process [173, 174].

Definition 3.9 (Isothermality). *A graph G is called isothermal if all the column sums of W are identical—that is,*

$$W([\![N]\!], i) = 1 \tag{3.3.13}$$

for all $i \in [\![N]\!]$, or equivalently, W is doubly stochastic. Note that all symmetric population structures are isothermal.

An easy calculation shows that a graph is isothermal if and only if

$$W(S, S^c) = W(S^c, S) \tag{3.3.14}$$

for all $\varnothing \neq S \subsetneq [\![N]\!]$, where $S^c := [\![N]\!] \backslash S$ is the set complement. The condition (3.3.14) and its equivalence to isothermality is at the core of the proof of the isothermal theorem. The term "isothermal" originates from an interpretation of the sum of the ingoing edge weights as heat, with "hotter" vertices changing more frequently in the Moran process. Thus, a graph satisfying (3.3.14) is isothermal because the ingoing and outgoing flow of heat is equal and all subsets S are in "thermal equilibrium." We now restate the forward direction of the original isothermal theorem.

Theorem 3.10 (Isothermal theorem). *Suppose that a graph G is isothermal, then the fixation probability of a randomly placed mutant of fitness f is equal to ρ_{M}.*

Proof. The proof is very similar to our derivation in the proof of Theorem 3.3: we show that $y_t := f^{-|S_t|}$ is a martingale with respect to S_t. First note that

$$\mathbb{P}\left\{|S_{t+1}| = |S_t| + 1 | S_t\right\} = \sum_{i \notin S_t} f W(S_t, i)/F_t = f W(S_t, S_t^c)/F_t, \tag{3.3.15}$$

as changing any individual of type α to type β increases $|S_t|$ by 1. Similarly,

$$\mathbb{P}\left\{|S_{t+1}| = |S_t| - 1 | S_t\right\} = W(S_t^c, S_t)/F_t. \tag{3.3.16}$$

Now we prove Equation (3.3.14):

$$\begin{aligned}
\sum_{i \in S} \sum_{j \notin S} W(i,j) &= \sum_{i \in S} \left(1 - \sum_{j \in S} W(i,j)\right) \\
&= |S| - \sum_{j \in S} \sum_{i \in S} W(i,j) \\
&= |S| - \sum_{j \in S} \left(1 - \sum_{i \notin S} W(i,j)\right) \\
&= \sum_{j \in S} \sum_{i \notin S} W(i,j).
\end{aligned} \tag{3.3.17}$$

Therefore, calculating, we see

$$\begin{aligned}
\mathbb{E}\left[f^{-\left|S_{t+1}\right|}|S_t\right] &= f^{-|S_t|-1}\frac{fW(S_t,S_t^c)}{F_t} + f^{-|S_t|+1}\frac{W(S_t^c,S_t)}{F_t} + f^{-|S_t|}\left(1-\frac{fW(S_t,S_t^c)}{F_t}-\frac{W(S_t^c,S_t)}{F_t}\right) \\
&= f^{-|S_t|} + f^{-|S_t|}\frac{W(S_t,S_t^c)}{F_t}\left(1+f-f-1\right) \\
&= f^{-|S_t|}. \qquad (3.3.18)
\end{aligned}$$

Note that in general $c^{|S_t|}$ is a martingale whenever

$$\frac{\mathbb{P}\left\{|S_{t+1}| = |S_t|-1|S_t = S\right\}}{\mathbb{P}\left\{|S_{t+1}| = |S_t|+1|S_t = S\right\}} \qquad (3.3.19)$$

is a constant c for all S. The proof is then easily completed with Doob's optional stopping in the same way as the proof of Theorem 3.3. ■

We ask, can we relax the assumptions of Theorem 3.10? That is, perhaps an approximate result can be obtained for W that are only approximately doubly stochastic in the following sense:

$$|W(\llbracket N\rrbracket, j) - 1| \leqslant \varepsilon \qquad (3.3.20)$$

for all $j \in \llbracket N\rrbracket$ and some small positive quantity ε. However, the graph $G_\varepsilon = (\{1,2,3,4\},W)$, where $0<\varepsilon<1$ and

$$W = \begin{bmatrix} 0 & 1-\varepsilon & 0 & \varepsilon \\ 1-\varepsilon & 0 & \varepsilon & 0 \\ 0 & \varepsilon^2 & 0 & 1-\varepsilon^2 \\ \varepsilon^2 & 0 & 1-\varepsilon^2 & 0 \end{bmatrix}, \qquad (3.3.21)$$

shows we cannot, since as $\varepsilon \to 0$

$$\frac{W(\{1,2\},\{3,4\})}{W(\{3,4\},\{1,2\})} = \frac{2\varepsilon}{2\varepsilon^2} = \varepsilon^{-1} \to \infty. \qquad (3.3.22)$$

That is, W is approximately doubly stochastic, but the ratio of the outgoing and ingoing edge weights is unbounded for some subset S. Moreover, it is easy to show that the fixation probability is such that

$$\lim_{\varepsilon \to 0} \rho_{G_\varepsilon}(r) = \frac{1}{2} \frac{1 - r^{-1}}{1 - r^{-2}}, \tag{3.3.23}$$

which is far from $\rho_{\mathrm{M}} = (1 - r^{-1})/(1 - r^{-4})$. Conditioning on whether the process starts at vertex 1 or 3, the difference is even more pronounced and given by

$$\lim_{\varepsilon \to 0} \rho^1_{G_\varepsilon}(r) = \frac{1 - r^{-1}}{1 - r^{-2}} \quad \text{and} \quad \lim_{\varepsilon \to 0} \rho^3_{G_\varepsilon}(r) = 0 \tag{3.3.24}$$

respectively. Thus, we need stronger assumptions for an approximate theorem—we take up this challenge in Chapter 6.

While Theorem 3.10 classifies population structures whose fixation probability is close to that of the well-mixed, the main interest in structured populations is how they can change the behavior of evolutionary dynamics. To illustrate some of these ideas, we study two special graphs that deviate from the well-mixed in two different directions.

Consider the following graph with weight matrix

$$\begin{pmatrix} 0 & 1/N & \cdots & 1/N \\ \vdots & \vdots & \cdots & \vdots \\ 0 & 1/N & \cdots & 1/N \end{pmatrix}. \tag{3.3.25}$$

We call this graph the *tennis racket*. Calculating the fixation probability of this graph is very easy, as the state $(\beta, \ldots, \beta)$ is only accessible from the initial state $(\beta, \alpha, \ldots, \alpha)$, that is, when the mutant arrises at vertex 1. This is because the vertex 1 is a root of the graph. Moreover, from the state $(\beta, \alpha, \ldots, \alpha)$, the state $(\alpha, \ldots, \alpha)$ cannot be reached by the Markov chain. Thus, the fixation probability is

$$\rho = \frac{1}{N}. \tag{3.3.26}$$

for all f. The interpretation of Equation (3.3.26) is that this population structure completely removes any dependence on the fitness difference between types—it is identical to the fixation probability of a neutral mutant in a well-mixed population. In this sense, it suppresses selection [4].

The effect of selection can also be amplified. The simplest example of an amplifying population structure is the *star*, which is a graph with vertex set $[\![N+1]\!]$ and weight matrix

$$\begin{pmatrix} 0 & 1/N & \cdots & 1/N \\ 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ 1 & 0 & \cdots & 0 \end{pmatrix}. \tag{3.3.27}$$

The fixation probability for the star is approximately

$$\frac{1-f^{-2}}{1-f^{-2n-2}}. \tag{3.3.28}$$

The vertex 1 is referred to as the center and the other vertices are referred to as leaves. Fitness differences are effectively squared when compared to the well-mixed population structure. For example, a 10% fitness advantage on the star yields roughly the same fixation probability as a 21% fitness advantage in the well-mixed population structure [4].

Before we prove (3.3.28), we build some intuition for the result. Let $c_t \in \{0,1\}$ denote whether the central vertex is type β and $l_t \in \{0,1,\ldots,N\}$ denote the number of type β on the leaves. We consider the possible actions that happen on the star and the rates at which they occur. In a typical step on the star, a leaf is chosen for reproduction, since there are many more leaves on a large star. Thus, the center changes type very rapidly compared to the leaves, as they can only change when the center reproduces. Let us imagine that the center is an independent process that changes between the states 0 and 1. We can then ask what is the probability that it is in state 0 (denoted by ϕ)? This is simply the proportion of the rates of transition (in exactly the same way to the calculation in (3.1.22)). Here we are essentially assuming that the timescales at which the center and leaves change are separated.

We find the following rates for the transitions of l_t and c_t:

Note that the top two rows imply

$$\phi = \frac{fl_t}{fl_t + N - l_t}. \tag{3.3.29}$$

Action	Rate (up to constants)
$c_{t+1} = c_t + 1$	fl_t
$c_{t+1} = c_t - 1$	$N - l_t$
$l_{t+1} = l_t + 1$	$\frac{f}{N}\phi\frac{N-l_t}{N}$
$l_{t+1} = l_t - 1$	$\frac{1-\phi}{N}\frac{l_t}{N}$

Table 3.1: Transition rates for the star.

Thus, looking at the ratio of l_{t+1} increasing over decreasing from l_t, just as we did in (3.3.19), we find

$$\frac{\frac{f}{N}\phi\frac{N-L}{N}}{\frac{1-\phi}{N}\frac{L}{N}} = \frac{f(N-L)\frac{fL}{fL+N-L}}{L\frac{N-L}{fL+N-L}} = f^2. \tag{3.3.30}$$

While this is only a heuristic and we cannot use a martingale and optional stopping to prove (3.3.28), it does suggest why we see this behavior. Now, we turn to a real proof.

THEOREM 3.11. *The fixation probability for the star of size $N+1$ is given by*

$$\rho_{\text{STAR}} = \frac{N^2\left(f^2-1\right)(N+f)+(Nf+1)\left(f^2-1\right)}{(N+1)(Nf+1)\left(f(N+f)-(Nf+1)\left(\frac{N+f}{Nf^2+f}\right)^N\right)}. \tag{3.3.31}$$

In particular ,

$$\lim_{N\to\infty}\rho_{\text{STAR}} = \left(1-f^{-2}\right)\mathbf{1}(f \geqslant 1). \tag{3.3.32}$$

To prove this we again use a martingale approach [4, 175, 176].

LEMMA 3.12. *Define*

$$g(x,y) = \left(\frac{fN+1}{fN+f^2}\right)^x\left(\frac{N+f}{Nf^2+f}\right)^y, \tag{3.3.33}$$

then $y_t := g(c_t, t_t)$ is a martingale with respect to $\mathbf{x}_t$.

PROOF. The martingale condition must be verified for all states, but the calculations are similar so we just show one.

We see

$$\begin{aligned}
\mathbb{E}\left[g(c_{t+1}, l_{t+1} | c_t = 0, l_t = i\right] &= g(1,i)\frac{fi}{F_t} + g(0,i-1)\frac{1}{F_t}\frac{i}{N} + g(0,i)\left(1 - \frac{fi}{F_t} - \frac{i}{F_t N}\right) \\
&= g(0,i) + \frac{fi}{F_t}\left(\frac{N+f}{Nf^2+f}\right)^i\left[\frac{fN+1}{fN+f^2} - 1\right] + \frac{i}{F_t N}\left(\frac{N+f}{Nf^2+f}\right)^i\left[\frac{Nf^2+f}{N+f} - 1\right] \\
&= g(0,i) + \frac{i}{F_t}\left(\frac{N+f}{Nf^2+f}\right)^i\left[f\frac{1-f^2}{fN+f^2} + \frac{1}{N}\frac{Nf^2-N}{N+f}\right] \\
&= g(0,i).
\end{aligned}$$

■

Proof of Theorem 3.11. Let T be the absorption time. Then when the mutant originates on a leaf, we see

$$g(0,1) = \mathbb{E}Y_0 = \mathbb{E}Y_T = \rho_l g(1,N) + (1-\rho_l)g(0,0). \tag{3.3.34}$$

Solving for ρ_l yields the fixation probability given that the mutant originates at a leaf,

$$\frac{g(0,1)-g(0,0)}{g(1,N)-g(0,0)} = \frac{N\left(f^2-1\right)(N+f)}{(Nf+1)\left(f(N+f)-(Nf+1)\left(\frac{N+f}{Nf^2+f}\right)^N\right)} \to \left(1-f^{-2}\right)\mathbf{1}(f \geqslant 1)\,. \tag{3.3.35}$$

Similarly, we find the fixation probability given that the mutant originates at the center, ρ_c,

$$\frac{g(1,0)-g(0,0)}{g(1,N)-g(0,0)} = \frac{f^2-1}{f(N+f)-(Nf+1)\left(\frac{N+f}{Nf^2+f}\right)^N} \to 0\,. \tag{3.3.36}$$

Thus, taking an expectation over the initial conditions, we find the expression for the fixation probability

$$\rho_{\text{STAR}} = \frac{N^2\left(f^2-1\right)(N+f) + (Nf+1)\left(f^2-1\right)}{(N+1)(Nf+1)\left(f(N+f)-(Nf+1)\left(\frac{N+f}{Nf^2+f}\right)^N\right)}. \tag{3.3.37}$$

Taking the limit as $N \to \infty$, we see (3.3.35) converges to $\left(1-f^{-2}\right)\mathbf{1}(f \geqslant 1)$ and (3.3.36) converges to 0. ■

3.3.1. Absorption times for evolutionary graph theory. As we have seen, population structure can distort the fixation probability, which is a statistic of primary interest, and shown that in some cases it leaves fixation probabilities unchanged. We now show that population structure can also distort absorption times and it can do so even when the fixation probability remains unchanged. While studying absorption times is interesting from a mathematical point of view, it also has relevant biological motivations. In particular, the molecular clock is defined as the rate at which neutral mutations fix in a population. Normally, this is expressed as

$$\varepsilon N \rho, \tag{3.3.38}$$

that is, the product of the rate at which neutral mutants enter the population, εN, and the fixation probability [177]. As we saw, for $f = 1$, $\rho = 1/N$, simplifying (3.3.38) to ε. However, in models where the fixation time also varies significantly the molecular clock is generalized to

$$\frac{1}{\frac{1}{\varepsilon N \rho} + t_1}, \tag{3.3.39}$$

which simplifies to $1/\left(1/\varepsilon + t_1\right)$ [177]. This means that when the fixation time is large it can dominate the molecular clock.

Another important and practical application of absorption times is how they bare on the efficiency of Monte Carlo simulations of evolutionary graph theory. The computational complexity of approximating the fixation probability has been extensively studied [174, 178–180]. A very natural algorithm to approximate the fixation probability, is to simulate the process, record outcomes, and then average these outcomes to estimate ρ. The efficiency of this algorithm is directly impacted by the expected absorption time.

Studying absorption times in structured populations is certainly more difficult than in unstructured populations. This is because often it is not possible to explicitly solve the recurrence equations as in (3.1.15) or to employ a diffusion approximation as in (3.2.7). However, we can perform some informal calculations when $f = 1$ and as $f \to \infty$ (just as we did in the proof of Theorem 3.4). We already found the absorption time for the well-mixed population in Theorem 3.4. So we start with the cycle defined in (3.3.8).

Let x_t be the number of individuals of type β. When $f = 1$, $y_t := x_t^2 - \frac{2t}{N}$ is a martingale with respect to $\mathbf{x}_t$:

$$\mathbb{E}\left[y_{t+1}|\mathbf{x}_t\right] = (x_t + 1)^2 \frac{2}{N}\frac{1}{2} + (x_t - 1)^2 \frac{2}{N}\frac{1}{2} + x_t^2\left(1 - 2\frac{2}{N}\frac{1}{2}\right) - \frac{2(t+1)}{N} = x_t^2 - \frac{2t}{N} = y_t. \tag{3.3.40}$$

Thus, let T be the absorption time. Then by Doob's optional stopping, we see

$$1 = \mathbb{E}y_0 = \mathbb{E}y_T = \rho N^2 + (1-\rho)\cdot 0 + \frac{2}{N}\mathbb{E}T. \tag{3.3.41}$$

Rearranging we get

$$t_1 = \frac{1}{2}N(N-1), \tag{3.3.42}$$

since $\rho = 1/N$. Note that when we measure t_1 in generational time, we find it takes on average $\mathcal{O}(N)$ generation for absorption, which is much longer than the $\mathcal{O}(\log N)$ generations required for well-mixed populations. So indeed, the rate of evolution in this case is

$$\frac{1}{\frac{1}{\varepsilon N\rho} + t_1} = \frac{1}{\frac{1}{\varepsilon N\frac{1}{N}} + \frac{1}{2}N(N-1)} = \Theta\left(\min\left\{\varepsilon, \frac{1}{N^2}\right\}\right). \tag{3.3.43}$$

When f is very large, it is very unlikely that the number of mutants ever decreases, thus we can decompose T as a sum of geometric random variables. Define $T_i := \min\{t : X_t = i\}$, then

$$\mathbb{E}T = \mathbb{E}T_N \approx \sum_{i=1}^{N-1}\mathbb{E}[T_{i+1}-T_i] \approx \sum_{i=1}^{N-1} i = \frac{N(N-1)}{2} = \Theta\left(N^2\right), \tag{3.3.44}$$

since $T_{i+1} - T_i \sim \text{Geo}\left(\frac{2}{i}\frac{1}{2}\right)$. This is again different than t_1 in a well-mixed population.

Now we turn to the tennis racket. First, let $f = 1$. We have to consider four things:

- Where does the mutant originate?
- How long does it take the root to reproduce and migrate to the well-mixed part?
- How many attempts does it take for the migrant to fix in the well-mixed part?
- What is the fixation time of the well-mixed part?

Putting this together, we find the expected absorption time to be approximately

$$\frac{1}{N}\cdot N\cdot N\cdot N\log N + \frac{N-1}{N}\left(\frac{1}{N}\cdot N\cdot N\cdot N\log N + \frac{N-1}{N}\cdot N\log N\right) = \Theta\left(N^2\log N\right). \tag{3.3.45}$$

Second, consider $f \to \infty$. If the mutant originates at the root, it migrates in one step and then it take approximately $N\log N$ steps (in expectation) to fix in the well-mixed part by Theorem 3.4. However, if the mutant originates in the well-mixed part, it fixes there, remains fixed, and cannot replace the root. So we never reach fixation or extinction of the mutant, hence

$$t_1 = \frac{1}{N}\left(1 + N\log N\right) + \frac{N-1}{N}\cdot\infty. \tag{3.3.46}$$

Finally, consider the star when $f = 1$. Let t_i be the expected absorption time starting from i mutants on the leaves. Here we ignore the time taken for the center to take the correct value and consider absorption when the leaves have 0 or N mutants. Let ϕ be the probability that the center is a mutant, as we saw before $\phi = i/N$. Then

$$P_i^+ \propto \frac{1}{N}\frac{i}{N}\frac{N-i}{N} \tag{3.3.47}$$

and

$$P_i^- \propto \frac{1}{N}\frac{N-i}{N}\frac{i}{N}. \tag{3.3.48}$$

Now we can write down a similar recurrence equation for t_i as we did in the well-mixed case in Equation (3.1.14):

$$t_i = 1 + \frac{1}{N}\frac{i}{N}\frac{N-i}{N}t_{i+1} + \frac{1}{N}\frac{N-i}{N}\frac{i}{N}t_{i-1} + \left(1 - \frac{1}{N}\frac{i}{N}\frac{N-i}{N} - \frac{1}{N}\frac{N-i}{N}\frac{i}{N}\right)t_i, \tag{3.3.49}$$

which implies $t_i^{\text{star}} = Nt_i^{\text{WM}}$.

When $f \to \infty$, in a similar way to before

$$T_{i+1} - T_i \sim \text{Geo}\left(\frac{1}{i+1}\frac{N-i}{N}\right), \tag{3.3.50}$$

thus

$$t_1 = N\sum_{i=1}^{N-1}\frac{i+1}{N-i} \approx N^2\log N. \tag{3.3.51}$$

Where the sum is approximated with an integral.

Graph	$t_1\ (r=1)$	$t_1\ (r\to\infty)$
Well-mixed	$N\log N$	$N\log N$
Cycle	$N^2/2$	$N^2/2$
Star	$N^2\log N$	$N^2\log N$
Tennis racket	$N^2\log N$	∞

Table 3.2: We summarize the calculations of absorption times from this section up to leading orders.

We have studied absorption time for several special populations structures, but some more general results are known. In [178] a supermartingale is derived from a potential function in the case of simple, undirected graphs. Using this supermartingale, they find that

$$t_1 \leqslant \frac{1}{1-f} N^3 \tag{3.3.52}$$

when $f < 1$,

$$t_1 \leqslant \frac{f}{f-1} N^4 \tag{3.3.53}$$

when $f > 1$, and

$$t_1 \leqslant N^4 \left(\sum_{i=1}^{N} \frac{1}{d_i} \right)^2 \tag{3.3.54}$$

when $f = 1$, where d_i is the degree of vertex i. In [180], the absorption time of simple, directed, regular graphs is proved to be $\Omega(N \log N)$ and $\mathcal{O}(N^2)$. They also define an infinite family of simple, directed graphs for which the expected absorption time is exponential in N, showing that in general the absorption time can be very long. These results have immediate implication for Monte Carlo algorithms for approximating ρ; [174] builds on these results and constructs even more efficient algorithms. Complexity questions have also been considered for evolutionary graph theory [179], but for a slightly different model than the birth-death Moran-type model we have defined here. For the proofs some non-monotonicity is required, so some sort of frequency dependence on fitness is assumed—the exact complexity of the model considered in this section remains unknown for directed, weighted graphs.

3.4. General evolutionary processes

In each of the above models, the process was updated as follows: Some subset R of the population is selected to be replaced. Each individual i in this subset, is replaced by the offspring of some other individual j. We denote this event $r(i) = j$. Precisely, there is some probability distribution $\mathbb{D}$ over the set of (R, r) such that $R \subseteq [\![N]\!]$ and $r : R \to [\![N]\!]$. The process is then updated by sampling from this distribution, which can depend on the current state of the process, that is,

$$\mathbb{P}\left\{\mathbf{x}_{t+1} = \mathbf{x}' | \mathbf{x}_t = \mathbf{x}\right\} = \sum_{(R,r)} \mathbb{D}_{\mathbf{x}}(R, r) \prod_{i \in R} \mathbf{1}\left(\mathbf{x}'(i) = \mathbf{x}(r(i))\right) \prod_{i \notin R} \mathbf{1}\left(\mathbf{x}'(i) = \mathbf{x}(i)\right). \tag{3.4.1}$$

That is, given $\mathbf{x}_t$ and a sample (R, r), the new state is $\mathbf{x}_{t+1}(i) = \mathbf{x}_t(r(i))$ for $i \in R$ and $\mathbf{x}_{t+1}(i) = \mathbf{x}_t(i)$ for $i \notin R$.

Looking at the models we studied in the previous sections, we can specify $\mathbb{D}$ in those cases. First, it is useful to introduce some notation. Previously, we limited ourselves to two types, α and β, but we gave several example of genotype spaces with many types in Chapter 2. In Section 2.6, we associated fitnesses with each genotype in the genotype space—denoted by $\mathcal{F}(\alpha)$ for a genotype α. The population structure and update rule are fixed, in that the replacement events at different points in time are chosen independently from the same distribution. This function is key to defining the distribution $\mathbb{D}$ for many models. As before, we use

$$F_{\mathbf{x}} := \sum_{k=1}^{N} \mathcal{F}(\mathbf{x}(k)) \tag{3.4.2}$$

to denote the average fitness of a population $\mathbf{x}$ and F_t for the population $\mathbf{x}_t$ at time t.

Definition 3.13 (Moran process in a well-mixed population with birth-death updating). *The update distribution is given by*

$$\mathbb{D}_{\mathbf{x}}(R,r) = \begin{cases} \frac{1}{N}\frac{\mathcal{F}(\mathbf{x}(r(i)))}{F_{\mathbf{x}}} & \text{if } R=\{i\} \\ 0 & \text{otherwise} \end{cases}, \tag{3.4.3}$$

where $\mathcal{F}(\mathbf{x}(k))$ is just the fitness of individual i.

Definition 3.14 (Wright-Fisher process in a well-mixed population). *The update distribution is given by*

$$\mathbb{D}_{\mathbf{x}}(R,r) = \begin{cases} F_{\mathbf{x}}^{-N} \prod_{i=1}^{N} \mathcal{F}\left(\mathbf{x}\left(r(i)\right)\right) & \text{if } R=[\![N]\!] \\ 0 & \text{otherwise} \end{cases}. \tag{3.4.4}$$

where $\mathcal{F}(\mathbf{x}(k))$ is just the fitness of individual i.

Definition 3.15 (Moran process on a graph with birth-death updating). *The update distribution is given by*

$$\mathbb{D}_{\mathbf{x}}(R,r) = \begin{cases} \frac{\mathcal{F}(\mathbf{x}(r(i)))W(r(i),i)}{F_{\mathbf{x}}} & \text{if } R=\{i\} \\ 0 & \text{otherwise} \end{cases}. \tag{3.4.5}$$

where $G = ([\![N]\!], W)$ is a graph representing a population structure.

The fitness dependence of the transition probabilities of the process are encoded in the dependence of the distribution $\mathbb{D}_{\mathbf{x}}$ on the current state $\mathbf{x}$, as we saw in the examples above. The distribution $\mathbb{D}_{\mathbf{x}}$ also incorporates any effects of populations structure. Using this formalism, we can give a simple and general definition of neutrality. A process is neutral if the distribution $\mathbb{D}_{\mathbf{x}}$ does not depend on the current state $\mathbf{x}$. We state this formally in Definition 3.17.

3.4.1. Adding mutation. In the examples we have considered so far, the type of the offspring is always identical to the type of the parent. Mutation breaks this assumption; we explored mutation without evolutionary dynamics extensively in Chapter 2.

When we define the dynamics from the distribution $\mathbb{D}$ in (3.4.1), we assumed no mutation. To add mutation, we write the following

$$\mathbb{P}\left\{\mathbf{x}_{t+1} = \mathbf{x}' | \mathbf{x}_t = \mathbf{x}\right\} = \sum_{(R,r)} \mathbb{D}_{\mathbf{x}}(R,r) \prod_{i \in R} \mathcal{M}\left(\mathbf{x}(r(i)), \mathbf{x}'(i)\right) \prod_{i \notin R} \mathbf{1}\left(\mathbf{x}'(i) = \mathbf{x}(i)\right). \tag{3.4.6}$$

So in general, we have some space of genotypes Γ and a transition kernel $\mathcal{M}$ that describes the probability of each type of mutation. Specifically, when the number of genotype types is finite, mutations are described by a Markov chain $(\Gamma, \mathcal{M})$, that we defined as the mutation process in Definition 2.2. Note that the process without mutation can be recovered by setting $\mathcal{M}(\alpha, \beta) = \mathbf{1}(\alpha, \beta)$ for all $\alpha, \beta \in \Gamma$. This leads us to a general way to define a stochastic evolutionary process in a finite population of fixed size.

Definition 3.16 (General evolutionary process). *A general evolutionary process is a triple* $\left(\Gamma, \mathcal{M}, (\mathbb{D}_{\mathbf{x}})_{\mathbf{x}}\right)$, *where* Γ *and* $\mathcal{M}$ *form a genotype space as defined in Definition 2.1, and* $\mathbb{D}_{\mathbf{x}}$ *is a distribution on the set of* (R, r) *such that* $R \subseteq [\![N]\!]$ *and* $r : R \to [\![N]\!]$ *for each* $\mathbf{x} \in \Gamma^N$. *The process is a Markov chain with state space* Γ^N *and its value at time* t *is denoted by* $\mathbf{x}_t$. *The transition probabilities of the process are given by Equation* (3.4.6).

The way in which mutations are incorporated here is essentially the same as they are in [25], except that here Γ and $\mathcal{M}$ are much more general and mutations need not be symmetric with respect to the genotypes.

Definition 3.17 (Neutral evolutionary process). *An evolutionary process is called neutral if*

$$\mathbb{D}_{\mathbf{x}} = \mathbb{D}_{\mathbf{x}'} \tag{3.4.7}$$

for all $\mathbf{x}, \mathbf{x}' \in \Gamma^N$.

In the next chapter, we develop many results for neutral evolutionary processes in this general setting.

4

Neutral evolution

In this chapter, we study a general setting of neutral evolution that was defined in Definition 3.17. This definition of neutrality results in a certain independence between the evolutionary dynamics and the mutation process. This independence enables lots of analysis that is not possible in the general case. However, this simplifying assumption does not make the theory trivial—there is a rich literature on neutral evolutionary theory [26–28, 181]. We consider population of finite size and that may have spatial structure. Many different forms of genotype space are covered, including those discussed in Chapter 2, and both discrete and continuous genotype spaces.

Under minimal assumptions, we show that the frequencies of different genotypes in the stationary distribution of the evolutionary process are independent of population size, spatial structure, and evolutionary update rule. Specifically, these frequencies in the stationary distribution are given by the stationary distribution of the mutation process. Thus, we demonstrate that the stationary frequencies of a neutral evolutionary process can always be calculated by evaluating a simple stochastic process describing a population of size one. We then characterize the convergence of the mutation process along the various lineages in terms of demographic variables of the evolutionary process. After that we consider the mixing time of the evolutionary process as a whole. Finally, we examine correlations between individuals in the stationary distribution in some special cases—in particular, whether the population is localized in genotype space.

As we have seen by considering the Moran process and the Wright-Fisher process, evolutionary dynamics asks how the details of evolutionary processes affect long-run and stationary statistics of our stochastic models of evolution. For example, microscopic quantities including mutation rates, fitness differences between competing types, and population structure can have profound effects on these statistics [123,169,182–186], but the precise relationships are often difficult to calculate. A number of recent studies have addressed this issue using neutral populations (those without selective differences between the genotypes) as a baseline comparison to determine the effect of a property like population structure on natural selection [25, 187–197]. Our main results in this chapter, study the underlying neutral process used in this comparison and derive the long-term genotype frequencies for any mutation process, population size, structure, and evolutionary update rule.

To motivate the discussion, consider the Moran process with a mutation process $(\Gamma, \mathcal{M})$. We assume that the process is neutral, which for the Moran process means that $\mathcal{F}(\alpha) = 1$ for all $\alpha \in \Gamma$. When evolution is neutral, there is a meaningful comparison between the evolutionary process and the mutation process since then neither one involves different types reproducing at different rates. In effect, the mutation process is just an evolutionary process in a population of size $N = 1$. A natural question to ask is whether population size, structure, or update rule affect average genotype frequency relative to the mutation process: is the average frequency of genotype α in a population of any size, $N > 1$, different from that of a population of size $N = 1$?

Under neutral drift, competing types have the same reproductive rates. Suppose that $|\Gamma| = n$ and that

$$\mathcal{M}(\alpha, \beta) = \varepsilon \frac{1}{n} \tag{4.0.1}$$

for all $\beta \neq \alpha$. Specifically, with probability ε a mutation to a uniformly random type occurs. This kind of mutation, which is common in evolutionary game theory, is symmetric in the sense that it acts on all competing types in the same way. Thus, we should expect that each genotype is equally abundant in the stationary distribution of the evolutionary process. As we saw in Chapter 2, with mutations of this form, all types are equally abundant in the stationary distribution of the mutation process [188, 189]. Thus, the frequency of the genotypes is the same in the stationary distribution of the mutation process and the evolutionary process. Here, we extend this result to any mutation process and any neutral evolutionary process. In particular, we show that the average frequency of a type in a population of any size is the same as that in a population of size $N = 1$. Consequently, reproduction, dispersal patterns, and population size have no effect on the average frequency of a particular genotype.

In some ways, this result is not surprising. When there are no selective differences between the genotypes, there is evidently nothing in the population that drives one genotype to a higher average frequency in a population than in the original mutation process itself. However, this heuristic intuition is not a proof. Given the increasing interest in using neutral frequency as a baseline measure to understand strategy selection, we provide a general proof here. Our results apply to not only any population structure but also to any genotype space, Γ.

Our focus here is on evolving populations of fixed size and structure. These assumptions are not strictly necessary, but they do make the notation more convenient since an evolutionary process can then be described by a probability distribution over simple replacement events [25]. Extinction also becomes a possibility when the population size fluctuates, which can make analyzing such a process more complicated. Nonetheless, the intuition from our analysis here carries over to other situations: as long as an individual arises on a lineage with sufficiently many prior birth events, the mutation process along this lineage will converge to its stationary distribution, π. If all individuals in the population have this property sufficiently far into the future, then π must describe the long-term average genotype frequencies in the population.

In and of themselves, models with neutral mutations form an important part of evolutionary theory [27, 198–202]. Their utility, however, goes beyond describing systems in which there are truly no selective differences between types. Models with weak selection, which assume that any reproductive differences between types are small, may be viewed as perturbations of neutral processes. Studying neutral models is therefore relevant to understanding evolutionary dynamics even with selection. We have concentrated on the lineages generated by neutral evolution, which exhibit unsurprising behavior with respect to genotype distributions but can have interesting effects on mixing times. Selection can affect these dynamics in complicated ways, and how it changes both marginal distributions and mixing times is an interesting question for future research.

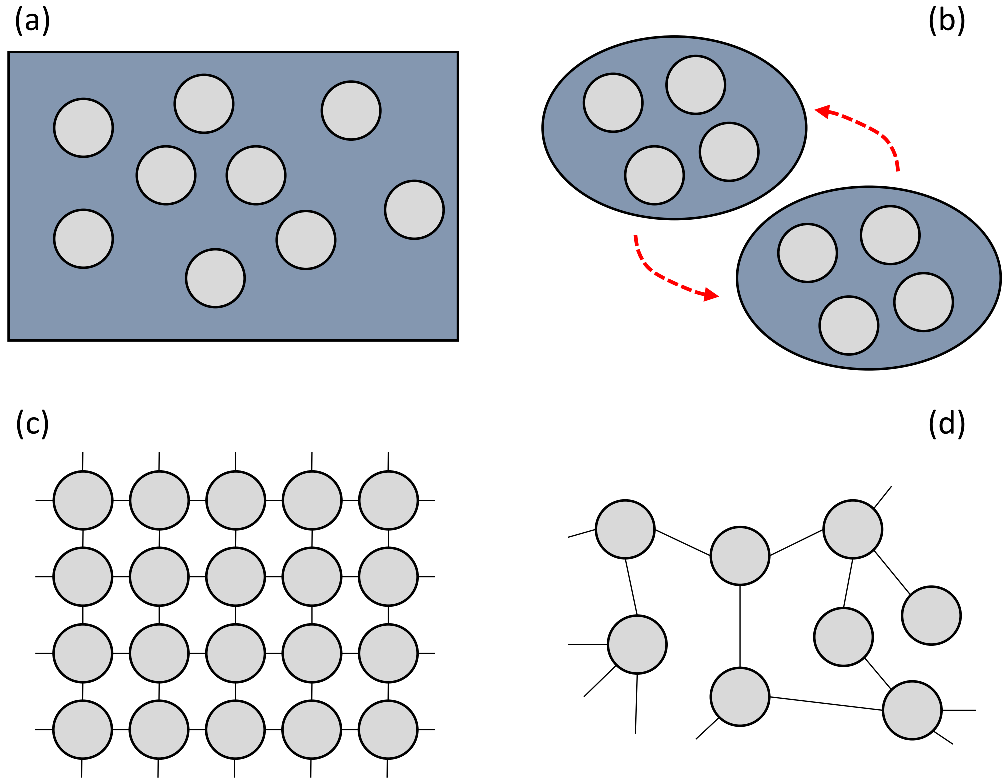

Figure 4.1: Four examples of evolving populations, ordered by increasing complexity of the spatial structure. In (a), the population is unstructured. (b) depicts a subdivided (or "deme-structured") population that consists of unstructured subpopulations (blue) with migration between them (red arrows). (c) shows a regular grid (square lattice) in which all players have exactly four neighbors. (d) illustrates a heterogeneous graph-structured population. Each individual resides on the vertex of a graph, and links indicate who is a neighbor of whom. The number of neighbors can vary from individual to individual, which results in structural asymmetries. In general, if the graph indicates an offspring-dispersal structure, then $\mathbb{D}(R,r)>0$ only if $\alpha(i)$ is a neighbor of i whenever $i\in R$.

4.1. Coherence assumption

This assumption formalizes the notion that the population evolves as a coherent unit, and that every individual can produce a lineage that takes over the entire population. Importantly, it does not imply that the population structure is trivial. Many interesting population structures, including heterogeneous graphs, sets, and subdivided populations with migration, satisfy the unity condition; we refer the reader to [25] for further examples. In Example 4.2 below, we give an example of a population that does not satisfy the unity condition.

For each $i\in [\![N]\!]$, there exists a sequence of replacement events, $\{(R_k,r_k)\}_{k=1}^{\ell}$, such that $\mathbb{D}\left((R_k,r_k)\right)>0$ for each

$k \in [\![\ell]\!]$ and for each j

$$\tilde{r}_1 \circ \cdots \circ \tilde{r}_\ell \left(j\right) = i, \tag{4.1.1}$$

where the mapping $\tilde{r}_k : \{1, \ldots, N\} \to \{1, \ldots, N\}$ is defined by $\tilde{r}_k\left(m\right) = r_k\left(m\right)$ for $m \in R_k$ and $\tilde{r}_k\left(m\right) = m$ for $m \notin R_k$. This assumption states that for any pair of individuals i and j in the population, there is a chain of replacement events with positive probability such that individual i replaces individual j in a finite number of steps.

4.2. Stationary distribution

The distribution $\{\mathbb{D}(R,r)\}_{(R,r)}$ gives rise to several useful demographic variables [25], namely

$$e_{ij} := \sum_{\substack{(R,r) \\ j \in R,\ r(j)=i}} \mathbb{D}(R,r); \qquad b_i := \sum_{j=1}^{N} e_{ij}; \qquad d_i := \sum_{j=1}^{N} e_{ji}. \tag{4.2.1}$$

e_{ij} is the probability that i transmits its offspring to j; b_i is the birth rate of i; and d_i is the death rate of i.

Theorem 4.1. *If the unity condition* (4.1.1) *holds, then* $(\mathbf{x}_t)_{t \geqslant 0}$ *is irreducible and aperiodic whenever the underlying mutation process* $(\Gamma, \mathcal{M})$ *is irreducible and aperiodic.*

Proof. Consider two states, $\mathbf{x}, \mathbf{y} \in \Gamma^N$. If $(\Gamma, \mathcal{M})$ is ergodic, then there exists m_0 such that $\mathcal{M}^m(\alpha, \beta) > 0$ whenever $\alpha, \beta \in \Gamma$ and $m \geqslant m_0$. By the unity condition, we can find an ordered sequence of replacement events $\left(R_1, r_1\right), \ldots, \left(R_{m_0}, r_{m_0}\right)$ with $\mathbb{D}(R_k, \alpha_k) > 0$ for each k, together with a collection $\left(i_0, i_1, \ldots, i_{m_0}\right)$ such that $r_k\left(i_{k-1}\right) = i_k$ for each $k = 1, \ldots, m_0$. After starting in state $\mathbf{x}$, and given this sequence of replacement events, the probability that individual i_0 has type α is $\mathcal{M}^{m_0}(\mathbf{x}(i_0), \alpha) > 0$. By the unity condition, i_0 can propagate its offspring to all other nodes in a finite number of steps; together with the fact that $\mathcal{M}^m$ is positive whenever $m \geqslant m_0$, we see that there is a positive probability of reaching state $\mathbf{y}$. Thus, $(\mathbf{x}_t)_{t \geqslant 0}$ is irreducible. Aperiodicity of $(\mathbf{x}_t)_{t \geqslant 0}$ follows from essentially the same argument because, after a similarly chosen sequence of replacements, the probability of staying in the state in which the chain started is positive. ■

If the unity condition (4.1.1) does not hold, then Theorem 4.1 does not necessarily hold, as we see in the following example:

Example 4.2 (Line graph). *As an example of a process that does not satisfy the unity condition, consider a population arranged on a line. If reproduction and replacement flows in only one direction, then individual 1, the player at the start of the line, is never replaced. Therefore, for each* $\alpha \in \Gamma$*, there exists a stationary distribution,* $\mu^{(\alpha)}$*, for* $\{\mathbf{x}_t\}_{T\geqslant 0}$ *with* $\mathbb{P}_{\mathbf{x}\sim\mu^{(\alpha)}}\{\mathbf{x}(1)=\alpha\}=1$*. In particular, there is more than one stationary distribution even when the mutation process is irreducible and aperiodic, so Theorem 4.1 need not hold when the unity condition is not satisfied.*

When the unity condition holds and the mutation process is irreducible and aperiodic, Theorem 4.1 implies that the evolutionary process has a unique stationary distribution, μ. In the next section, we show that if π is the stationary distribution of the mutation process and μ the stationary distribution of the evolutionary process, then

$$\mathbb{P}_{\mathbf{x}\sim\mu}\{\mathbf{x}(i)=\alpha\}=\pi_\alpha \tag{4.2.2}$$

for each $\alpha \in \Gamma$. When the evolutionary process has more than one stationary distribution, it need not be the case that every such distribution satisfies Equation (4.2.2), which we illustrate using an irreducible mutation chain (with a unique stationary distribution) that is periodic:

Example 4.3 (Irreducible but periodic mutation process). *Suppose that* $\Gamma=\{\alpha,\beta\}$ *and let* $\mathcal{M}$ *be the matrix*

$$\mathcal{M}:=\begin{pmatrix}0 & 1\\ 1 & 0\end{pmatrix}. \tag{4.2.3}$$

the mutation process is irreducible but periodic with period 2*, and its (unique) stationary distribution is* $\pi=(1/2,1/2)$*. Consider a birth-death process on a star graph (see Equation* (3.3.27)*). Each player is chosen uniformly-at-random to reproduce. If a peripheral player reproduces, their offspring is subjected to the mutation operator and propagated to the central node. If the player at the central node reproduces, then this player propagates an offspring to every peripheral location (and, again, each offspring is subjected to the mutation operator). It is easy to see that the state with type* α *at the central node and type* β *at all of the peripheral nodes is stationary (see Figure 4.2). For this stationary distribution,* $\frac{1}{N}\sum_{i=1}^{N}\mathbb{P}_{\mathbf{x}\sim\mu}\left(\mathbf{x}(i)=\alpha\right)=1/N$*, which is not equal to* $\pi(\alpha)=1/2$ *when* $N>2$*.*

The following result is the main tool we need to show that population size and structure do not influence stationary genotype frequencies. The essence of this result is that for any $L\geqslant 0$, if one looks sufficiently far into the future, then every individual in the population will be able to have a lineage of length at least L:

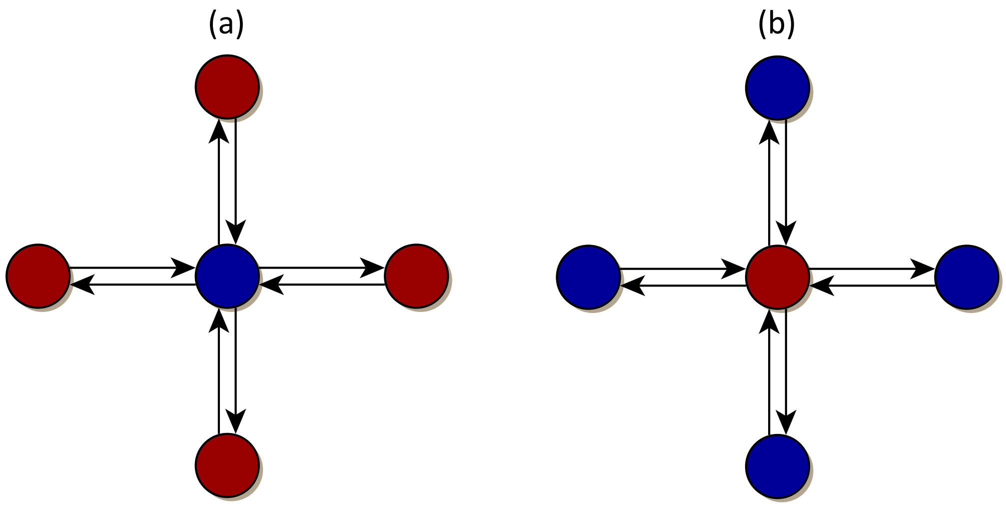

Figure 4.2: A birth-death process on a star graph of size $N = 5$. When a player reproduces, the offspring has the opposite type of the parent (i.e. if the parent is blue, the child is red; if the parent is red, the child is blue). When a peripheral individual reproduces, the central individual dies and is replaced by the offspring. If the central individual reproduces, then all four peripheral players die and are replaced by offspring of the central player. (a) and (b) both give stationary states, which are unaffected by the evolutionary process. The frequency of blue is 1/5 in (a) and 4/5 in (b), while the frequency of blue in the stationary distribution of the mutation process (i.e. in a population of size $N = 1$) is 1/2. Therefore, even though the mutation process is irreducible, the evolutionary process has multiple stationary distributions due to the periodicity of $\mathcal{M}$, and these stationary distributions need not all exhibit the same trait frequencies as the original mutation process. However, there does exist a stationary distribution for the birth-death process with trait frequency 1/2 for each of blue and red, namely the distribution that assigns probability 1/2 to state (a) and 1/2 to state (b).

Lemma 4.4. *Let $B_t^{(i)}$ be the number of birth events after t updates in the lineage leading to individual i (a random variable). Then, for any $i \in [\![N]\!]$ and any $L \geqslant 0$, we have*

$$\lim_{t\to\infty} \mathbb{P}\{B_t^{(i)} \geqslant L\} = 1. \tag{4.2.4}$$

Proof. The probability that i is replaced in any given update step (i.e. the death rate of i) is d_i (Equation 4.2.1), which is the same at every update since the process is neutral. Let $d_* := \min_{1\leqslant i\leqslant N} d_i$ and $d^* := \max_{1\leqslant i\leqslant N} d_i$. Let $B_t^{(i)}$ be the number of birth events after t updates in the lineage leading to individual i.

Obviously, $B_*^t := \min_i B_t^{(i)} \leqslant B_t^{(i)}$. Note that when individual j is replaced by individual i at time t, we have $B_j^{t+1} = B_i^t + 1$. Moreover, B_*^t is nondecreasing with t, unlike the individual $B_t^{(i)}$. Thus, if B_*^t is achieved in k individuals, $i_1, \ldots, i_k$, there is a probability of at least d_* that for one i_j, we have $B_{i_j}^{t+1} > B_{i_j}^t$. Thus, in k steps the probability that $B_*^{t+k} > B_*^t$ is lower bounded by $d_*^k \geqslant d_*^N > 0$. Thus,

$$\mathbb{P}\left[B_*^{tN} < L\right] \leqslant \sum_{k=0}^{L-1} \binom{t}{k} (1 - d_*^N)^{t-k} \to 0 \tag{4.2.5}$$

as $t \to \infty$. ■

Now we use this simple fact to derive the long-run genotype frequencies.

While the assumption that Γ is finite is reasonable in many cases, there are also scenarios in which one would like to consider continuous genotype spaces (see Chapter 2). Here we let Γ be a measurable space, which contains as special cases finite, denumerable, and uncountably infinite genotype spaces.

Throughout the rest of this section, we denote by $\mathbb{P}_\nu$ and $\mathbb{E}_\nu$ the distribution and expectation, respectively, of a random variable that depends on another distribution, ν. Such as the case when ν is the initial distribution of a Markov chain and the random variable under consideration is the time-t state of the chain. As a shorthand, we write $\mathbb{P}_{\mathbf{x}}$ and $\mathbb{E}_{\mathbf{x}}$ for $\mathbb{P}_{\delta_{\mathbf{x}}}$ and $\mathbb{E}_{\delta_{\mathbf{x}}}$, where $\delta_{\mathbf{x}}$ is a delta mass at α so that the chain must start at $\mathbf{x}$.

When dealing with a general genotype space, we can no longer necessarily represent a mutation chain by a transition matrix. Instead, such a mutation process is described by a transition kernel. Let $\sigma(\Gamma)$ be a σ-algebra of subsets on Γ and denote by $\Delta(\Gamma)$ the space of probability measures on Γ. A mutation process on Γ, $\{\alpha_t\}_{t\geqslant 0}$, is then defined by a Markov kernel, $\mathcal{M} : S \to \Delta(S)$, where for $\alpha \in \Gamma$ and $E \in \sigma(\Gamma)$, $\mathcal{M}(\alpha, E)$ is the probability that the chain is in E after being in state α. If Γ is finite and the transition matrix for the mutation process is $\tilde{\mathcal{M}}$, then $\tilde{\mathcal{M}}(\alpha, \beta) = \mathcal{M}(\alpha, \{\beta\})$ for each $\alpha, \beta \in \Gamma$. To extend the notion of ergodicity to a Markov chain on a general state space, one needs the notion of a Harris chain, which we recall from [203]:

Definition 4.5. *A Markov chain, $\{\alpha_t\}_{t\geqslant 0}$, on Γ with kernel $\mathcal{M}$ is a Harris chain if there exist $A, B \in \sigma(\Gamma)$, $\varepsilon > 0$, a function $q : A \times B \to \mathbb{R}$ with $q(\alpha, \beta) \geqslant \varepsilon$ for each $\alpha \in A$ and $\beta \in B$, and $\rho \in \Delta(B)$ such that*

(i) $\mathbb{P}_s\{\tau_A < \infty\} > 0$ for each $\alpha \in \Gamma$, where $\tau_A = \inf\{t \geqslant 0 \mid \alpha_t \in A\}$;

(ii) $\mathcal{M}(\alpha, C) \geqslant \int_{\beta\in C} q(\alpha, \beta)\, d\rho(\beta)$ for each $\alpha \in A$ and $C \in \sigma(B)$.

Recurrence and aperiodicity are defined in the same way for Harris chains as they are for Markov chains on a finite state space. We refer to a recurrent, aperiodic Harris chain as ergodic. The key result we need is the following:

if $\mathcal{M}$ defines an ergodic Harris chain with stationary distribution $\pi \in \Delta(\Gamma)$, then

$$\lim_{t\to\infty} \sup_{E\in\sigma(\Gamma)} \left|\kappa^t(\alpha, E) - \pi(E)\right| = 0 \tag{4.2.6}$$

whenever $\alpha \in \Gamma$ satisfies $\mathbb{P}_\alpha\{R_A < \infty\} = 1$, where $R_A = \inf\{t \geqslant 1 \mid \alpha_t \in A\}$. In other words, $\mathcal{M}^t(\alpha, \cdot)$, the t-step transition kernel when starting from α, converges in total variation to π (see [203]).

Our main result with respect to genotype frequency can be stated succinctly as follows:

THEOREM 4.6. *For any $i \in [\![N]\!]$, initial distribution μ_0, and $E \in \sigma(\Gamma)$, we have*

$$\lim_{t\to\infty} \mathbb{P}_{\mu_0}\{\mathbf{x}_t(i) \in E\} = \pi(E). \tag{4.2.7}$$

PROOF. Since the mutation process $\{\alpha_t\}_{t\geqslant 0}$ is ergodic, for each $\alpha \in \Gamma$ and $\varepsilon > 0$, there exists $L \geqslant 0$ such that whenever $\ell \geqslant L$,

$$\sup_{E\in\sigma(\Gamma)} \left|\mathbb{P}_\alpha\{\alpha_t \in E\} - \pi(E)\right| < \varepsilon. \tag{4.2.8}$$

Thus, for each $E \in \sigma(\Gamma)$ and $\ell \geqslant L$, we have $\pi(E) - \varepsilon < \mathbb{P}_\alpha\{\alpha_t \in E\} < \pi(E) + \varepsilon$, which gives

$$(\pi(E) - \varepsilon)\,\mathbb{P}\{B_t^{(i)} \geqslant L\} < \sum_{\ell=L}^{T} \mathbb{P}_\alpha\{\alpha_\ell \in E\}\mathbb{P}\{B_t^{(i)} = \ell\} < (\pi(E) + \varepsilon)\,\mathbb{P}\{B_t^{(i)} \geqslant L\}. \tag{4.2.9}$$

It then follows from the limit $\lim_{t\to\infty} \mathbb{P}\{B_t^{(i)} \geqslant L\} = 1$ (see Lemma 4.4) that

$$\begin{aligned}
\lim_{t\to\infty} \mathbb{P}_{\mu_0}\{\mathbf{x}_t(i) \in E\} &= \lim_{t\to\infty} \sum_{\alpha\in\Gamma} \sum_{\ell=0}^{t} \mathbb{P}_\alpha\{\alpha_\ell \in E\}\mathbb{P}\{B_t^{(i)} = \ell\}\mathbb{P}_{\mathbf{x}\sim\mu_0}\{\mathbf{x}(i) = \alpha\} \\
&= \sum_{\alpha\in\Gamma} \sum_{\ell=0}^{L-1} \mathbb{P}_\alpha\{\alpha_\ell \in E\}\mathbb{P}_{\mathbf{x}\sim\mu_0}\{\mathbf{x}(i) = \alpha\} \lim_{t\to\infty} \mathbb{P}\{B_t^{(i)} = \ell\} \\
&\quad + \lim_{t\to\infty} \sum_{\alpha\in\Gamma} \sum_{\ell=L}^{t} \mathbb{P}_\alpha\{\alpha_\ell \in E\}\mathbb{P}\{B_t^{(i)} = \ell\}\mathbb{P}_{\mathbf{x}\sim\mu_0}\{\mathbf{x}(i) = \alpha\} \\
&= \sum_{\alpha\in\Gamma} \mathbb{P}_{\mathbf{x}\sim\mu_0}\{\mathbf{x}(i) = \alpha\} \lim_{T\to\infty} \sum_{\ell=L}^{t} \mathbb{P}_s\{\alpha_\ell \in E\}\mathbb{P}\{B_t^{(i)} = \ell\} \\
&\in (\pi(E) - \varepsilon, \pi(E) + \varepsilon),
\end{aligned} \tag{4.2.10}$$

which gives $\lim_{t\to\infty} \mathbb{P}_{\mu_0}\left\{\mathbf{x}_t(i) \in E\right\} = \pi(E)$ since ε was arbitrary. ■

The frequency of individuals whose genotype lies in $E \in \sigma(\Gamma)$ at time t is

$$\mathbb{E}_{\mu_0}\left[\frac{1}{N}\sum_{i=1}^{N} \mathbf{1}_{\mathbf{x}_t(i)(t)}(E)\right] = \frac{1}{N}\sum_{i=1}^{N} \mathbb{P}_{\mu_0}\left\{\mathbf{x}_t(i) \in E\right\}, \tag{4.2.11}$$

which approaches $\pi(E)$ as $t \to \infty$ by Theorem 4.6; therefore, π gives the long-term genotype frequencies. In the setting of Theorem 4.1, if $\{\mathbf{x}_t\}_{t\geqslant 0}$ has a unique stationary distribution, μ, then $\mathbb{P}_{\mathbf{x}\sim\mu}\left\{\mathbf{x}(i) = \alpha\right\} = \pi(\alpha)$ for each $i = [\![N]\!]$ and each $\alpha \in \Gamma$, that is, the marginal distribution of every individual i is exactly π.

Although our focus has been on haploid individuals, we note that Theorem 4.6 can be extended to populations with diploid or haplodiploid genetics, with either sexual or asexual reproduction (or a combination of both) as long as there is no recombination within the locus in question. The idea of this extension is to let $i \in [\![N]\!]$ index *genetic sites* rather than individuals. These genetic sites are distributed among individuals, so that each individual has a number of genetic sites equal to its ploidy. Each genetic site i contains a single allele $\alpha \in \Gamma$, where Γ represents the set of possible alleles on a particular locus. During transitions, the alleles in a subset R of genetic sites are replaced by (possibly mutated) copies of the alleles in other genetic sites, as determined by the chosen replacement event (R, r). The probability distribution over replacement events, $\{\mathbb{D}(R, r)\}_{(R,r)}$, encodes all necessary information about ploidy, sexes, and mating.

Formally, the framework based on genetic sites is mathematically equivalent to the framework based on haploid individuals. All of our results therefore carry over to sexually-reproducing populations without any additional mathematical assumptions. In applying these results to sexually-reproducing populations, there is, however, an implicit *biological* assumption that there is no recombination. That is, each allele in a new offspring is a copy of a single allele in one parent, not a mixture of two (or more) alleles, which is reasonable if the locus represented by Γ is small enough that linkage within the locus can be assumed complete.

Even if the mutation process does not have a unique stationary distribution (such as when there are multiple absorbing states), the proof of Theorem 4.6 can still be used to derive the marginal distributions of every individual when the starting condition is monomorphic. Suppose that every individual in the population initially has genotype $\alpha \in \Gamma$. There will eventually be enough births along every lineage for the mutation process to reach its limiting distribution (provided the limit exists). Moreover, since the lineage leading to i at time t can start at any $j \in [\![N]\!]$

at time 0, we know the initial condition of this chain (α) provided the population starts out monomorphic. We state this observation as a proposition:

Proposition 4.7. *If $\alpha \in \Gamma$, $E \in \sigma(\Gamma)$, and $i \in [\![N]\!]$, then, provided the limits exist,*

$$\lim_{t\to\infty} \mathbb{P}_{(\alpha,\dots,\alpha)} \left\{\mathbf{x}_t(i) \in E\right\} = \lim_{t\to\infty} \mathbb{P}_{\alpha} \left\{\alpha_t \in E\right\}. \tag{4.2.12}$$

The result on stationary genotype frequencies is well-established in the special case of two competing types on a homogeneous population structure [204, 205]. Similarly, in population genetics, it has been noted that the so-called "common ancestor" process is the same as the mutation process when there is no selection [206, 207]. Comparisons between neutrally-evolving populations and their mutation processes have also been used to show that evolution favors genotypes that are robust against mutation [208]. Given the coupling of birth and mutation, how neutral evolution affects genotype frequencies is a natural question to ask. Under mild assumptions, we have seen that neutral evolution does not change mean genotype frequencies at all.

Figure 4.3 shows that heterogeneity induced by population structure can influence a genotype's fluctuation around its mean, which is reminiscent of spatial structure's effects on the molecular clock, whose rate can be altered by population asymmetries [209]. Indeed, many properties depend on a population's structure and update rule, even under neutral evolution. Genome frequency is evidently one of the rare quantities that does not. However, the framework we use to formally establish this result also gives the long-run marginal distributions for initially-monomorphic populations, even if the mutation process is not ergodic (Proposition 4.7).

The study of genotype dynamics along lineages leads to an interesting question: how does selection change the stationary distribution of individual i? Under neutral drift, this distribution is just π, but selection could potentially change this distribution in complicated ways. In populations with heterogeneous structure, selection could also affect these marginal distributions in different ways at different locations. The standard method of measuring the effects of selection on overall genotype frequency does not necessarily capture these subtleties. Nevertheless, understanding the locations at which a given genotype is more likely to be found in a population is highly relevant to the study of how spatial structure influences evolutionary dynamics.

Matrix games and the regeneration process have finitely many possible types, but our result on long-term genotype frequency holds for mutation processes with continuous state spaces as well (see Chapter 2). A type of this form could be an element of an interval such as $[0,1]$, for example, representing partial expression of a trait or the tendency to be of a particular binary type. Mutations can also be supported on uncountably infinitely many points in the genotype space, one example being when an individual with type $x \in [0,1]$ mutates to a nearby type with probability determined by a Gaussian distribution centered at x [210, 211]. On a generic state space, the technical condition we require is that the mutation process be an ergodic Harris chain [203].

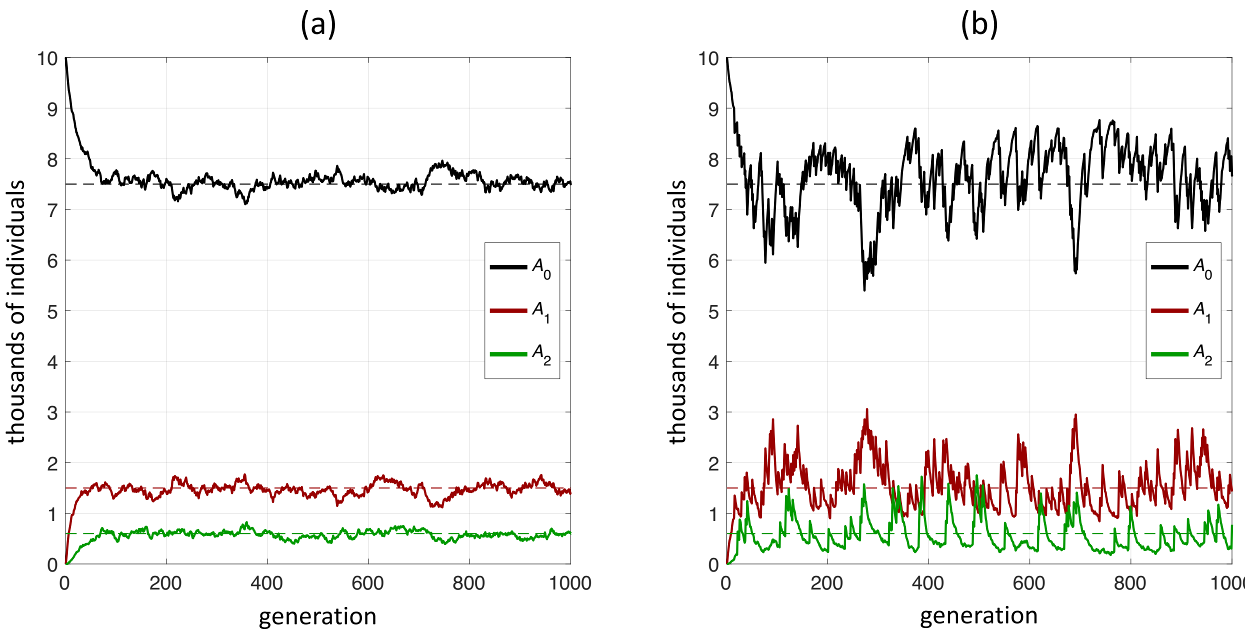

Figure 4.3: The frequency of three cell types in a population of size $N = 10^4$. The mean frequency of each type (for a population of any size) is indicated by a dashed line. Mutations in both panels are governed by the regeneration process with $m = 10^3$ and mutation rates $w = 0.01$, $u = 0.02$, and $v = 0.03$. In (a), the population is updated according to a Wright-Fisher rule. (b) illustrates a modified (and asymmetric) Wright-Fisher rule in which there exists a single cell in every generation that is more likely to reproduce than the others (independent of cell type). In (a), each offspring chooses a parent uniformly-at-random from the previous generation (i.e. each with probability $1/N = 10^{-4}$). In (b), one marked individual is chosen as the parent with probability 0.05 and each of the $N-1$ remaining individuals is chosen with probability $(1-0.05)/(N-1) \ll 0.05$. Despite the heterogeneity in (b), the average trait frequencies are the same; only the fluctuations change.

4.3. Marginal mixing times

Until this point, we have focused on neutral genotype frequencies, which are not affected by the distribution over replacement events, $\{\mathbb{D}(R,r)\}_{(R,r)}$. We now turn to rate of convergence to these equilibrium frequencies along the lineages, characterized in terms of mixing times. The proof of Theorem 4.6 is based on the fact that, eventually, all lineages will contain sufficiently many birth events for the mutation process to mix. Here, we consider the amount of time–measured in update steps–for this mixing to occur along a given lineage. Note that this is a different concept than the mixing time of the evolutionary process, which is considered later.

Let $\pi \in \Delta(\Gamma)$ be the stationary distribution for the mutation process. Let $t_{\mathrm{mix}}(\varepsilon)$ be the number of steps in the mutation process until the chain is within a distance of ε of π in total variation. Formally, let

$$\psi_t := \sup_{\alpha \in \Gamma} \sup_{E \in \sigma(\Gamma)} \left| \mathbb{P}_{\alpha} \left\{ \alpha_t \in E \right\} - \pi(E) \right|. \tag{4.3.1}$$

The ε-mixing time of the mutation process is then $t_{\mathrm{mix}}(\varepsilon) := \min\{t \geqslant 0 \ : \ \psi_t < \varepsilon\}$. We use the symbol ψ to denote this distance rather than the usual d in order to avoid confusion with death rates.

At the population level, we can consider an analogue of mixing time along lineages. Recall that $B_t^{(i)}$ is the number of birth events in the lineage leading to individual i after t update steps. Since the marginal distributions all converge to π, the distance between distribution of i's genotype at time t and the stationary distribution is

$$\Psi_t^{(i)} = \sup_{\mathbf{x} \in \Gamma^N} \sup_{E \in \sigma(\Gamma)} \left| \mathbb{P}_{\mathbf{x}} \left\{ \mathbf{x}_t(i) \in E \right\} - \pi(E) \right|. \tag{4.3.2}$$

The analogue of $t_{\mathrm{mix}}(\varepsilon)$ in this case is $t_{\mathrm{mix}}^{(i)} := \min\left\{t \geqslant 0 \ : \ \Psi_t^{(i)} < \varepsilon\right\}$, which is the ε-mixing time of the process along the lineage leading to individual i. We have the lower bound $t_{\mathrm{mix}}^{(i)}(\varepsilon) \geqslant t_{\mathrm{mix}}(\varepsilon)$ for all $i \in [\![N]\!]$.

Example 4.8 (Non-overlapping generations). *If generations do not overlap (for example Wright-Fisher updating), then* $\mathbb{P}\left\{B_t^{(i)} = t\right\} = 1$ *for all* $i \in [\![N]\!]$. *Trivially, then, we have* $\Psi_t^{(i)} = \psi_t$ *and* $t_{mix}^{(i)}(\varepsilon) = t_{mix}(\varepsilon)$ *for every* $i \in [\![N]\!]$.

Suppose, for instance, that the death rate is constant and equal to $d \in (0,1)$ so that each individual is updated every $1/d$ update steps (on average). Since every individual has the same probability of being replaced, $t_{\mathrm{mix}}^{(i)}(\varepsilon)$ is independent of i. If t_{mix} is the mixing time of the mutation process, then a natural guess for $t_{\mathrm{mix}}^{(i)}$ is simply t_{mix}/d since, on average, the lineage experiences a mutation (i.e. a step in the mutation chain) only every $1/d$ updates. However, it turns out that t_{mix}/d is generally a bad approximation of $t_{\mathrm{mix}}^{(i)}$ because it doesn't take into account enough information about the distribution of lineage length. In fact, one cannot even use t_{mix}/d to establish a general upper or lower bound on $t_{\mathrm{mix}}^{(i)}$. Figure 4.4 illustrates this comparison for death-birth updating with $d = 1/N$, where N is the population size.

Instead, we can bound $\Psi_t^{(i)}$ by the average of $\psi_{B_t^{(i)}}$, which we state as a simple lemma:

LEMMA 4.9. *For every $i \in [\![N]\!]$ and $t \geqslant 1$, we have*

$$\Psi_t^{(i)} \leqslant \mathbb{E}\left[\psi_{B_t^{(i)}}\right]. \tag{4.3.3}$$

PROOF. Since $\mathbb{P}_{\mathbf{x}}\{\mathbf{x}_t(i) \in E\} = \sum_{L=0}^{t} \mathbb{P}\left\{B_t^{(i)} = L\right\} \mathbb{P}_{\mathbf{x}}\{\alpha_L \in E\}$, we see that for every $\mathbf{x} \in \Gamma^N$ and $E \in \sigma(\Gamma)$,

$$\begin{aligned} |\mathbb{P}_{\mathbf{s}}\{\mathbf{x}_t(i) \in E\} - \pi(E)| &\leqslant \sum_{L=0}^{t} \mathbb{P}\left\{B_t^{(i)} = L\right\} |\mathbb{P}_{\mathbf{x}}\{\alpha_L \in E\} - \pi(E)| \\ &= \mathbb{E}\left[\left|\mathbb{P}_{\mathbf{x}}\left\{\alpha_{B_t^{(i)}} \in E\right\} - \pi(E)\right|\right] \\ &\leqslant \mathbb{E}\left[\psi_{B_t^{(i)}}\right]. \end{aligned} \tag{4.3.4}$$

Taking the supremum over all $E \in \sigma(\Gamma)$ and $\mathbf{x} \in \Gamma^N$ gives $\Psi_t^{(i)} \leqslant \mathbb{E}\left[\psi_{B_t^{(i)}}\right]$. ■

To calculate $\mathbb{E}\left[\psi_{B_t^{(i)}}\right]$, we note that $\mathbb{P}\left\{B_t^{(i)} = L\right\}$ satisfies the multivariate recurrence relation,

$$\mathbb{P}\left\{B_t^{(i)} = L\right\} = \mathbb{P}\left\{B_{T-1}^{(i)} = L\right\}(1 - d_i) + \sum_{j=1}^{N} \mathbb{P}\left\{B_{T-1}^{(j)} = L - 1\right\} e_{ji}, \tag{4.3.5}$$

with boundary conditions $\mathbb{P}\left\{B_t^{(i)} = 0\right\} = (1 - d_i)^t$ for $i \in [\![N]\!]$. Explicitly, for $1 < L \leqslant t$, we have

$$\mathbb{P}\left\{B_t^{(i)} = L\right\} = \sum_{\substack{k_0,\dots,k_L \geqslant 1 \\ k_0+\cdots+k_L = T}} (1 - d_i)^{k_L - 1} \sum_{j_{L-1}=1}^{N} \left(1 - d_{j_{L-1}}\right)^{k_{L-1}-1} e_{j_{L-1},j_L} \cdots \sum_{j_0=1}^{N} (1 - d_{j_0})^{k_0 - 1} e_{j_0,j_1}. \tag{4.3.6}$$

While the general expression for $\mathbb{P}\left\{B_t^{(i)} = L\right\}$ Equation (4.3.6) can appear complicated (depending on the replacement rule, $\{\mathbb{D}(R,r)\}_{(R,r)}$, and the resulting demographic variables), we can further simplify the bound given by Lemma 4.9 if some additional properties hold. We consider two cases in which one can be more explicit.

4.3.1. Finite, ergodic mutation chains. If Γ is finite and the mutation process $\{\alpha_t\}_{t \geqslant 0}$ is irreducible and aperiodic, then there exists $C > 0$ and $c \in (0,1)$ such that $\psi(L) \leqslant Cc^L$ for every $L \geqslant 1$ [67,203]. So from Lemma 4.9,

$$\Psi_t^{(i)} \leqslant C\mathbb{E}\left[\alpha^{B_i^t}\right]. \tag{4.3.7}$$

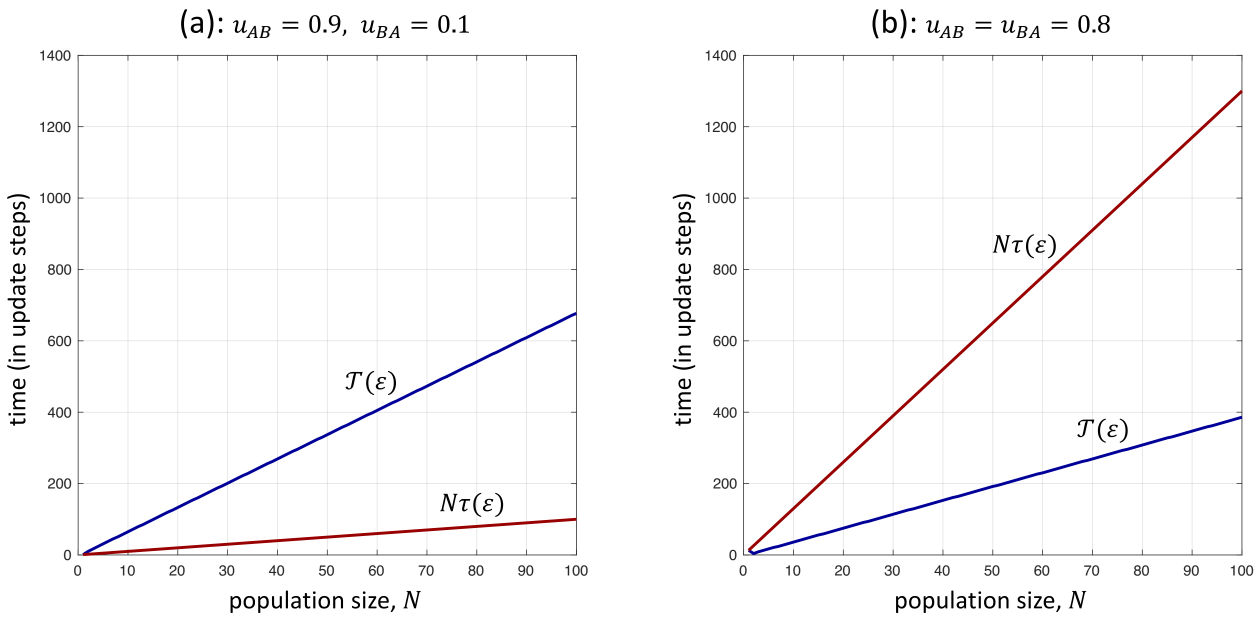

Figure 4.4: Intra-lineage mixing time, $t_{\text{mix}}^{(i)}$, relative to the mixing time of the underlying mutation process, t_{mix}, for death-birth updating (shown here for $\varepsilon = 10^{-3}$). Every individual has one of two genotypes, α or β, and upon reproduction α mutates to β with probability $u_{\alpha\beta}$ and β mutates to α with probability $u_{\beta\alpha}$. The stationary distribution of this process puts probability $\frac{u_{\beta\alpha}}{u_{\alpha\beta}+u_{\beta\alpha}}$ on state α and probability $\frac{u_{\alpha\beta}}{u_{\alpha\beta}+u_{\beta\alpha}}$ on state β. Since the death rate is constant and equal to $1/N$ (regardless of the population's spatial structure), each individual is updated every N steps (on average). However, $t_{\text{mix}}^{(i)}(\varepsilon)$ differs significantly from the rescaled mixing time of the mutation process, $Nt_{\text{mix}}(\varepsilon)$. In fact, $Nt_{\text{mix}}(\varepsilon)$ neither a general upper nor lower bound on $t_{\text{mix}}^{(i)}(\varepsilon)$, which can be seen by simply varying $u_{\alpha\beta}$ and $u_{\beta\alpha}$.

Example 4.10 (Constant death rate). *If the death rate is constant, meaning there exists d with $d_i = d$ for $i \in [\![N]\!]$, then $\mathbb{P}\left\{B_i^T = L\right\} = \binom{T}{L}(1-d)^{T-L} d^L$ (see Equation 4.3.6). From Equation* (4.3.7)*, we have the upper bound*

$$\Psi_t^{(i)} \leqslant C\left(1-d\left(1-c\right)\right)^t, \tag{4.3.8}$$

which decays exponentially to 0. Under death-birth updating, for example, the death rate is constant and equal to $1/N$. The death rate is also clearly constant when generations are non-overlapping, and in this case $1-d(1-c) = c$, which gives back the same upper bound of $C\alpha^t$. On the other hand, when $d < 1$ we have $1-d(1-c) > c$, and so the bound on mixing is slower than our assumption Cc^t.

4.3.2. Reversible mutation chains and spectral theory. Suppose that the mutation process, $\mathcal{M}$, is reversible with respect to π. The matrix $\tilde{\mathcal{M}}$ defined by $\tilde{\mathcal{M}}(\alpha,\beta) = \pi(\alpha)^{1/2}\mathcal{M}(\alpha,\beta)\pi(\beta)^{-1/2}$ is then symmetric with all of its eigenvalues real, $1 = \lambda_1 > \lambda_2 \geqslant \cdots \geqslant \lambda_{|\Gamma|} \geqslant -1$. These eigenvalues are the same as those of $\mathcal{M}$ since $\mathcal{M}$ and $\tilde{\mathcal{M}}$ represent the same linear transformation. Let $\pi_* := \min_{\alpha\in\Gamma}\pi(\alpha)$. By standard results on mixing times in reversible Markov chains [212], we have

$$\psi_t \leqslant \frac{1}{2}\sqrt{\frac{1-\pi_*}{\pi_*}}\max\left\{|\lambda_2|, \left|\lambda_{|\Gamma|}\right|\right\}^t. \tag{4.3.9}$$

To see how the evolutionary process affects mixing along its lineages, we can again consider the case of a constant death rate for simplicity. If d is the death rate for every individual, i, then the mutation process along any lineage is a Markov chain with transition matrix $\mathcal{M}_d := (1-d)\,I + d\mathcal{M}$. Moreover, $\mathcal{M}$ is reversible with respect to π since $\mathcal{M}$ is (see Section 2.5.3). Letting $\tilde{\mathcal{M}}_d := (1-d)\,I + d\tilde{\mathcal{M}}$, we see that if $\tilde{\mathcal{M}}\boldsymbol{v} = \lambda\boldsymbol{v}$, then

$$\tilde{\mathcal{M}}_d\boldsymbol{v} = \left(1 - d\,(1-\lambda)\right)\boldsymbol{v}. \tag{4.3.10}$$

Since the map $\lambda \mapsto 1 - d\,(1-\lambda)$ sends eigenvalues of $\tilde{\mathcal{M}}$ to eigenvalues of $\tilde{\mathcal{M}}_d$, and since

$$\max_{2\leqslant i\leqslant|\Gamma|}\left|1 - d\,(1-\lambda_i)\right| = \max\left\{\left|1 - d\,(1-\lambda_2)\right|, \left|1 - d\left(1-\lambda_{|\Gamma|}\right)\right|\right\}, \tag{4.3.11}$$

we have

$$\Psi_t^{(i)} \leqslant \frac{1}{2}\sqrt{\frac{1-\pi_*}{\pi_*}}\max\left\{\left|1 - d\,(1-\lambda_2)\right|, \left|1 - d\left(1-\lambda_{|S|}\right)\right|\right\}^t. \tag{4.3.12}$$

On the other hand, if the death rate depends on i, we cannot necessarily transform the mutation process into one of the form $\mathcal{M}_\varepsilon = (1-\varepsilon)\,I + \varepsilon\mathcal{M}$ for some $\varepsilon \in (0,1]$. Instead, starting at location i, one needs to keep track of where in the population each ancestor arises. Since $\mathcal{M}$ is reversible, its time-reversal is again just $\mathcal{M}$. Consider the enriched Markov chain y_t on $[\![N]\!] \times \Gamma$, $\{\alpha_t\}_{t\geqslant 0}$, where the first coordinate records the ancestor, with transitions given by

$$T\left((i,\alpha), (j,\beta)\right) = \begin{cases} e_{ji}\mathcal{M}(\alpha,\beta) & i \neq j \text{ or } \alpha \neq \beta, \\ \\ e_{ii}M_{\alpha,\alpha} + 1 - d_i & i = j \text{ and } \alpha = \beta. \end{cases} \tag{4.3.13}$$

Starting at location i, one can then ask how many steps is required for the marginal genotype distribution to be close to π (in total variation). From the definition of $\Psi_t^{(i)}$, it is easily seen that

$$\Psi_t^{(i)} = \sup_{\alpha\in\Gamma} \sup_{E\in\sigma(\Gamma)} \left|\mathbb{P}_{(i,\alpha)}\left\{y_t \in [\![N]\!] \times E\right\} - \pi\left(E\right)\right|. \tag{4.3.14}$$

As a result, the mixing time of the marginal genotype distribution of $\{y_t\}_{t\geqslant 0}$ starting from i is $t_{\mathrm{mix}}^{(i)}(\varepsilon)$.

We have seen this framework can also be used to characterize convergence to the stationary distribution along the lineages. Since each lineage can contain at most one birth event per update step, the lineages mix at least as slowly as does the original mutation process. In general, if $\Psi_t^{(i)}$ is the distance between the stationary distribution and the lineage leading to individual i, then $\Psi_t^{(i)} \leqslant \mathbb{E}\left[\psi_{B_t^{(i)}}\right]$, where ψ_t is the distance between the mutation process and the stationary distribution after t steps and $B_t^{(i)}$ is a random variable giving the number of birth events along the lineage leading to i at time t. The distribution of $B_t^{(i)}$ can be given explicitly in terms of the demographic variables of the update rule (Equation (4.3.6)), which gives a bound on $\Psi_t^{(i)}$ that is straightforward to calculate (although the expression itself is not necessarily simple).

4.4. Mixing times

In this section, we turn to the mixing time of the whole evolutionary process $\{\mathbf{x}_t\}_{t\geqslant 0}$, not only the marginals. Again, let $(\Gamma, \mathcal{M})$ be the mutation process. Let $\mathbb{D}$ define a neutral replacement distribution. The next Theorem bounds the mixing time of the evolutionary process in terms of a property of the mutation process and the replacement distribution.

Theorem 4.11. *Define*

$$\phi := \max_{\alpha,\beta\in\Gamma} \left\|\mathcal{M}(\alpha,\cdot) - \mathcal{M}(\beta,\cdot)\right\|_{\mathrm{TV}} \quad \textit{and} \quad R := \min_{i\in[\![N]\!]} \sum_{j=1}^{N} \left(e_{ji} - \phi e_{ij}\right), \tag{4.4.1}$$

then

$$t_{mix}(\varepsilon) \leqslant \frac{-\log\left(\varepsilon\right) + \log N}{R}. \tag{4.4.2}$$

Proof. Let $\mathbf{x}, \mathbf{y} \in \Gamma^N$ be states of the evolutionary process. Then we define the transportation metric

$$\rho(\mathbf{x},\mathbf{y}) := \sum_{i=1}^{N} \mathbf{1}\left(\mathbf{x}(i) \neq \mathbf{y}(i)\right), \tag{4.4.3}$$

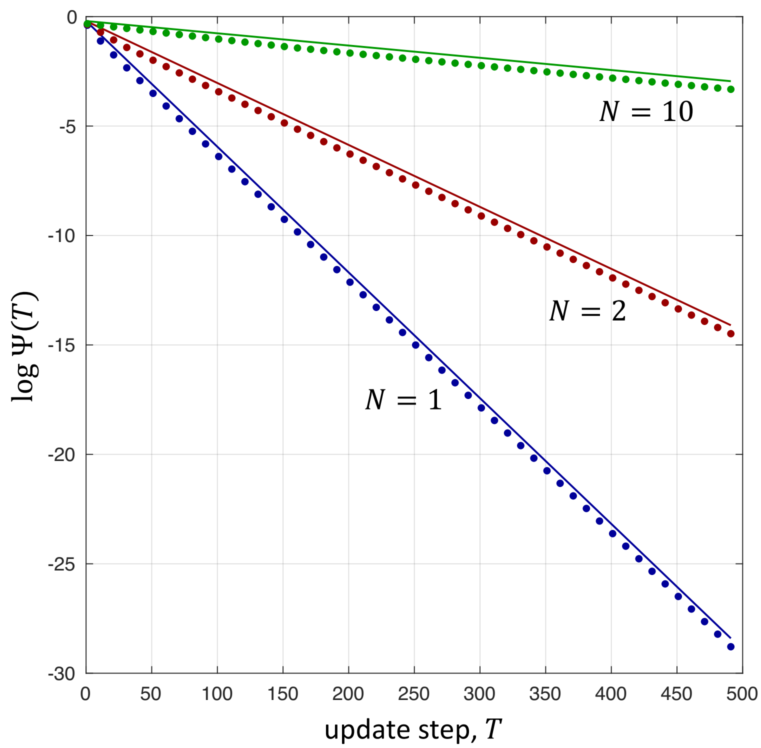

Figure 4.5: The distance between individual i's genotype and the stationary distribution after T update steps, $\Psi(T)$, under death-birth updating. All of the marginal distributions converge to π as $T \to \infty$, but they do so at different rates depending on the population size, structure, and update rule. A population of size $N = 1$ is the same as the underlying mutation process itself, and in this case $\Psi(T) = \psi(T) \leqslant Cc^T$ for some $C > 0$ and $c \in (0, 1)$. The mutation process depicted here is reversible with $|\Gamma| = 3$ genotypes, so we can set $C = \frac{1}{2}\sqrt{\frac{1-\pi_*}{\pi_*}}$ and $c = \max\{|\lambda_2|, |\lambda_3|\}$ (see Equation (4.3.9)). If the population has size N, then we know that $\Psi(T) \leqslant C\left(1 - \frac{1}{N}(1-c)\right)^T$ since death rate for all individuals is $1/N$ under death-birth updating. Since $\Psi(T)$ converges rapidly to 0 in all cases, we show both the predicted upper bounds (solid lines) and the actual values of $\Psi(T)$ (dots) on a logarithmic scale.

that is, the number of individuals that have different genotypes in states $\mathbf{x}$ and $\mathbf{y}$. Define a graph whose vertices are the elements of Γ^N and where $\{\mathbf{x}, \mathbf{y}\}$ is an edge if $\rho(x, y) = 1$, that is, they differ for exactly one individual. Note that this graph is a hypercube with alphabet Γ and the diameter of this graph is N. We want to show that ρ contracts on neighbors of the graph (see Chapter 14 of [67]).

Define the following coupling for the distribution of the next step of the evolutionary process $\mathbf{x}_1$ and $\mathbf{y}_1$, when they are conditioned such that $\mathbf{x}_0 = \mathbf{x}$ and $\mathbf{y}_0 = \mathbf{y}$. The replacement sample is identical for both processes and is sampled according to $\mathbb{D}$. This is correct marginally for both processes, as the process is neutral and the distribution of the replacement does not depend on the current state. Then for each replacement $i \in R$, we do the following: Denote the replacement by $r(i) = j$ and suppose i takes on type α after replacement, that is, $\mathbf{x}_1(i) = \alpha$. Note that this α is sampled normally according to $\mathcal{M}(\mathbf{x}(j), \cdot)$.

Now we describe how to sample $\mathbf{y}_1$. If $\mathbf{x}(j) = \mathbf{y}(j)$, the mutation is the same in both processes and $\mathbf{x}_1(i) = \mathbf{y}_1(i) = \alpha$. Otherwise $\mathbf{x}(j) \neq \mathbf{y}(j)$. There are two cases: (1) if $\mathcal{M}(\mathbf{y}(j), \alpha) \geqslant \mathcal{M}(\mathbf{x}(j), \alpha)$, then set $\mathbf{y}_1(i) = \alpha$; (2) if $\mathcal{M}(\mathbf{y}(j), \alpha) < \mathcal{M}(\mathbf{x}(j), \alpha)$, then with probability $\mathcal{M}(\mathbf{y}(j), \alpha)/\mathcal{M}(\mathbf{x}(j), \alpha)$, $\mathbf{y}(i) = \alpha$ and otherwise $\mathbf{y}(i)$ is sampled from $\mathcal{M}(\mathbf{y}(j), \cdot)$ conditional on not being α.

Note that under this coupling, if i differs in $\mathbf{x}$ and $\mathbf{y}$, then the probability that i's child differs is

$$\begin{aligned}\sum_{\alpha\in\Gamma} \mathcal{M}(\mathbf{x}(i), \alpha) \left(1 - \min\left\{1, \frac{\mathcal{M}(\mathbf{y}(i), \alpha)}{\mathcal{M}(\mathbf{x}(i), \alpha)}\right\}\right) &= \sum_{\alpha\in\Gamma} \mathcal{M}(\mathbf{x}(i), \alpha) - \min\left\{\mathcal{M}(\mathbf{x}(i), \alpha), \mathcal{M}(\mathbf{y}(i), \alpha)\right\} \\ &\leqslant \|\mathcal{M}(\mathbf{x}(i), \cdot) - \mathcal{M}(\mathbf{y}(i), \cdot)\|_{\mathrm{TV}}. \end{aligned} \tag{4.4.4}$$

Suppose $\mathbf{x}$ and $\mathbf{y}$ differ only by individual i (that is $\rho(\mathbf{x}, \mathbf{y}) = 1$), then

$$\begin{aligned}\mathbb{E}_{\mathbf{x},\mathbf{y}}\rho(X_1, Y_1) &\leqslant \mathbb{P}\left\{i \notin R\right\} + \|\mathcal{M}(\mathbf{x}(i), \cdot) - \mathcal{M}(\mathbf{y}(i), \cdot)\|_{\mathrm{TV}} \sum_{j=1}^{N} \mathbb{P}\left\{r(j) = i\right\} \\ &\leqslant 1 + \phi \sum_{j=1}^{N} e_{ij} - \sum_{j=1}^{N} e_{ji} \\ &\leqslant 1 - R \\ &\leqslant \exp\left(-R\right). \end{aligned} \tag{4.4.5}$$

Therefore, by Corollary 14.7 of [67], the mixing time of the evolutionary process is bounded by

$$t_{\mathrm{mix}}(\varepsilon) \leqslant \frac{-\log\left(\varepsilon\right) + \log N}{\max\{0, R\}}. \tag{4.4.6}$$

■

For the Moran process, we have

$$R = \frac{1-\phi}{N}, \tag{4.4.7}$$

since $e_{ij} = 1/N^2$. For the Wright-Fisher process, we have

$$R = 1 - \phi, \tag{4.4.8}$$

since $e_{ij} = 1/N$.

However, the quantity ϕ does not capture the mixing time of the mutation process in general. In fact, it can be that $\phi = 1$. Moreover, it is possible that $R \leqslant 0$, in which case, the bound on t_{mix} is vacuous.

4.5. Correlations in the stationary distributions

So far we have characterized the marginals of the stationary distribution. In this section, we try to find more detailed information about the stationary distribution and in particular how genotypes of individuals are correlated in the stationary distribution. To do this we have to study the coalescence probabilities generated by the evolutionary process. The basic idea is that after the coalescence, the mutation process along the lineages of two individuals are almost independent copies of the mutation process. Thus, after getting the distribution of the coalescence time, we can ask how many reproductions occurred along each lineage. After that, we simply study how the correlation between two independent copies of the mutation process decays over time.

Take two individuals i and j in a population. Imagine the family tree of each individual and eventually, when traced far enough backward in time, the two trees must join. This follows directly from the coherence assumption (see Section 4.1). We call this event *coalescence* [213]. Where these trees join is the *most recent common ancestor* (MRCA) and we may ask, how far back in time is this MRCA? Many interesting questions can be answered using information about this time, such as the viability of cooperation in structured populations and localization of a population in genotype space [214–216]. We take up localization in genotype space in Section 4.6.

In real populations the coalescence time has a definite answer, but it is often impossible to obtain. Thus, mathematical models are useful to study the statistics of the time. Considering the simplest case, the Wright-Fisher process: a population of N haploid individuals. We sample the parents of the next generation with replacement from the previous generation and each time a parent is selected, it produces one offspring [120, 217]. What is the coalescence time, which we denote T? The probability that any two individuals have the same parent is simply $1/N$—that is, the probability that $T = 1$. If they do not have the same parent, then the only way the two original individual can coalesce in the next generation, is if their parents have the same parent, which again has probability $1/N$. Continuing the argument backward, we find the probability that $T = t$ is exactly

$$\frac{1}{N}\left(1 - \frac{1}{N}\right)^{t-1} \tag{4.5.1}$$

or T is distributed geometrically with a mean of N generations. Note that the distribution of T does not depend on the individuals that are coalescing.

In the Moran process, we find a similar distribution for the time to the MRCA. The only way two individuals can coalesce in the Moran process is for one to replace the other. Either individual i can replace j or j can replace i. Each event happens with probability $1/N^2$, as the process is neutral. Then importantly, if coalesce does not occur, we are effectively in the same situation. Thus,

$$T \sim \mathrm{Geo}\left(\frac{2}{N^2}\right). \tag{4.5.2}$$

Again, the distribution of T does not depend on the individuals that are coalescing.

This argument can be generalized. Define the following demographic variables

$$c_{ij} := \mathbb{P}_{\mathbb{D}}\left\{r(i) = j, j \notin R\right\} + \mathbb{P}_{\mathbb{D}}\left\{r(j) = i, i \notin R\right\} + \sum_{k \in [\![N]\!]} \mathbb{P}_{\mathbb{D}}\left\{r(i) = r(j) = k\right\}. \tag{4.5.3}$$

As we have seen above $c_{ij} = 1/N$ in the Wright-Fisher process and $c_{ij} = 2/N^2$ in the Moran process.

THEOREM 4.12. *Suppose that for all $i, j \in [\![N]\!]$ we have $c_{ij} = c$ for some constant $c \in [0,1]$. Then the time to the* MRCA *of any two individuals i and j is distributed as*

$$T \sim \mathrm{Geo}\left(c\right). \tag{4.5.4}$$

PROOF. For any two individuals i and j to coalesce in the next time step, there are three mutually exclusive possibilities. 1) i replaces j and i is not replaced; 2) j replaces i and j is not replaced; 3) some individual k replaces i and j. The probability of any of these possibilities occurring is c_{ij}. By assumption, this does not depend on i or j. Thus, if there is no coalescence, we are in effectively the same situation. So $T \sim \mathrm{Geo}\left(c\right)$. ∎

While T is not always geometric, we can lower bound T by a geometric random variable. The following theorem generalizes Theorem 4.12.

THEOREM 4.13. *Define* $c_* := \min_{i,j:i\neq j} c_{ij}$ *and* $c^* := \max_{i,j:i\neq j} c_{ij}$. *Then the time to the* MRCA *of any two individuals* i *and* j *satisfies*

$$(1-c^*)^t \leqslant \mathbb{P}\{T \geqslant t\} \leqslant (1-c_*)^t. \tag{4.5.5}$$

PROOF. Decompose $\mathbb{P}\{T \geqslant t\}$ as follows

$$\mathbb{P}\{T \geqslant t\} = \mathbb{P}\{T \neq 1\}\mathbb{P}\{T \neq 2|T \neq 1\}\cdots\mathbb{P}\{T \neq t|T \neq 1,\ldots,T \neq t-1\}. \tag{4.5.6}$$

Let $\mathcal{A}_t^{(i)}$ denote the ancestor of i at time t. Now, note

$$\mathbb{P}\{T \neq t|T \neq 1,\ldots,T \neq s\} = \sum_{k\neq l}(1-c_{kl})\mathbb{P}\left\{\mathcal{A}_t^{(i)} = k, \mathcal{A}_t^{(j)} = j|T \neq 1,\ldots,T \neq t-1\right\} \leqslant (1-c_*), \tag{4.5.7}$$

since the union of the events $\left\{\mathcal{A}_t^{(i)} = k, \mathcal{A}_t^{(j)} = j\right\}$ is the whole probability space . The lower bound is similar. Plugging the bound from Equation (4.5.7) into Equation (4.5.6) completes the proof. ■

4.5.1. Joint distribution from coalescence statistics. Let π be the stationary distribution and $t_{\text{mix}}(\varepsilon)$ be the mixing of the mutation process $(\Gamma, \mathcal{M})$. As before, we let T be the time to the most recent common ancestor of distinct individuals i and j under the neutral update distribution $\mathbb{D}$. Let the stationary distribution of the evolutionary process be μ.

The following theorem is analogous to the expression (4.2.2), but for the joint distribution of the genotype of two individuals rather than the marginal of a single individual.

THEOREM 4.14. *Recall that* $d_* := \min_i d_i$. *Assume that for all* $\delta, t > 0$, *there exists* N *large enough such that* $\mathbb{P}\{T < t/d_*\} < \delta$, *then*

$$\lim_{N\to\infty} \mathbb{P}_\mu\{(\mathbf{x}(i), \mathbf{x}(j)) = (\alpha,\beta)\} = \pi(\alpha)\pi(\beta) \tag{4.5.8}$$

for all distinct $i, j \in [\![N]\!]$. *That is, the join distribution of two individuals is approximately independent.*

PROOF. First, define T_i and T_j as the number of reproductive steps in the lineages of individuals i and j since time

T. Then note

$$\mathbb{P}_\mu\left\{(\mathbf{x}(i),\mathbf{x}(j))=(\alpha,\beta))\right\}=\sum_{t,t_1,t_2=1}^{\infty}\sum_{\alpha'\in\Gamma}\mathbb{P}_{\alpha'}\left\{\alpha_{t_1}=\alpha\right\}\mathbb{P}_{\alpha'}\left\{\beta_{t_2}=\beta\right\}$$

$$\cdot\,\mathbb{P}\left\{\alpha_{\text{MRCA}}=\alpha'|T=t,T_i=t_1,T_j=t_2\right\}\mathbb{P}\left\{T=t,T_i=t_1,T_j=t_2\right\}, \tag{4.5.9}$$

where α_t and β_t are independent copies of the mutation process. Note that for $t>t_1$, we have

$$\mathbb{P}\left\{T_i<t_1|T=t\right\}\leqslant\sum_{i=0}^{\lfloor t_1\rfloor}\binom{t}{i}(1-d_*)^{t-i}\leqslant C(1-d_*)^{t-\lfloor t_1\rfloor}t^{\lfloor t_1\rfloor} \tag{4.5.10}$$

for some constant $C>0$. Now fix δ and choose t large enough such that $C(1-d_*)^{t-\lfloor t_1\rfloor}t^{\lfloor t_1\rfloor}\leqslant\delta_1$, which is equivalent to

$$t\geqslant\frac{C}{d_*}\log(C/\delta_1) \tag{4.5.11}$$

for some large constant $C>0$. Note that by assumption, we can choose N large enough so that

$$\mathbb{P}\left\{T<\frac{C}{d_*}\log(C/\delta_1)\right\}\leqslant\delta. \tag{4.5.12}$$

Therefore, for this choice of N and t, we see

$$\begin{aligned}\mathbb{P}\left\{T_i<t_1\right\}&=\mathbb{P}\left\{T_i<t_1|T<t\right\}\mathbb{P}\{T<t\}+\mathbb{P}\left\{T_i<t_1|T\geqslant t\right\}\mathbb{P}\{T\geqslant t\}\\&\leqslant\delta+C(1-d_*)^{t-\lfloor t_1\rfloor}t^{\lfloor t_1\rfloor}\\&\leqslant\delta+\delta_1.\end{aligned} \tag{4.5.13}$$

Similarly, we can choose N large enough such that $\mathbb{P}\left\{T_i<t_{\text{mix}}(\varepsilon)\text{ or }T_j<t_{\text{mix}}(\varepsilon)\right\}<\delta$, then we can bound the left-hand side of equation (4.5.8) by

$$
\begin{aligned}
&\mathbb{P}\left\{T_i < t_{\text{mix}}(\varepsilon) \text{ or } T_j < t_{\text{mix}}(\varepsilon)\right\} + \sum_{t,t_1,t_2=t_{\text{mix}}(\varepsilon)}^{\infty} \sum_{\alpha'\in\Gamma} \left|\mathbb{P}_{\alpha'}\left\{\alpha_t = \alpha\right\}\mathbb{P}_{\alpha'}\left\{\beta_t = \beta\right\} - \pi(\alpha)\pi(\beta)\right| \\
&\qquad \cdot \mathbb{P}\left\{x_{\text{MRCA}} = \alpha' | T = t, T_i = t_1, T_j = t_2\right\}\mathbb{P}\left\{T = t, T_i = t_1, T_j = t_2\right\} \\
&\quad \leqslant \delta + \left(\mu(\alpha)\varepsilon + \pi(\beta)\varepsilon + \varepsilon^2\right) \sum_{t,t_1,t_2=t_{\text{mix}}(\varepsilon)}^{\infty} \mathbb{P}\left\{T = t, T_i = t_1, T_j = t_2\right\} \sum_{\alpha'\in\Gamma} \mathbb{P}\left\{\alpha_{\text{MRCA}} = \alpha' | T = t, T_i = t_1, T_j = t_2\right\} \\
&\quad \leqslant \delta + \left(\mu(\alpha)\varepsilon + \mu(\beta)\varepsilon + \varepsilon^2\right). \qquad (4.5.14)
\end{aligned}
$$

Sending $\varepsilon, \delta \to 0$ completes the proof. ■

Theorem 4.13 implies the following bound

$$
\mathbb{P}\left\{T < t/d_*\right\} = 1 - \mathbb{P}\{T \geqslant r/d_*\} \leqslant 1 - (1-c^*)^{t/d_*}. \qquad (4.5.15)
$$

It is easy to see using the union bound that $c^* \leqslant 2d^*$ and $c_* \leqslant 2d_*$. However, the assumption that $\mathbb{P}\left\{T < t/d_*\right\} \to 0$ for all t does not follow from this bound. The assumption is satisfied whenever $c^* \ll d_*$. This is the case for both the Wright-Fisher and Moran processes.

Remark 4.15. With a basically identical argument, it is possible to extend this independence of the joint distribution to larger subsets of individuals provided the time to the first coalesce increases unboundedly as N becomes large.

4.6. Localization in genotype space

In this section, we consider how populations spread out over genotype space for some specific neutral evolutionary processes [218]. We recover some know results about average pairwise Hamming distance between genotypes in the Moran and Wright-Fisher processes [214]. Typically, there is a sharp threshold for localization: If the mutation rate ε is such that $\varepsilon \ll 1/N$, then the average distance between genotypes is $o(1)$; if the mutation rate ε is such that $\varepsilon \gg 1/N$, then the average distance between genotypes is on the order of the diameter of the genotype space.

This transition is to be expected based on our results from the previous section. When the assumptions of Theorem 4.14 hold, statistics depending on the average of pairwise functions over the populations are close to evaluating the same function on two points drawn independently according to the stationary distribution of the mutation process π.

That is,

$$\mathbb{E}_\mu \langle g(\mathbf{x}(i), \mathbf{x}(j)) \rangle \approx \mathbb{E}_\pi g(\alpha, \beta), \tag{4.6.1}$$

where $\mathbf{x} \sim \mu$, $\alpha, \beta \sim \pi$ independently, and $\langle \cdot, \cdot \rangle$ represents an empirical average over $i, j \in [\![N]\!]$. The main point of this section is that you can obtain exact formulae for these statistics in many interesting cases.

4.6.1. Wright-Fisher process with the independent point mutation process on the κ-hypercube. We consider the mutation process defined in Definition 2.13 and Wright-Fisher process defined in Definition 3.14. The Markov chain $(\Gamma, \mathcal{M})$ is reversible with respect to the uniform distribution, since $\mathcal{M}$ is symmetric, thus we know that sampling the type of any single individual from the stationary distribution is also uniform on Γ by Theorem 4.6. In particular, the average abundance of each type over a long period of time is also uniform. We now ask about the structure of the stationary distribution of the evolutionary process μ. Under μ, is the population concentrated in a small fraction of sequence space or is the population diffused throughout the space? One measure of this quantity is the mean-square distance of the population or equivalently its variance. Here distance is measured according to the Hamming distance. Thus, for $\mathbf{x} \in \Gamma^N$, define

$$\bar{D} := \frac{1}{N(N-1)/2} \sum_{i=1}^{N} \sum_{j:i<j} \mathcal{D}(\mathbf{x}(i), \mathbf{x}(j)) = \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{j}^{(i)} \bar{D}(\mathbf{x}(i), \mathbf{x}(j)) \,. \tag{4.6.2}$$

We consider $\bar{D}$ as a random variable, where $(\mathbf{x}(1), \ldots, \mathbf{x}(N)) \sim \mu$.

Theorem 4.16. *For the Wright-Fisher process in a well-mixed population with the independent point mutation process on the κ-hypercube, we have*

$$\mathbb{E}_{\mathbf{x}\sim\mu} \bar{D} = n \frac{\kappa - 1}{\kappa} \frac{N(2\varepsilon - \varepsilon^2)}{N(2\varepsilon - \varepsilon^2) + (1-\varepsilon)^2}. \tag{4.6.3}$$

Proof. Using the linearity of expectation and the symmetry of $(R, r) \sim \mathbb{D}$, we have

$$\mathbb{E}_\mu \bar{D} = \mathbb{E}_\mu \bar{D}(\mathbf{x}(1), \mathbf{x}(2)). \tag{4.6.4}$$

We proceed using coalescence times. Note that if $(\mathbf{x}_0(1), \ldots, \mathbf{x}_0(N)) \sim \mu$, then $(\mathbf{x}_t(1), \ldots, \mathbf{x}_t(N)) \sim \mu$. Let T be the time to the most recent common ancestor of individuals 1 and 2. Then

$$\mathbb{E}_\mu \bar{D}(\mathbf{x}(1), \mathbf{x}(2)) = \mathbb{E}\mathbb{E}\left[\bar{D}(\mathbf{x}_t(1), \mathbf{x}_t(2)) | T = t\right]. \tag{4.6.5}$$

Let $\boldsymbol{\alpha}(t)$ and $\boldsymbol{\beta}(t)$ be independent copies of the mutation process $(\Gamma, \mathcal{M})$ such that $\boldsymbol{\alpha}_0 = \boldsymbol{\beta}_0 = \mathbf{0}$. Note

$$\bar{D}(\boldsymbol{\alpha}_t, \boldsymbol{\beta}_t) = \sum_{i=1}^{L} \mathbf{1}(\boldsymbol{\alpha}_t(i) \neq \boldsymbol{\beta}_t(i)). \tag{4.6.6}$$

Let $R_t^{\mathbf{x}}(i)$ and $R_t^{\mathbf{y}}(i)$ be the event that there was no mutation in coordinate i at time t in $\boldsymbol{\alpha}_t$ and $\boldsymbol{\beta}_t$ respectively. Then, the events are independent and each has probability $1-\varepsilon$. Note that

$$\begin{aligned}
\mathbb{P}\{\boldsymbol{\alpha}_t(i) \neq \boldsymbol{\beta}_t(i)\} &= \mathbb{P}\{\boldsymbol{\alpha}_t(i) \neq \boldsymbol{\beta}_t(i) | R_i^{\mathbf{x}}(1), \ldots, R_t^{\mathbf{x}}(i), R_i^{\mathbf{y}}(1), \ldots, R_t^{\mathbf{y}}(i)\} \, \mathbb{P}\{R_i^{\mathbf{x}}(1), \ldots, R_t^{\mathbf{x}}(i), R_i^{\mathbf{y}}(1), \ldots, R_t^{\mathbf{y}}(i)\} \\
&\quad + \mathbb{P}\{\boldsymbol{\alpha}_t(i) \neq \boldsymbol{\beta}_t(i) | \left(R_i^{\mathbf{x}}(1), \ldots, R_t^{\mathbf{x}}(i), R_i^{\mathbf{y}}(1), \ldots, R_t^{\mathbf{y}}(i)\right)^c\} \, \mathbb{P}\{\left(R_i^{\mathbf{x}}(1), \ldots, R_t^{\mathbf{x}}(i), R_i^{\mathbf{y}}(1), \ldots, R_t^{\mathbf{y}}(i)\right)^c\} \\
&= \frac{\kappa - 1}{\kappa} \left(1 - (1-\varepsilon)^{2t}\right), \qquad (4.6.7)
\end{aligned}$$

since if there have been no mutations the coordinates must be equal and otherwise the coordinates are uniformly random in $[\![\kappa]\!]$. Moreover, the mutations on each coordinate are independent, thus we have $\bar{D}(\mathbf{x}_t, \mathbf{y}_t) \sim \mathrm{Bin}(n, p_t)$, where

$$p_t := \frac{\alpha - 1}{\alpha} \left(1 - (1-u)^{2t}\right). \tag{4.6.8}$$

Recall the distribution of T from Equation (4.5.1). Then

$$\begin{aligned}
\mathbb{E}\mathbb{E}\left[\bar{D}(\mathbf{x}_t(1), \mathbf{x}_t(2)) | T = t\right] &= \sum_{t=1}^{\infty} n \, p_t \frac{1}{N} \left(1 - \frac{1}{N}\right)^{t-1} \\
&= n \frac{\kappa-1}{\kappa} \sum_{t=1}^{\infty} \frac{1}{N} \left(1 - \frac{1}{N}\right)^{t-1} - \frac{n}{N-1} \frac{\kappa - 1}{\kappa} \sum_{t=1}^{\infty} \left(\left(1 - \frac{1}{N}\right) (1-\varepsilon)^2 \right)^t \\
&= n \frac{\kappa - 1}{\kappa} \frac{N(2\varepsilon - \varepsilon^2)}{N(2\varepsilon - \varepsilon^2) + (1-\varepsilon)^2}. \qquad (4.6.9)
\end{aligned}$$

■

The expression (4.6.3) is interpretable. Hold n and κ fixed. Note that if the mutation rate is very small, that is, $\varepsilon \ll 1/N$, then

$$n \frac{\kappa - 1}{\kappa} \frac{N(2\varepsilon - \varepsilon^2)}{N(2\varepsilon - \varepsilon^2) + (1-\varepsilon)^2} \to 0, \tag{4.6.10}$$

as $N \to \infty$. So, we see that the population is grouped close together in genotype space. This is because the correlations from the evolutionary process are very strong. When mutation is frequent, that is, $\varepsilon \gg 1/N$, we see

$$n\frac{\kappa-1}{\kappa}\frac{N(2\varepsilon-\varepsilon^2)}{N(2\varepsilon-\varepsilon^2)+(1-\varepsilon)^2} \to n\frac{\kappa-1}{\kappa}, \tag{4.6.11}$$

as $N \to \infty$. This is exactly the expected distance between two points that are sampled independently and uniformly from the genotype space (see Equation (2.1.10)). In this regime the correlations from the evolutionary process are overwhelmed by the frequent mutations.

4.6.2. Moran process with the independent point mutation process on the κ-hypercube. We consider the mutation process defined in Definition 2.13 and Moran process defined in Definition 3.13.

THEOREM 4.17. *For the Moran Process in a well-mixed population with the independent point mutation process on the κ-hypercube, we have*

$$\mathbb{E}_\mu \bar{D} = n\frac{\kappa-1}{\kappa}\frac{\varepsilon N}{\varepsilon(N-1)+1}. \tag{4.6.12}$$

PROOF. In the Wright-Fisher process, we know that as we go back in time, every individual is unborn and produces a parent. Thus, time and number of reproductive steps are linked, since

$$\mathbb{P}_{\mathbb{D}}\left\{R = [\![N]\!]\right\} = 1. \tag{4.6.13}$$

In the calculation above, we used this fact that given the coalescence time, we know exactly how many transitions have been made in the individual's past according to the mutation matrix. The Moran process is more complicated. We must condition on both the time to the MRCA and the number of reproductive steps that have been made by individuals 1 and 2. We call these random variables T, T_1, and T_2 respectively. Thus, as before

$$\mathbb{E}_\mu \bar{D}(\mathbf{x}(1),\mathbf{x}(2)) = \mathbb{E}\mathbb{E}\left[\bar{D}(\boldsymbol{\alpha}_{t_1},\boldsymbol{\beta}_{t_2})|T=t, T_1=t_1, T_2=t_2\right]. \tag{4.6.14}$$

First,

$$T_1+T_2|T \sim 1+\mathrm{Bin}\left(T-1, \frac{2(N-1)}{N^2-2}\right) \text{ and } T \sim \mathrm{Geo}(2/N^2) \tag{4.6.15}$$

since for each time step the individuals do not coalesce, there is a chance for one of the individuals to make a reproductive step. For each time step, these chances are independent and have probability

$$\frac{\frac{2}{N}\frac{N-1}{N}}{1-\frac{2}{N}\frac{1}{N}} = \frac{2(N-1)}{N^2-2}. \tag{4.6.16}$$

Moreover, on the final time step when the two coalesce, there must be a reproductive step. Second, note (4.6.7) easily generalizes to

$$p_{t_1 t_2} := \mathbb{P}\left\{\boldsymbol{\alpha}_{t_1}(i) \neq \boldsymbol{\beta}_{t_2}(i)\right\} = \frac{\kappa-1}{\kappa}\left(1-(1-\varepsilon)^{t_1+t_2}\right). \tag{4.6.17}$$

Thus,

$$\begin{aligned}
\mathbb{E}_\mu \bar{D}(\mathbf{x}(1),\mathbf{x}(2)) &= \mathbb{E}\mathbb{E}\left[np_{T1,T_2}|T=t\right] \\
&= n\frac{\kappa-1}{\kappa}\left(1-\mathbb{E}\mathbb{E}\left[(1-\varepsilon)^{T_1+T_2}|T=t\right]\right) \\
&= n\frac{\kappa-1}{\kappa}\left(1-(1-\varepsilon)\mathbb{E}\left(1-\frac{2\varepsilon(N-1)}{N^2-2}\right)^{T-1}\right) \\
&= n\frac{\kappa-1}{\kappa}\left(1-(1-\varepsilon)\sum_{t=1}^{\infty}\frac{2}{N^2}\left(1-\frac{2}{N^2}\right)^{t-1}\left(1-\frac{2\varepsilon(N-1)}{N^2-2}\right)^{t-1}\right) \\
&= n\frac{\kappa-1}{\kappa}\frac{\varepsilon N}{\varepsilon(N-1)+1}.
\end{aligned} \tag{4.6.18}$$

■

As before, we see sharp threshold for localization at $\varepsilon \sim \frac{1}{N}$, despite the difference in the details of the evolutionary process.

4.6.3. Wright-Fisher Process with a random walk on the cycle. We again consider the Wright-Fisher process. The mutation process is defined by a random walk on the cycle $\Gamma := \mathbb{Z}/n\mathbb{Z}$, so the mutation kernel is given by

$$\mathcal{M}(\alpha,\beta) := \begin{cases} 1-\varepsilon & \text{if } \alpha=\beta \\ \varepsilon/2 & \text{if } \alpha=\beta\pm 1 \\ 0 & \text{otherwise} \end{cases}, \tag{4.6.19}$$

for $\alpha,\beta \in \Gamma$. That is a lazy random walk on Γ that moves with probability ε independently at each time. For simplicity, we consider odd n; the calculation is identical for even n with one small change. We use the distance metric $\mathcal{D}(\alpha,\beta) := |\alpha-\beta| \pmod n$, which is just the graph distance.

Theorem 4.18. *For the Wright-Fisher process in a well-mixed population with the mutation process a random walk on the n-cycles, we have*

$$\mathbb{E}_\mu \bar{D}(\alpha,\beta) \approx \frac{n^2-1}{4n} - \frac{1}{2nN}\frac{\sin^{-2}\left(\frac{\pi}{2n}\right)}{1-\left(1-\varepsilon+\varepsilon\cos\left(\frac{\pi}{2n}\right)\right)^{-2}}. \tag{4.6.20}$$

Proof. Note that Γ is an Abelian group under addition and that our mutation process $(\Gamma,\mathcal{M})$ is a random walk on this group. We proceed with group representation techniques (see [63] for details). The group has degree one, irreducible representations

$$\rho_j(\alpha) = \exp\left(\frac{2\pi \mathrm{i} j\alpha}{n}\right), \tag{4.6.21}$$

for $j,\alpha \in \mathbb{Z}_n$. Denote $\mathcal{M}^t(\alpha) := \mathcal{M}^t(0,\alpha)$, then its Fourier transform $\hat{\mathcal{M}}^1$ at ρ_j is

$$1-\varepsilon+\varepsilon\cos\left(2\pi j/n\right). \tag{4.6.22}$$

Note also $\hat{\mathcal{M}}^t(\rho_j) = \left(\hat{M}^1(\rho_j)\right)^t$. Similarly, the Fourier transform of $\mathcal{D}(\alpha) := \mathcal{D}(0,\alpha)$ is

$$\hat{\mathcal{D}}(\rho_j) = \sum_{\alpha=0}^{n}\mathcal{D}(\alpha)\rho_j(\alpha) = \sum_{\alpha=1}^{\frac{n-1}{2}} 2\alpha\cos\left(\frac{2\pi\alpha j}{n}\right) = \frac{\cos(\pi j)\cos\left(\pi j/n\right)-1}{2\sin^2(\pi j/n)} + \frac{n\sin(\pi j)}{2\sin(\pi j/n)}, \tag{4.6.23}$$

since $\mathcal{D}$ an even function. There is an issue in equation (4.6.23) when $j=0$. To define the right-hand side then, we take the limit $j\to 0$ and get

$$\frac{n(n-1)}{4}. \tag{4.6.24}$$

Define

$$\delta(t_1,t_2) := \mathbb{E}_\mu\,\mathcal{D}(\alpha_{t_1},\beta_{t_2}), \tag{4.6.25}$$

where α_t and β_t are independent copies of the Markov chain $(\Gamma,\mathcal{D})$ such that $\alpha_0=\beta_0\sim \mathrm{Unif}(\Gamma)$. Note that

$$\delta(t_1,t_2) = \mathbb{E}_0\,\mathcal{D}(\alpha_{t_1+t_2},0) =: \delta(t_1+t_2), \tag{4.6.26}$$

that is, for $\mathbb{P}\{\alpha_0 = 0\} = 1$. Using the Plancherel formula and equations (4.6.22) and (4.6.23), we see

$$\delta(t) = \sum_{\alpha=0}^{n-1} \mathcal{D}(\alpha)\mathcal{M}^t(\alpha) \tag{4.6.27}$$

$$= \frac{1}{n}\sum_{j=0}^{n-1}\left(\frac{\cos(\pi j)\cos(\pi j/n) - 1}{2\sin^2(\pi j/n)} + \frac{n\sin(\pi j)}{2\sin(\pi j/n)}\right)(1-\varepsilon+\varepsilon\cos(2\pi j/n))^t\ . \tag{4.6.28}$$

Now we are ready to calculate the expected pairwise distance under the stationary distribution. Again, conditioning on the time to the MRCA T and then using equation (4.6.27), we see

$$\begin{aligned}
\mathbb{E}_\mu \bar{D}(\alpha,\beta) &= \mathbb{E}\mathbb{E}\left[\mathcal{D}(\alpha_t,\beta_t)|T=t\right] \\
&= \mathbb{E}\delta(2T) \\
&= \sum_{t=1}^{\infty}\frac{1}{N}\left(1-\frac{1}{N}\right)^{t-1}\delta(2t) \\
&= \frac{1}{n(N-1)}\sum_{j=0}^{n-1}\left(\frac{\cos(\pi j)\cos(\pi j/n) - 1}{2\sin^2(\pi j/n)} + \frac{n\sin(\pi j)}{2\sin(\pi j/n)}\right) \\
&\quad\cdot\sum_{t=1}^{\infty}\left(\frac{N-1}{N}\right)^t(1-\varepsilon+\varepsilon\cos(2\pi j/n))^{2t} \\
&= \frac{1}{nN}\sum_{j=0}^{n-1}\left(\frac{\cos(\pi j)\cos(\pi j/n) - 1}{2\sin^2(\pi j/n)} + \frac{n\sin(\pi j)}{2\sin(\pi j/n)}\right)\frac{(1-\varepsilon+\varepsilon\cos(2\pi j/n))^2}{1-\frac{N-1}{N}(1-\varepsilon+\varepsilon\cos(2\pi j/n))^2}, \qquad (4.6.29)
\end{aligned}$$

where we summed the geometric series in t. Next, we separate out the summand for $j = 0$ and use the fact that the summands are again even functions in j. Thus,

$$\begin{aligned}
\mathbb{E}_\mu \bar{D}(\alpha,\beta) = \frac{n^2-1}{4n} + \frac{2}{nN}\sum_{j=1}^{\frac{n-1}{2}}&\left(\frac{\cos(\pi j)\cos(\pi j/n) - 1}{2\sin^2(\pi j/n)} + \frac{n\sin(\pi j)}{2\sin(\pi j/n)}\right) \\
&\cdot\frac{(1-\varepsilon+\varepsilon\cos(2\pi j/n))^2}{1-\frac{N-1}{N}(1-\varepsilon+\varepsilon\cos(2\pi j/n))^2}\ . \qquad (4.6.30)
\end{aligned}$$

The summands do not oscillate too much and, in fact, the sum is dominated by its first term when N is large, so we can approximate

$$\mathbb{E}_\mu \bar{D}(\alpha,\beta) \approx \frac{n^2-1}{4n} - \frac{1}{2nN}\frac{\sin^{-2}\left(\frac{\pi}{2n}\right)}{1-\left(1-\varepsilon+\varepsilon\cos\left(\frac{\pi}{2n}\right)\right)^{-2}} \tag{4.6.31}$$

neglecting the $o(1/N)$ term also. ■

4.6.4. Wright-Fisher process with a random walk on the complete graph. We again consider the Wright-

Fisher process. The mutation process is defined by a random walk on the complete graph $\Gamma = [\![n]\!]$, so the mutation kernel is given by

$$\mathcal{M}(\alpha,\beta) := \begin{cases} 1-\varepsilon & \text{if } \alpha=\beta \\ \\ \frac{\varepsilon}{n-1} & \text{otherwise} \end{cases}, \tag{4.6.32}$$

for $\alpha, \beta \in \Gamma$. That is a lazy random walk on Γ that moves with probability ε independently at each time. We use the distance metric $\mathcal{D}(\alpha,\beta) := \mathbf{1}(\alpha \neq \beta)$, which is just the graph distance.

Theorem 4.19. *For the Wright-Fisher process in a well-mixed population with the mutation process a random walk on the complete graph of size n, we have*

$$\mathbb{E}_\mu \bar{D}(\alpha,\beta) = \frac{n-1}{n}\left(\frac{N\left(1-\left(1-n\frac{n}{n-1}\right)^2\right)}{N\left(1-\left(1-n\frac{n}{n-1}\right)^2\right)-\left(1-n\frac{n}{n-1}\right)^2}\right) \tag{4.6.33}$$

Proof. The proof is similar to before. Project Markov chain with $\alpha \mapsto \mathbf{1}(\alpha \neq 1)$. Then

$$\delta(t) = \frac{n-1}{n}\left(1-\left(1-\varepsilon\frac{n}{n-1}\right)^t\right) \tag{4.6.34}$$

Thus, conditioning and summing over the time to the MRCA, we see

$$\begin{aligned}
\sum_{t=1}^{\infty}\frac{1}{N}\left(1-\frac{1}{N}\right)^{2t}\delta(2t) &= \frac{n-1}{n}\left(1-\frac{1}{N-1}\frac{\left(1-\varepsilon\frac{n}{n-1}\right)^2\frac{N-1}{N}}{1-\left(1-\varepsilon\frac{n}{n-1}\right)^2\frac{N-1}{N}}\right)\\
&= \frac{n-1}{n}\left(1-\frac{\left(1-\varepsilon\frac{n}{n-1}\right)^2}{N-\left(1-\varepsilon\frac{n}{n-1}\right)^2(N-1)}\right)\\
&= \frac{n-1}{n}\left(\frac{N\left(1-\left(1-\varepsilon\frac{n}{n-1}\right)^2\right)}{N\left(1-\left(1-\varepsilon\frac{n}{n-1}\right)^2\right)-\left(1-\varepsilon\frac{n}{n-1}\right)^2}\right).
\end{aligned} \tag{4.6.35}$$

■

Note that when n becomes large, we see

$$\mathbb{E}_\mu \bar{D}(\alpha,\beta) \sim \frac{N\left(2\varepsilon-\varepsilon^2\right)}{N\left(1-(1-\varepsilon)^2\right)-(1-\varepsilon)^2} \approx \frac{2\varepsilon N}{2\varepsilon N-1}. \tag{4.6.36}$$

4.6.5. Mutation process given by a random walk on a group. Define the following class of functions

$$\mathbb{F}_M := \left\{ g : g(t) = \sum_{k=0}^{M} a_k b_k^t : \text{for } a_k, b_k \in \mathbb{R} \right\}. \tag{4.6.37}$$

In the above examples, we have been able to get explicit formulae for the average pairwise distance between individuals in the stationary distribution. Note that there are three essential ingredients to this. First, we require the following for the distribution of the coalescence time:

$$p(t) := \mathbb{P}\left\{T = t\right\} \in \mathbb{F}_M, \tag{4.6.38}$$

which held in both the Moran and Wright-Fisher case as $T \sim \text{Exp}\left(2/N^2\right)$ and $T \sim \text{Exp}\left(1/N\right)$ respectively. Second, the function

$$\delta(t_1, t_2) := \mathbb{E}\left[\mathcal{D}(\alpha_{t_1}, \beta_{t_2}) | \alpha_0 = \beta_0 \sim \pi\right] \tag{4.6.39}$$

is such that $\delta(t_1, t_2) = \delta(t_1 + t_2) \in \mathbb{F}_M$. An expression of this form for δ can be found using Fourier analysis whenever the mutation process given by a random walk on a group. Third, for all $c \in \mathbb{R}$

$$\mathbb{E}\left[c^{T_1+T_2} | T\right] = g_c(T) \tag{4.6.40}$$

for some $g_c \in \mathbb{F}_M$. Equivalently, we require that the m.g.f. of $T_1 + T_2$ conditional on T is in $\mathbb{F}_M$. We saw that this held in the Moran process, as $(T_1 + T_2 | T) \sim 1 + \text{Bin}(T - 1, 2(N-1)/(N^2 - 2)$.

While these conditions may seem very specific, they hold in some important cases and can be useful for obtaining bounds even when they do not hold exactly. The utility of these assumptions can be seen in the following calculation: Let

$$p(t) = \sum_{k=0}^{M} a_{p,k} b_{p,k}^t, \tag{4.6.41}$$

$$\delta(t) = \sum_{k=0}^{M} a_{\delta,k} b_{\delta,k}^t, \tag{4.6.42}$$

and

$$g_c(t) = \sum_{k=0}^{M} a_{c,k} b_{c,k}^t, \tag{4.6.43}$$

where the coefficients can depend on i and j. Then

$$\begin{aligned}
\mathbb{E}\bar{D} &= \frac{1}{N(N-1)}\sum_{i\neq j}\mathbb{E}\mathbb{E}\left[\delta(T_1+T_2)|T\right]\\
&= \frac{1}{N(N-1)}\sum_{i\neq j}\mathbb{E}\mathbb{E}\left[\sum_{k=0}^{M} a_{\delta,k} b_{\delta,k}^{T_1+T_2}|T\right]\\
&= \frac{1}{N(N-1)}\sum_{i\neq j}\sum_{k=0}^{M} a_{\delta,k}\mathbb{E}\mathbb{E}\left[b_{\delta,k}^{T_1+T_2}|T\right]\\
&= \frac{1}{N(N-1)}\sum_{i\neq j}\sum_{k=0}^{M} a_{\delta,k}\mathbb{E}\left[g_{b_{\delta,k}}(T)\right]\\
&= \frac{1}{N(N-1)}\sum_{i\neq j}\sum_{k=0}^{M} a_{\delta,k}\sum_{t=0}^{\infty} p(t) g_{b_k}(t)\\
&= \frac{1}{N(N-1)}\sum_{i\neq j}\sum_{k,l,m=0}^{M} a_{\delta,k} a_{p,l} a_{\delta_k,m}\sum_{t=0}^{\infty}\left(b_{p,m} b_{\delta_k,m}\right)^t\\
&= \frac{1}{N(N-1)}\sum_{i\neq j}\sum_{k,l,m=0}^{M} \frac{a_{\delta,k} a_{p,l} a_{\delta_k,m}}{1-b_{p,m} b_{\delta_k,m}}. \qquad (4.6.44)
\end{aligned}$$

So we can get a closed formula for $\mathbb{E}\bar{D}$.

4.6.6. General localization bound. The following is a simple and general bound that implies localization. It can be useful when $T_1 + T_2$ depends on i and j.

LEMMA 4.20. *Let α_t be a mutation process. Suppose for all α_0 that $\mathbb{E}_{\alpha_0}\mathcal{D}(\alpha_0, \alpha_1) \leqslant \varepsilon$ and suppose for some evolutionary process that*

$$\mathbb{E}_{\alpha_0}\mathcal{D}(\alpha_0, \alpha_1) \leqslant \varepsilon, \qquad (4.6.45)$$

where T_1 and T_2 are the number of reproduction in the lineages of i and j since the MRCA*, then $\mathbb{E}\bar{D} = o(1)$*

PROOF. First,

$$\mathbb{E}\mathcal{D}(\alpha_0, \alpha_t) \leqslant \sum_{s=0}^{t-1}\mathbb{E}\,\mathcal{D}(\alpha_s, \alpha_{s+1}) \leqslant t\varepsilon \qquad (4.6.46)$$

and $\delta(t_1, t_2) \leqslant \varepsilon(t_1 + t_2)$. Therefore,

$$\mathbb{E}\bar{D} = \frac{1}{N(N-1)}\sum_{i\neq j}\mathbb{E}\delta(T_1+T_2) = \frac{\varepsilon}{N(N-1)}\sum_{i\neq j}\mathbb{E}\left[T_1+T_2\right] \qquad (4.6.47)$$

and $\mathbb{E}\bar{D} = o(1)$, since $\sum_{i\neq j}\mathbb{E}[T_1+T_2] \gg N^2\varepsilon$. ∎

5

Macroscopic evolutionary dynamics

In this chapter, we consider a model for evolution over long timescales to address three fundamental biological questions: (1) What is the timescale required for evolution to discover novel functionalities? (2) What types of functionality are in principal consistently discoverable by evolution? (3) What mechanisms does evolution use to discover novel functionality? All of these questions are quite different in character than those addressed by the models of Chapter 3. Those models help characterize the timescales of microscopic evolutionary events, like the fixation time of new mutants arising in populations of certain size and structure. In this chapter, we want to see how these microscopic event accumulate to produce a macroscopic picture of evolution, whereby zooming out from the details of the evolutionary dynamics within populations, we obtain a view of populations exploring a genotype space. We call this macroscopic process an *origin-fixation process.* We formalize novel functionality as the subset of genotypes whose phenotype has that functionality, and we call this subset the *target set.*

Into this macroscopic model we build key observations from Chapters 2, 3, and 4. From Chapter 2, we pay particular attention to the effects of high-dimensionality on genotype space like rapid mixing and the large expansion of the boundary of subsets. We also assume reversibility of mutation processes. Based on our discussion on fitness in

Section 2.6, we introduce a random model for target sets and reasonable assignments of quasi-neutral fitnesses to our genotype space. We use our calculations of fixation probabilities and absorption times to derive the origin-fixation process as a limit of the microscopic evolutionary dynamics from Chapter 3, under the assumptions of low mutation rate and a time-invariant genotype space. Finally, Chapter 4 gives us another way to derive the origin-fixation process as a limit of neutral evolutionary processes under potentially larger mutation rates. We show these derivation and obtain some key dynamical properties of the origin-fixation process in Sections 5.1 and 5.2.

Our key variable is n that indexes a sequence of genotype spaces of increasing size. In computer science there is a crucial distinction between problems that require algorithms that take polynomial or exponential time [12,24,118,219]. The latter are considered to be intractable. We motivate our study using the hypercube as an example in Subsection 5.3.1. Next, we characterize the expected discover time of a single genotype in Subsection 5.3.2 and then extend this non-singleton target sets in Subsection 5.3.4. We show the dependence of this discovery time on geometric properties of the target set, and show that in high-dimensional spaces the expected time is close to the inverse density of the target set under the stationary distribution. Thus, in many genotype spaces, where the target set has a vanishing density in the space, discoveries take time exponential in n in expectation.

We strengthen the above results to consider the full distribution of discovery times in Subsection 5.3.5. In particular, there is a clear comparison between the distribution of discovery times in the origin-fixation process and to how long it would take to find the same target set by random sampling. All results are asymptotic as n becomes large, but these results often provide good approximations for finite n. These results lead us to search for specific mechanisms that allow evolution to work on efficient timescales in Section 5.4. We call this mechanism the *regeneration process* and show that it enables discoveries in polynomial time in a very general setting. Biologically, the regeneration process relates to gene duplication, as a mechanism for the emergence of novel genes, and suggests why new genes often resemble old ones.

The specific model we derive of macroscopic evolution is part of a large class of models called origin-fixation models [220]. It has been argued that origin-fixation models form a coherent class of models of evolutionary change under certain assumptions and that these models contrast sharply with models that consider populations with standing genetic variation [220]. In these models, the role of the population in evolution is relegated by assuming all the genotypes of the population are the same (monomorphic), except during fast bouts of selection. Their defining feature is to break evolutionary change into two parts—often by invoking a separation of time scales argument.

These two parts are mutation and selection (fixation or extinction of genetic variants in the population). Importantly, other than mutation rarely producing a single mutant genotype to compete in a otherwise monomorphic wild-type population, mutation is assumed to not interfere with selection [221–224]. This results in a model of evolution where populations transition between monomorphic states at a rate determined explicitly by the rates of different mutations and their effect on fitness. This is possible because of well known formulae for the fixation probability (see Chapter 3 and [13, 225]). Without this separation of mutation and selection, explicitly determining fixation probabilities becomes mathematically difficult—even when only three competing types are considered [13].

There is a wide literature on models from this class, going back to at least 1969 [226, 227], and [220] provides an excellent review. However, their empirical applicability to evolving natural populations remains an open question [80, 130–133].

5.1. Defining the model

The key assumption for origin-fixation models is that the rate of mutation is sufficiently low. This is required both to view the state of population as monomorphic, and to explicitly determine the rates of transition between these monomorphic states using known formulae for the fixation probability. This is sometimes referred to in the literature as weak mutation.

We now describe mathematically how these models can be obtained as a limit of the evolutionary dynamics from Chapter 3 under specific assumptions. In the mutation processes we considered in Chapter 2, mutations occur during reproduction with probability ε. When ε is small, the population is concentrated in a small area of genotype space (see Section 4.6). Taking low mutation to an extreme, most of the time the population is in a monomorphic state, but sometimes a mutation occurs resulting in two different genotypes in the populations. Focusing on the dynamics once a mutation has occurred, we see that we want to avoid a third mutation entering the population during this period of selection. So we have to ask, how many reproduction typically occur during this period of selection, as a mutation is equal likely for each reproduction. Previously in Chapter 3, we calculated the absorption time for various models. While these times were measured in the number of times steps in the stochastic process, this is easily converted into generational time, where a generation is defined as the averaged number of time steps required for N reproduction to occur. For the Moran process, a single generation is given by N times steps; for the Wright-Fisher process, a single generation is given by 1 time step.

Let T be the absorption time, measured in generations, then the probability that a second mutation enters the population before the first has either fixed or gone extinct is

$$\mathbb{E}\left[1-(1-\varepsilon)^{NT}\right] \leqslant \mathbb{E}\left[\varepsilon NT + \mathcal{O}\left(\varepsilon NT\right)^2\right]. \tag{5.1.1}$$

However, we require that the probability (5.1.1) is small regardless of the fate of the first mutation; that is, we require that $\varepsilon N \mathbb{E}\left[T|F\right]$ when the event F is extinction or fixation, and for all possible effects the mutation can have on fitness. Consider the Moran process in a well-mixed population. In this case, $\mathbb{E}\left[T|F\right] = \mathcal{O}(N)$ is largest when the mutation is neutral and F is fixation. A simple intuition for this is that the number of mutants performs a random walk, and a random walk takes time $\mathcal{O}(N^2)$ to travel a distance of $\mathcal{O}(N)$. Therefore, the condition on ε is

$$\varepsilon \ll \frac{1}{N^2}, \tag{5.1.2}$$

see [220]. However, for other evolutionary processes the condition may vary. In particular, see Subsection 3.3.1 for examples of populations structure distorting absorption times.

While the exact condition is model dependent, the general principle is that low mutation rates allows us to characterize the state of the evolutionary process as a single genotype, unlike in Chapter 3 where we recorded the whole state of the evolutionary process as a vector $\mathbf{x} \in \Gamma^N$. Recording the whole state of the evolutionary process in this way, results in a Markov chain whose state space grows exponentially with the population size—this often means additional assumptions are required to prove analytic results. By considering the evolutionary dynamics from Chapter 3 in the low mutation limit, we obtain a new process that transition between monomorphic states on a different timescale. Moreover, we can obtain exact expressions for the transitions between these monomorphic states, such that the process agrees with the longterm behavior of our original evolutionary process.

These transition probabilities can be specified because in the limit of low mutation, we only see two competing genotypes in the population. So once a mutation is introduced, its fate is given by the fixation probability based on its effect on fitness compared to the wild-type. We only update the state of the process, once a new mutation has fixed. The rate at which different mutations occur is given by the mutation process $(\Gamma, \mathcal{M})$ under consideration.

Thus, the transition probabilities for the origin-fixation model are

$$\mathbb{T}(\alpha,\beta) := N\mathcal{M}(\alpha,\beta)\rho\left(\frac{\mathcal{F}(\beta)}{\mathcal{F}(\alpha)}\right) \tag{5.1.3}$$

and $\mathbb{T}(\alpha,\alpha) = 1-\sum_{\beta}^{(\alpha)}\mathbb{T}(\alpha,\beta)$. Its state space is simply Γ, where the state α represents all genotypes in the population being α. The factor N in Equation (5.1.10), means that a single step of the process corresponds to 1 generation.

Note that in the large population limit $N \to \infty$, the definition (5.1.3) can exclude deleterious mutations, since for many choices of ρ

$$\rho\,(f) \leqslant c_f^N \to 0 \tag{5.1.4}$$

when $f < 1$ for some constant $c_f < 1$. Whereas $\rho(f) \to c_f$ for some constant $c_f > 0$. So in the large population limit when selection is strong, this produces a process that is not reversible in the limit. The limiting process is also not irreducible in general, as genotypes with lower fitnesses are inaccessible from genotypes of higher fitness. Ultimately, the limiting process simply climbs to neighbors of higher fitness—resulting in quite different dynamics [49,50,228,229]. Questions about the long term behavior of this limiting process in the strong selection regime have a different flavor than those we take up in this chapter, as once the process reaches a local maximum (in terms of fitness) it is stuck there forever.

For this reason and others we discuss later, we consider the origin-fixation process either under neutrality or in the weak selection (also called quasi-neutral) regime. For weak selection, we assume that

$$\mathcal{F}(\alpha) = 1 + \frac{f_\alpha}{N} \tag{5.1.5}$$

for some constant $f_\alpha \in \mathbb{R}$ for all $\alpha \in \Gamma$. This means that the transition probabilities $\mathbb{T}(\alpha,\beta)$ and $\mathbb{T}(\beta,\alpha)$ are of the same order. Consider the Moran process for example. Since if $f_\alpha = f_\beta$, we have $\rho\left(\frac{\mathcal{F}(\beta)}{\mathcal{F}(\alpha)}\right)/\rho\left(\frac{\mathcal{F}(\alpha)}{\mathcal{F}(\beta)}\right) = 1$, and if $f_\alpha \neq f_\beta$, we have

$$\frac{\rho\left(\frac{\mathcal{F}(\beta)}{\mathcal{F}(\alpha)}\right)}{\rho\left(\frac{\mathcal{F}(\alpha)}{\mathcal{F}(\beta)}\right)} \to \exp\left(f_\beta - f_\alpha\right). \tag{5.1.6}$$

See Example 5.4 for details. Indeed, in this case,

$$N\rho\left(\frac{1+\frac{f_\beta}{N}}{1+\frac{f_\alpha}{N}}\right) \sim N\rho\left(1+\frac{f_\beta - f_\alpha}{N} + \mathcal{O}(N^{-2})\right) \sim N\frac{1-\frac{N}{N+f_\beta-f_\alpha}}{1-\frac{1}{\left(1+\frac{f_\beta-f_\alpha}{N}\right)^N}} \sim \frac{f_\beta - f_\alpha}{1-\exp(f_\alpha - f_\beta)} > 0. \tag{5.1.7}$$

Thus, we have a limiting process that is not neutral and allows both beneficial and deleterious mutations to fix.

Definition 5.1 (Origin-fixation process). *Let* $(\Gamma, \mathcal{M}, \mathcal{F})$ *be a genome space in the sense of Definition 2.1 with mutation rate* $\varepsilon = 1 - \mathcal{M}(\alpha, \alpha)$ *for all* $\alpha \in \Gamma$. *Suppose that for all* $\alpha \in \Gamma$, *we have*

$$\mathcal{F}(\alpha) = 1 + \frac{f_\alpha}{N} \tag{5.1.8}$$

for some constant $f_\alpha \in \mathbb{R}$. *Let* $(\mathbb{D}_{\mathbf{x}})_{\mathbf{x} \in \Gamma^N}$ *be an evolutionary process in the sense of Definition 3.16 with fixation probability* ρ *such that*

$$\mathbb{E}\left[T|F\right] \ll \frac{1}{\varepsilon N}. \tag{5.1.9}$$

where T *is the absorption time and* F *is the event of fixation or extinction of a mutant. Then the origin-fixation process* $(\Gamma, \mathbb{T})$ *is a Markov chain with state space* Γ *and transition matrix*

$$\mathbb{T}(\alpha, \beta) := N\mathcal{M}(\alpha, \beta)\rho\left(\frac{\mathcal{F}(\beta)}{\mathcal{F}(\alpha)}\right) \tag{5.1.10}$$

for all $\alpha, \beta \in \Gamma$.

Remark 5.2. There is another possibility, which we do not develop in detail here, that can still give some bounds on discovery times. As we have seen, very low rates of mutation are required to both localize the population in genotype space and explicitly specify transition probabilities $\mathbb{T}(\alpha, \beta)$ for the monomorphic population. However, for neutral processes, we can give up on the second requirement and still understand the longterm behavior of an evolutionary process. Suppose mutation is low enough to localize the population and contain it to a ball in genotype space of diameter D (see Section 4.6). Note that this is a weaker bound on mutation and so may increase the applicability of these models. Then if we are interested in the first time a genotype α occurs in the population, we know that all genotypes in the population must be within distance D of α at this time. Thus,

$$H(\alpha) \geqslant H\left(B_\alpha(D)\right), \tag{5.1.11}$$

where H the discovery time of a target set, defined in Equation (5.3.1). Moreover, by focusing on a single lineage in the evolving population, we have seen that marginally this is simply a copy of the mutation process on a different timescale (see Chapter 4). These two assumptions—low enough mutation rates to localize the population into a small diameter region of genotype space, and neutral evolution—are a different approach to studying evolution over long timescales.

5.2. Stationary distributions and mixing times

Now we have defined our evolutionary dynamics and discussed some sufficient assumptions for the model's validity, we can start to probe its properties. The next theorem verifies a key property, reversibility, for the model's analytic tractability.

Theorem 5.3. *Let* $(\Gamma, \mathbb{T})$ *be an origin-fixation process. Suppose that*

$$\frac{\rho\left(y/x\right)}{\rho\left(x/y\right)} = \frac{g(y)}{g(x)} \tag{5.2.1}$$

for some function $g : \mathbb{R}^+ \to \mathbb{R}^+$*, and that the mutation process* $(\Gamma, \mathcal{M})$ *is reversible. Then the origin-fixation process is reversible with respect to the stationary distribution*

$$\mu(\alpha) = \frac{\pi(\alpha) g\left(\mathcal{F}(\alpha)\right)}{\sum_\beta \pi(\beta) g\left(\mathcal{F}(\beta)\right)}. \tag{5.2.2}$$

Proof. We verify that the process is reversible with respect to μ:

$$\begin{aligned}
\mu(\alpha)\mathbb{T}(\alpha,\beta) &= \frac{\pi(\alpha) g\left(\mathcal{F}(\alpha)\right)}{\sum_\beta \pi(\beta) g\left(\mathcal{F}(\beta)\right)} N\mathcal{M}(\alpha,\beta)\rho\left(\frac{\mathcal{F}(\beta)}{\mathcal{F}(\alpha)}\right) \\
&= \frac{N}{\sum_\beta \pi(\beta) g\left(\mathcal{F}(\beta)\right)} \pi(\alpha)\mathcal{M}(\alpha,\beta) g\left(\mathcal{F}(\alpha)\right) \frac{g(\mathcal{F}(\beta))}{g(\mathcal{F}(\alpha))} \rho\left(\frac{\mathcal{F}(\alpha)}{\mathcal{F}(\beta)}\right) \\
&= \frac{N}{\sum_\beta \pi(\beta) g\left(\mathcal{F}(\beta)\right)} \pi(\beta)\mathcal{M}(\beta,\alpha) g\left(\mathcal{F}(\beta)\right) \rho\left(\frac{\mathcal{F}(\alpha)}{\mathcal{F}(\beta)}\right) \\
&= \mu(\beta)\mathbb{T}(\beta,\alpha),
\end{aligned} \tag{5.2.3}$$

where we used the assumption (5.2.1) and the reversibility of $\mathcal{M}$. ■

Example 5.4. *Consider the Moran process, where we showed in Theorem 3.3 that*

$$\rho(f) = \frac{1 - \frac{1}{f}}{1 - \frac{1}{f^N}}. \tag{5.2.4}$$

In this case,

$$\frac{\rho\left(\frac{\mathcal{F}(\beta)}{\mathcal{F}(\alpha)}\right)}{\rho\left(\frac{\mathcal{F}(\alpha)}{\mathcal{F}(\beta)}\right)} = \frac{\mathcal{F}(\beta)^{N-1}(\mathcal{F}(\beta)-\mathcal{F}(\alpha))}{\mathcal{F}(\beta)^N-\mathcal{F}(\alpha)^N}\frac{\mathcal{F}(\alpha)^N-\mathcal{F}(\beta)^N}{\mathcal{F}(\alpha)^{N-1}(\mathcal{F}(\alpha)-\mathcal{F}(\beta))} = \frac{\mathcal{F}(\beta)^{N-1}}{\mathcal{F}(\alpha)^{N-1}}, \tag{5.2.5}$$

so $g : x \mapsto x^{N-1}$.

When $\pi(\alpha) = \Theta(\pi(\beta))$ *for all* $\alpha, \beta \in \Gamma$, *we see that the stationary distribution* μ *concentrates on the set*

$$\left\{\alpha : \mathcal{F}(\alpha) = \max_{\beta\in\Gamma}\mathcal{F}(\beta)\right\} \tag{5.2.6}$$

as $N \to \infty$, *since*

$$\mu(\alpha) = \frac{\pi(\alpha)\,\mathcal{F}(\alpha)^{N-1}}{\sum_\beta \pi(\beta)\,\mathcal{F}(\beta)^{N-1}} \ll \frac{\pi(\beta)\,\mathcal{F}(\beta)^{N-1}}{\sum_\alpha \pi(\alpha)\,\mathcal{F}(\alpha)^{N-1}} = \mu(\beta) \tag{5.2.7}$$

if $\mathcal{F}(\alpha) < \mathcal{F}(\beta)$. *This provides another reason to study the process in the limit of weak selection, as in this case, we let* $\mathcal{F}(\alpha) = 1 + \frac{f_\alpha}{N}$ *and find*

$$\left(1+\frac{f_\alpha}{N}\right)^{N-1} \to e^{f_\alpha}, \tag{5.2.8}$$

so

$$\mu(\alpha) \sim \frac{\pi(\alpha)e^{f_\alpha}}{\sum_\beta \pi(\beta)e^{f_\beta}}. \tag{5.2.9}$$

In the special case where π *is uniform,* μ *can be interpreted as a type of Boltzmann distribution with* f_α *the negative energy of the state* α *(see [220, 230]).*

Note that the applicability of this form for the fixation probability is increased by Theorems 6.3 and 6.5, which prove that it holds approximately for the Moran process in many population structures.

The existence of a function g can be shown for many forms of the fixation probability, not only for the Moran process. For more examples that lead to reversible dynamics for $(\Gamma, \mathbb{T})$ see [225].

We can also compare the mixing time of the origin-fixation process $(\Gamma, \mathbb{T})$ to the mutation process $(\Gamma, \mathcal{M})$ from which it is derived.

Lemma 5.5. *Let* $(\Gamma, \mathbb{T})$ *be an origin-fixation process and denote its mixing time be* t_{mix}. *Suppose that*

$$\frac{\rho\,(y/x)}{\rho\,(x/y)} = \frac{g(y)}{g(x)} \tag{5.2.10}$$

for some function $g : \mathbb{R}^+ \to \mathbb{R}^+$*, that*

$$N\rho(1 + x/N) \sim c(x) \tag{5.2.11}$$

for some function $c(x) = \mathcal{O}(1)$*, and that the mutation process* $(\Gamma, \mathcal{M})$ *is reversible. Then* t_{mix} *is bounded by*

$$\left(\max_{\alpha,\beta:\mathcal{D}(\alpha,\beta)=1} \frac{g(f_\alpha)c\left(f_\beta - f_\alpha\right)}{\sum_{\alpha'} \pi(\alpha')g(f_{\alpha'})}\right)\left(\log \sum_{\alpha'} \pi(\alpha')g(f_{\alpha'}) - \min_{\alpha} \log\left(g(f_\alpha)\pi(\alpha)\right)\right)\tilde{t}_{mix}, \tag{5.2.12}$$

where $\tilde{t}_{mix}$ *is the mixing time of* $(\Gamma, \mathcal{M})$*.*

Proof. Let $\mathcal{F}(\alpha) = 1 + \frac{f_\alpha}{N}$. Then using the expression for the stationary distribution from Equation (5.2.2) and Theorem 2.23, we see

$$\begin{aligned} B &= \max_{\alpha,\beta:\mathcal{D}(\alpha,\beta)=1} \frac{\pi(\alpha)g(f_\alpha)}{\pi(\alpha)\sum_{\alpha'} \pi(\alpha')g(f_{\alpha'})} \frac{N\mathcal{M}(\alpha,\beta)\rho\left(\frac{\mathcal{F}(\beta)}{\mathcal{F}(\alpha)}\right)}{\mathcal{M}(\alpha,\beta)} \\ &= \max_{\alpha,\beta:\mathcal{D}(\alpha,\beta)=1} \frac{g(f_\alpha)c_{f_\beta - f_\alpha}}{\sum_{\alpha'} \pi(\alpha')g(f_{\alpha'})}. \end{aligned} \tag{5.2.13}$$

Then applying Theorem 2.21, we see

$$\tilde{t}_{\text{mix}} \leqslant B\left(\log \sum_{\alpha'} \pi(\alpha')g(f_{\alpha'}) - \min_{\alpha} \log\left(g(f_\alpha)\pi(\alpha)\right)\right)t_{\text{mix}}. \tag{5.2.14}$$

■

Note in many cases, the bound in Equation (5.2.12) can be controlled. Suppose ρ is given by the Moran process, that $f_\alpha = \Theta(1)$ for all α, and that π is uniform, then $t_{\text{mix}} \leqslant C\tilde{t}_{\text{mix}}$ for some constant C not depending on N.

5.3. Target sets and hitting times

In this section, we take up the issue of discovery times. That is, how long does it take for the origin-fixation process to find a particular genotype or subset of genotypes? We consider this both with a specified initial condition α and starting the process in the stationary distribution. For some subset $\chi \subseteq \Gamma$, we define the random variable

$$H(\chi) := \{t : x_t \in \chi\}, \tag{5.3.1}$$

where x_t is the origin-fixation process $(\Gamma, \mathcal{M}, \mathbb{D})$. We also abbreviate $H(\{\alpha\})$ as $H(\alpha)$. We call χ the target set and H the hitting time or discovery time. Sometimes it is technically useful to exclude the time 0, so we also deifne

$$H^+(\chi) := \{t > 0 : x_t \in \chi\}. \tag{5.3.2}$$

To begin with, we consider the simplest case where the target set is a singleton. There are many interesting expressions and relationships for the expectation of hitting times—some of which we state below.

LEMMA 5.6. *Suppose* $(\Gamma, \mathbb{T})$ *is an origin-fixation process, and define the resolvent for* $\mathbb{T}$ *as*

$$\mathcal{G}(\alpha, \beta) := \sum_{t=0}^{\infty} \left(\mathbb{T}^t(\alpha, \beta) - \mu(\beta)\right). \tag{5.3.3}$$

Then we have the following expressions for hitting times:

$$\mathbb{E}_\beta H(\alpha) = \frac{\mathcal{G}(\alpha, \alpha) - \mathcal{G}(\beta, \alpha)}{\mu(\alpha)}, \tag{5.3.4}$$

$$\mathbb{E}_\mu H(\alpha) = \frac{\mathcal{G}(\alpha, \alpha)}{\mu(\alpha)}, \tag{5.3.5}$$

and

$$\mathbb{P}_{\alpha'}\{H(\alpha) < H(\beta)\} = \frac{\mathbb{E}_{\alpha'} H(\beta) + \mathbb{E}_\beta H(\alpha) - \mathbb{E}_{\alpha'} H(\alpha)}{\mathbb{E}_\alpha H(\beta) + \mathbb{E}_\beta H(\alpha)}. \tag{5.3.6}$$

for all $\alpha \in \Gamma$.

PROOF. See Chapter 2 of [231] or Chapter 10 of [67] for proofs. ■

Using the random variables $H(\alpha)$ for $\alpha \in \Gamma$, we could define a number of measures to measure how likely evolution is to find a genotype α. With these measures, we could then order genotypes. Theorem 5.7, states that when $(\Gamma, \mathbb{T})$ is reversible, three very natural ways of measuring the discoverability of genotypes are equivalent [80].

THEOREM 5.7. *Suppose* $(\Gamma, \mathbb{T})$ *is a reversible origin-fixation process. Then the following inequalities are equivalent*

$$\mathbb{E}_\mu H(\alpha) \leqslant \mathbb{E}_\mu H(\beta), \tag{5.3.7}$$

$$\mathbb{E}_{\beta} H(\alpha) \leqslant \mathbb{E}_{\alpha} H(\beta), \tag{5.3.8}$$

and

$$\mathbb{P}_{\mu}\left\{H(\beta) < H(\alpha)\right\} \leqslant \mathbb{P}_{\mu}\left\{H(\alpha) < H(\beta)\right\} \tag{5.3.9}$$

for all $\alpha, \beta \in \Gamma$.

Proof. Note that for a reversible process, we have

$$\mu(\alpha)\mathcal{G}(\alpha,\beta) = \sum_{t=0}^{\infty}\left(\mu(\alpha)\mathbb{T}^t(\alpha,\beta) - \mu(\alpha)\mu(\beta)\right) = \mu(\beta)\mathcal{G}(\beta.\alpha) \tag{5.3.10}$$

by the condition in Equation (2.3.1) applied to $\mathbb{T}$. Next we see

$$\begin{aligned}
\mathbb{E}_{\beta} H(\alpha) - \mathbb{E}_{\alpha} H(\beta) &= \frac{\mathcal{G}(\alpha,\alpha) - \mathcal{G}(\beta,\alpha)}{\mu(\alpha)} - \frac{\mathcal{G}(\beta,\beta) - \mathcal{G}(\alpha,\beta)}{\mu(\beta)} \\
&= \frac{\mathcal{G}(\alpha,\alpha)}{\mu(\alpha)} - \frac{\mathcal{G}(\beta,\beta)}{\mu(\beta)} \\
&= \mathbb{E}_{\pi} H(\alpha) - \mathbb{E}_{\pi} H(\beta)
\end{aligned} \tag{5.3.11}$$

by Lemma 5.6 and Equation (5.3.10). Thus, (5.3.7) and (5.3.8) are equivalent.

Note that if we multiply the left-hand side of Equation (5.3.6) by $\mu(\alpha')$ and sum over α', we find

$$\sum_{\alpha'} \mu(\alpha')\mathbb{P}_{\alpha'}\{H(\alpha) < H(\beta)\} = \mathbb{P}_{\mu}\{H(\alpha) < H(\beta)\}. \tag{5.3.12}$$

Thus, we see

$$\begin{aligned}
\frac{\mathbb{P}_{\pi}\left\{H(\alpha) < H(\beta)\right\}}{\mathbb{P}_{\pi}\left\{H(\beta) < H(\alpha)\right\}} &= \frac{\mathbb{E}_{\mu} H(\alpha) + \mathbb{E}_{\alpha} H(\beta) - \mathbb{E}_{\mu} H(\beta)}{\mathbb{E}_{\mu} H(\beta) + \mathbb{E}_{\beta} H(\alpha) - \mathbb{E}_{\mu} H(\alpha)} \\
&= \frac{\mathbb{E}_{\alpha} H(\beta)}{\mathbb{E}_{\beta} H(\alpha)},
\end{aligned} \tag{5.3.13}$$

which proves the equivalence of (5.3.8) and (5.3.9). ■

5.3.1. Hypercube as motivation for discovery times. In Section 2.4, we argued (and proved in particular cases) that we should expect genotype spaces to have a number of general properties. We argued that they are high-dimensional, so that their volume grows rapidly with some index n. We also argued that they should display rapid mixing—that is, $|\Gamma_n| \gg t_{\mathrm{mix}}(\varepsilon)$.

Using the expression for $\mathbb{E}_\beta H(\alpha)$ in Equation (5.3.5), we can immediately start to see the implications of these properties and frequently find

$$\mathbb{E}_\mu H(\alpha) = \frac{1}{\mu(\alpha)} \sum_{t=0}^{\infty} \left(\mathbb{T}^t(\alpha,\alpha) - \mu(\alpha)\right) \approx \frac{C}{\mu(\alpha)} + c\varepsilon. \tag{5.3.14}$$

Often, the measure of α under the stationary distribution is approximately $\mu(\alpha) \approx 1/|\Gamma_n|$, which means $\mathbb{E}_\mu H(\alpha)$ is very large.

As a first simple example, let us consider a case where we can calculate explicitly [2]. Let $(\Gamma_n, \mathcal{M})$ be the single point mutation process defined in Definition 2.4, let $\mathbb{D}$ be the Moran process in a well-mixed population, and $\mathcal{F}(\boldsymbol{\alpha}) = 1$ for all $\boldsymbol{\alpha} \in \Gamma_n$. So we are calculating the expected time to find a specific sequence of DNA or amino acids. We can use the projection to a 1-dimensional process we developed in Section 2.7 and the results of Appendix A, to see

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\beta}} H(\boldsymbol{\alpha}) &= \sum_{l=1}^{i} \sum_{k=0}^{n-1} \frac{\prod_{j=l+1}^{k} p_j^+}{\prod_{j=l+1}^{k+1} p_j^-} \\
&= \frac{\kappa n}{\varepsilon} \sum_{l=1}^{i} \sum_{k=0}^{n-1} (\kappa-1)^{k-l} \frac{l!(n-l)!}{(k+1)!(n-k-1)!} \\
&= \frac{\kappa n \left(\kappa^n - 1\right)}{\varepsilon} \sum_{l=1}^{i} \frac{l!(n-l)!}{n!(\kappa-1)^{l+1}} \\
&= \Theta\left(\frac{\kappa^n}{\varepsilon}\right),
\end{aligned} \tag{5.3.15}$$

where $i = \mathcal{D}(\boldsymbol{\alpha}, \boldsymbol{\beta}) \geqslant 1$. Suppose instead of trying to discover a specific sequence, we only ask for a sequence that agree with some target sequence in a percentage of its coordinates. Mathematically, define

$$\chi := B_{\boldsymbol{\alpha}}(d), \tag{5.3.16}$$

that is, the ball centered at $\boldsymbol{\alpha}$ of radius d. Then for all $r < (\kappa-1)/\kappa$, we find again

$$\mathbb{E}_{\boldsymbol{\beta}} H(\chi) \geqslant \mathcal{O}\left(\frac{\kappa^n}{\varepsilon}\right) \tag{5.3.17}$$

when $\boldsymbol{\beta} \notin \chi$, by a similar argument to above.

Now we consider a non-neutral genotype space. Specifically, consider a multiplicative fitness landscape such that a genotype's fitness increases multiplicatively as it gets closer to $\boldsymbol{\alpha}$, that is,

$$\mathcal{F}(\alpha) = (1 + f/N)^{n - \mathcal{D}(\boldsymbol{\alpha}, \boldsymbol{\beta})}. \tag{5.3.18}$$

Then

$$\begin{aligned} \mathbb{E}_{\boldsymbol{\beta}} H(\boldsymbol{\alpha}) &= \frac{\kappa n}{\varepsilon} \sum_{l=1}^{i} \sum_{k=0}^{n-1} (\kappa - 1)^{k-l} \frac{l!(n-l-1)!}{(k+1)!(n-k-1)!} \frac{1}{N\rho\,(1+f/N)} \left(1 + \frac{f}{N}\right)^{(N-1)(k-l)} \\ &=\sim \frac{1-e^{-f}}{f} \sum_{l=1}^{i} \sum_{k=0}^{n-1} e^{(f+\log(\kappa-1))(k-l)} \frac{l!(n-l-1)!}{(k+1)!(n-k-1)!} \\ &\sim \frac{e^{-f}\left(1-e^{-f}\right)}{f} \frac{\kappa n \left(\left(1 - e^{f} + \kappa\right)^{n} - 1\right)}{\varepsilon} \sum_{l=1}^{i} \frac{l!(n-l-1)!e^{f(l+1)}}{n!(\kappa-1)^{l+1}}, \end{aligned} \tag{5.3.19}$$

by Equations (5.1.6) and (5.1.7). Thus, $\mathbb{E}_{\boldsymbol{\beta}} H(\boldsymbol{\alpha})$ is exponential in n when

$$\log(\kappa) > f. \tag{5.3.20}$$

Similar results hold for $\mathbb{E}_{\mu} B_{\boldsymbol{\alpha}}(d)$ [2].

So we have seen that finding some target sets takes a very long time in expectation, unless the fitness landscape is steadily (and strongly) increasing in all directions toward the target. The biological interpretation of such a landscape is quite strange. It implies that the functionality that increases the fitness in the phenotype can be observed very far away from the target. While we have argued that genotypes are robust to mutations in some directions (see Section 2.6), it is unreasonable to assume that they are robust in all directions to large number of mutations. How many mutation is an empirical question that may well depend on the functionality in question.

5.3.2. Expected discovery times for singleton targets. The next theorem shows that under a simple condition on μ, $\mathbb{E}_{\mu} H(\alpha)$ is close to what we would expect from randomly sampling from μ. That is, if $x_t \sim \mu$ independently for each t, then $H(\alpha) \sim \text{Geo}(\mu(\alpha))$ and

$$\mathbb{E}_{\mu} H(\alpha) = \frac{1}{\mu(\alpha)}. \tag{5.3.21}$$

This serves as a fundamental point of comparison for $H(\alpha)$ generated by $\mathbb{T}$.

THEOREM 5.8. *Let $(\Gamma, \mathbb{T})$ be an origin-fixation process with stationary distribution μ. Then for all $\alpha \in \Gamma$ such that $\mu(\alpha) = o(1)$, we have*

$$\frac{1}{\mu(\alpha)(1-\mathbb{T}(\alpha,\alpha))} \leqslant \mathbb{E}_\mu H(\alpha) \leqslant \frac{t_{mix}}{\mu(\alpha)}. \tag{5.3.22}$$

Note that in many cases the upper bound also depends on the mutation rate ε through t_{mix}.

PROOF. Let λ_2 be the second largest eigenvalue of $\mathbb{T}$ and t_{rel} its relaxation time. For the lower bound, note

$$\mathbb{P}_\alpha\left\{x_t = \alpha\right\} - \mu(\alpha) \geqslant (1-\mu(\alpha))\left(\frac{\mathbb{T}(\alpha,\alpha)}{1-\mu(\alpha)} - \frac{\mu(\alpha)}{1-\mu(\alpha)}\right)^t, \tag{5.3.23}$$

since $1-\mathbb{T}(\alpha,\alpha)$ is the probability of leaving the state α. For the upper bound, we use Equation (2.4.4) to see

$$\mathbb{P}_\alpha\left\{x_t = \alpha\right\} - \mu(\alpha) \leqslant (1-\mu(\alpha))\left(1-\lambda_2\right)^t. \tag{5.3.24}$$

Summing over t in these bounds, we find

$$\begin{aligned}(1-\mu(\alpha))\frac{1-\mu(\alpha)}{1-\mathbb{T}(\alpha,\alpha)} &= \sum_{t=0}^{\infty}(1-\mu(\alpha))\left(\frac{\mathbb{T}(\alpha,\alpha)}{1-\mu(\alpha)} - \frac{\mu(\alpha)}{1-\mu(\alpha)}\right)^t \\ &\leqslant \mathcal{G}(\alpha,\alpha) \\ &\leqslant \sum_{t=0}^{\infty}(1-\mu(\alpha))\left(1-\lambda_2\right)^t \\ &= (1-\mu(\alpha))\frac{1}{1-\lambda_2}. \end{aligned} \tag{5.3.25}$$

Finally, Theorem 2.21 and Lemma 5.6 imply

$$\frac{1}{\mu(\alpha)(1-\mathbb{T}(\alpha,\alpha))} \sim \frac{(1-\mu(\alpha))^2}{\mu(\alpha)(1-\mathbb{T}(\alpha,\alpha))} \leqslant \mathbb{E}_\mu H(\alpha) \leqslant \frac{t_{\mathrm{mix}}(1-\mu(\alpha))}{\mu(\alpha)} \sim \frac{t_{\mathrm{mix}}}{\mu(\alpha)}. \tag{5.3.26}$$

■

Note that in all the mutation processes we have considered (and thus similarly for origin-fixation processes), we have $1-\mathbb{T}(\alpha,\alpha) \geqslant 1-\varepsilon = \Theta(1)$ for all α. Moreover, t_{mix} is often much, much less than $1/\mu(\alpha)$. This implies that the bound in Equation (5.3.22) is actual bounds $\mathbb{E}_\mu(\alpha)$ quite tightly. Theorem 5.8 should be interpreted as identifying the typical behavior in origin-fixation processes. Note that the lower bound implies that in expectation it takes a long time to discover specific sequences when starting from the stationary distribution. This is particularly informative when $\mu(\alpha) = \Theta(1/|\Gamma_n|)$ (see many examples in Chapter 2), in which case $\mathbb{E}_\mu H(\alpha) \geqslant \Theta(|\Gamma_n|)$. Since the size of Γ_n frequently grows exponentially in the parameter n, which it does when Γ_n is high-dimensional, it can mean it is infeasible to expect evolution to find specific genotype sequences in reasonable timescales when it is modeled well by an origin-fixation process.

The upper bound can be interpreted as follows: the Markov chain is close to the stationary distribution every t_{mix} steps, regardless of what has happened previously; moreover, after each independent sample from the stationary distribution, we have a probability $\mu(\alpha)$ of sampling α; thus, the expected number of independent samples to find α is $1/\mu(\alpha)$ and each sample takes at most t_{mix} time steps.

5.3.3. Models for target sets. So far we have considered two types of target sets—singletons and closed balls. We now define some different types of target set that vary in size and structure. For a collection of points $\alpha_1, \ldots, \alpha_K \in \Gamma$ define the target set

$$\chi := \bigcup_{i=1}^{K} B_d(\alpha_i). \tag{5.3.27}$$

Biologically, we can interpret this as there being K specific genotypes that perform a function, that is, have a specific phenotype that we are looking for. Moreover, each genotype is robust to some small number of mutations, so that the phenotype still performs the function. While the phenotype might not be robust to all mutations on the genotype, we take all genotypes within distance d of a target sequence as a bound. The size of K and d should be informed by empirical data.

If we assume that the function of the phenotype has a direct effect on fitness, we can then suppose that $\mathcal{F}$ is neutral or quasi-neutral outside of χ (see Section 2.6).

For high-dimensional genotype spaces, we can effectively bound the size of χ:

$$|\chi| \leqslant \mathcal{O}\left(K(cd)^n\right) \tag{5.3.28}$$

if we assume $|B_\alpha(d)| \leqslant \mathcal{O}((cd)^n)$, which we argued is typical of high-dimensional spaces in Subsection 2.4.2. For example, consider $\Gamma_n = [\![\kappa]\!]^n$. Then for $d < n(\kappa-1)/\kappa$, by the union bound

$$\frac{|\chi|}{|\Gamma|} \leqslant \frac{K}{\kappa^n} \sum_{k=0}^{\lfloor d \rfloor} \binom{n}{k} (\kappa-1)^k \leqslant \mathcal{O}\left(Kc^n\right) \tag{5.3.29}$$

for some small constant $c < 1$ for n large enough. Note that the quantity $\frac{|\chi|}{|\Gamma|}$ is actually just the measure of χ under the stationary distribution π—a quantity that is significant in the theorems that follow. Note that $\frac{|\chi|}{|\Gamma|} \leqslant \mathcal{O}(c^n)$ even when we allow K_n to grow polynomially with n.

A secondary question is how to choose the genotypes $\alpha_1, \ldots, \alpha_K \in \Gamma$. Since we want to understand evolution in general, it makes sense to ask these questions about target sets that are in some sense typical. This suggests we should consider choosing the genotypes $\alpha_1, \ldots, \alpha_K \in \Gamma$ randomly. Exactly how we define this randomness, leads to different statistical models for the targets sets χ. Since studying the structure of these sets empirically is very difficult due to the size and high-dimensionality of genotype space, it seems reasonable to sample K centers uniformly and independently from Γ. Denote the set of centers for a target set χ by C_χ.

Definition 5.9 (Random target set). *For some genotype space Γ, we define the random variable $\chi_{p,d} \subseteq \Gamma$ as*

$$\chi_{p,d} := \bigcup_{\alpha : I_\alpha = 1} B_\alpha(d), \tag{5.3.30}$$

where $I_\alpha \sim Bern(p)$ independently for $\alpha \in \Gamma$. We also define the set of target centers as

$$C_{\chi_{p,d}} := \{\alpha : I_\alpha = 1\}. \tag{5.3.31}$$

The geometry of the target set $\chi_{p,d}$ is interesting. Again consider $\Gamma_n = [\![\kappa]\!]^n$. Suppose there is some constant $\delta > 0$ such that

$$\delta > \frac{2d\kappa}{(\kappa - 1)n}, \tag{5.3.32}$$

for some other constant $c > 0$. The points of the target set are spread out through the space Γ_n so long as K and d are note too large. Since every point $\boldsymbol{\alpha} \in \chi_{p,d}$ is within distance d of some point in $C_{\chi_{p,d}}$, we have

$$\mathbb{P}\left\{\min_{\boldsymbol{\alpha}', \boldsymbol{\beta}' \in C_{\chi_{p,d}}} \min_{\boldsymbol{\alpha} \in B_{\boldsymbol{\alpha}'}(d), \boldsymbol{\beta} \in B_{\boldsymbol{\beta}'}(d)} \mathcal{D}(\boldsymbol{\alpha}, \boldsymbol{\beta}) \leqslant (1 - \delta)n\frac{\kappa - 1}{\kappa}\right\} \leqslant \mathbb{P}\left\{\min_{\boldsymbol{\alpha}, \boldsymbol{\beta} \in C_{\chi_{p,d}}} \mathcal{D}(\boldsymbol{\alpha}, \boldsymbol{\beta}) \leqslant (1 - \delta)n\frac{\kappa - 1}{\kappa} + 2d\right\}. \tag{5.3.33}$$

Then applying the union bound, we see

$$\mathbb{P}\left\{\min_{\boldsymbol{\alpha},\boldsymbol{\beta}\in C_{\chi_{p,d}}} \mathcal{D}(\boldsymbol{\alpha},\boldsymbol{\beta}) \leqslant (1-\delta)n\frac{\kappa-1}{\kappa}+2d\right\} \leqslant \sum_{\boldsymbol{\alpha},\boldsymbol{\beta}\in C_{\chi_{p,d}}} \mathbb{P}\left\{\mathcal{D}(\boldsymbol{\alpha},\boldsymbol{\beta}) \leqslant (1-\delta)n\frac{\kappa-1}{\kappa}+2d\right\}$$
$$\leqslant K^2 \exp\left(d-\delta\frac{(\kappa-1)n}{2\kappa}\right)$$
$$\leqslant K^2 e^{-cn}, \tag{5.3.34}$$

for some constant $c > 0$. Thus, the probability goes to zero if $K \ll e^{cn/2}$. Note that this implies that the lower bound in Equation (5.3.33) is optimal. However, a similar bound holds for any fixed point in Γ_n—that is, most points are far from the target set with high probability so long as $|\chi_{p,d}|$ is not too large.

Another important aspect of the geometry of χ is that its boundary is much larger than its interior. The boundary of χ is defined as

$$\partial(\chi) := \{\boldsymbol{\alpha} \in \chi : \exists \boldsymbol{\beta} \notin \chi,\ \mathcal{D}(\boldsymbol{\alpha},\boldsymbol{\beta}) = 1\} \tag{5.3.35}$$

and the interior is then $\chi\backslash\partial(\chi)$. Given that the balls $|B_\alpha(d)|$ do not intersect, we have

$$|\partial(\chi)| = cK^2\binom{n}{d}(\kappa-1)^d \tag{5.3.36}$$

and

$$|\chi\backslash\partial(\chi)| = K^2\sum_{k=0}^{d-1}\binom{n}{k}(\kappa-1)^k. \tag{5.3.37}$$

Therefore,

$$|\partial(\chi)| \gg |\chi\backslash\partial(\chi)|\,. \tag{5.3.38}$$

This can be generalized to cover the case where the balls $|B_\alpha(d)|$ do intersect, using Harper's Theorem [52].

5.3.4. Expected discovery times for target sets. In this subsection, we state some general theorems that bound discovery times for general target sets, not just singletons. We then apply these bounds to the target sets we defined in the previous subsection. The first theorem relates the expected return time of the set χ to the measure of χ under the stationary distribution.

Theorem 5.10 (Kac's formula). *Let $(\Gamma, \mathbb{T})$ be an origin-fixation process with stationary distribution μ, then for* *all target sets $\chi \subseteq \Gamma$, we have*

$$\mathbb{E}_{\mu_\chi} H^+(\chi) = \frac{1}{\mu(\chi)}, \tag{5.3.39}$$

where $\mu_\chi(\alpha) = \mu(\alpha)/\mu(\chi)$ for $\alpha \in \chi$ and 0 otherwise.

Proof. See Lemma 21.13 from [67]. ■

Lemma 5.11. *Recall the definition of $\mathcal{Q}$ for a Markov chain in Equation* (2.4.7)*. Let $(\Gamma, \mathbb{T})$ be an origin-fixation process with stationary distribution μ, then for all target sets $\chi \subseteq \Gamma$, we have*

$$\frac{1}{\mathcal{Q}(\chi, \chi^c)} \leqslant \mathbb{E}_\mu H(\chi) \leqslant \frac{t_{mix}(1 - \mu(\chi))}{\mu(\chi)} \tag{5.3.40}$$

Proof. The proof is similar to the proof of Theorem 5.8, except we bound the quantity

$$\mathbb{P}_\alpha \{x_t \in \chi\} - \mu(\chi). \tag{5.3.41}$$

For the lower bound, we can apply Kac's formula 5.10. For details see Proposition 21 in Chapter 3 of [231]. ■

The following theorem generalizes Theorem 5.8 to sets of size greater than 1.

Theorem 5.12. *Let $(\Gamma, \mathbb{T})$ be an origin-fixation process with stationary distribution μ. Let $\chi \subseteq \Gamma$ be a target set such that $\mu(\chi) = o(1)$. Suppose that there is some $\chi_1 \subseteq \chi$ such that $\mu(\chi \backslash \chi_1) = o(\mu(\chi))$ and that*

$$\mathbb{T}(\alpha, \chi \backslash \{\alpha\}) = o(1) \tag{5.3.42}$$

for all $\alpha \in \chi_1$. Then we have

$$\frac{1}{\mu(\chi) \min_\alpha \{1 - \mathbb{T}(\alpha, \alpha)\}} \lesssim \mathbb{E}_\mu H(\chi) \leqslant \frac{t_{mix}}{\mu(\chi)}. \tag{5.3.43}$$

Proof. We start by applying Lemma 5.11 to χ. The upper bound in Equation (5.3.43) follows immediately from the assumption $\mu(\chi) = o(1)$. For the lower bound in Equation (5.3.43), we have

$$\begin{aligned}\mathcal{Q}(\chi, \chi^c) &= \sum_{\alpha \in \chi} \mu(\alpha) \left(1 - \mathbb{T}(\alpha, \alpha) - (\mathbb{T}(\alpha, \chi) - \mathbb{T}(\alpha, \alpha))\right) \\ &\leqslant \varepsilon\mu(\chi) - \sum_{\alpha \in \chi \backslash \chi_1} \mu(\alpha)(\mathbb{T}(\alpha, \chi) - \mathbb{T}(\alpha, \alpha)) - \sum_{\alpha \in \chi_1} \mu(\alpha)(\mathbb{T}(\alpha, \chi) - \mathbb{T}(\alpha, \alpha)) \\ &\lesssim \varepsilon\mu(\chi) + \mu(\chi\backslash\chi_1)\varepsilon + \mu(\chi_1)\varepsilon o(1) \\ &\sim \varepsilon\mu(\chi). \end{aligned} \tag{5.3.44}$$

■

Example 5.13. *Consider how Theorem 5.12 applies to the hypercube and the Moran process: let* $\Gamma_n = [\![\kappa]\!]^n$ *and* $\mathcal{M}$ *be given by the single point mutation process with* $\mathcal{F}(\boldsymbol{\alpha}) = 1$ *for all* $\boldsymbol{\alpha} \in \Gamma_n$*. In this case, we find* $\mathbb{T}(\boldsymbol{\alpha}, \boldsymbol{\alpha}) = 1 - \varepsilon + \varepsilon\frac{1}{\kappa}$ *for all* $\boldsymbol{\alpha} \in \Gamma_n$*. Then suppose* χ *is sampled according to the randomization in Definition 5.9. Then* $\mu(\chi) = |\chi|/\kappa^n$*. Now let* $\chi_1 := \partial(\chi)$*, then the calculation in Equation* (5.3.38) *implies* $\mu(\chi\backslash\chi_1) = o(\mu(\chi))$*. Thus, we have*

$$\frac{\kappa^n}{\varepsilon|\chi|} \lesssim \mathbb{E}_\mu H(\chi) \leqslant \frac{\kappa^n n \log n}{\varepsilon|\chi|}. \tag{5.3.45}$$

So, under mild assumptions, in expectation evolution does no better than sampling uniformly from μ in terms of discovering target sets. However, it could be that the expectation is large because of very low probabilities events where discovery takes a very, very long time, and typically the target set is found quickly. So in the next subsection, we address not only the expected discovery time but its full distribution.

5.3.5. Distributions of discovery times. In this subsection, we show that under mild assumptions, the expectation of the discovery time asymptotically determines the full distribution of the discovery time. This validates the comparison we have drawn previously between the discovery time of the origin-fixation process and the discovery time of random, independent sampling from the stationary distribution.

The idea of the proof is to write

$$\begin{aligned}\mathbb{P}_\mu\{H(\chi) > t\} &= \mathbb{P}_\mu\{x_0 \notin \chi, \ldots, x_t \notin \chi\} \\ &= \mathbb{P}_\mu\{x_0 \notin \chi\}\,\mathbb{P}_\mu\{x_1 \notin \chi | x_0 \notin \chi\} \cdots \mathbb{P}_\mu\{x_t \notin \chi | x_0 \notin \chi, \ldots, x_{t-1} \notin \chi\}. \end{aligned} \tag{5.3.46}$$

Then, show that the distribution $\mathbb{P}_\mu \{x_{t+1} = \cdot | x_0 \notin \chi, \ldots, x_k \notin \chi\}$ is close to μ is total variation distance. The intuition for why conditioning does not change the distribution of x_{t+1} very much is simple: the mixing of the process is rapid, so information about the process more than t_{mix} steps in the past contains little information. Moreover, the set χ is not very dense when measured with μ, so the process hitting χ in such short time (that is, t_{mix} steps) is very unlikely. Thus this conditioning does not convey much information.

We state and prove a Theorem from [232] for our setting. Note that the hitting times of sets on the hypercube have been considered before [233–235]. These works were motivated by spin-glasses, and because of the symmetry of the hypercube, obtain much tighter bounds using very precise estimates for the hitting time of a singleton.

THEOREM 5.14. *Let* $(\Gamma, \mathbb{T})$ *be an origin-fixation process with stationary distribution* μ *and mixing time* t_{mix}*. Assume that*

$$\Delta := \frac{C t_{mix} \left(1 + \log^+ \left(\mathbb{E}_\mu H(\chi)/t_{mix}\right)\right)}{\mathbb{E}_\mu H(\chi)} < 1, \tag{5.3.47}$$

then

$$\sup_t \left|\mathbb{P}_\mu \{H(\chi) > t\} - \exp\left(-t/\mathbb{E}_\mu H(\chi)\right)\right| \leqslant \mathcal{O}(\Delta). \tag{5.3.48}$$

PROOF. Note that we let the constant C change from line to line in this proof for ease of notation. Define the distribution

$$\nu(\alpha) := \lim_{t\to\infty} \mathbb{P}_\mu \{x_t = \alpha | H(\chi) > t\}. \tag{5.3.49}$$

Clearly this limit exists when $\mathbb{T}$ restricted to $\Gamma \backslash \chi$ is irreducible, however, a simple limiting argument shows its existence even when it is reducible (see Remark 2.18 in [232]).

We can bound the left-hand sided of Equation (5.3.48) by two terms:

$$\begin{aligned} \left|\mathbb{P}_\mu \{H(\chi) > t\} - \exp\left(-t/\mathbb{E}_\mu H(\chi)\right)\right| &\leqslant \left|\mathbb{P}_\mu \{H(\chi) > t\} - \mathbb{P}_\nu \{H(\chi) > t\}\right| + \left|\mathbb{P}_\nu \{H(\chi) > t\} - \exp\left(-t/\mathbb{E}_\mu H(\chi)\right)\right| \\ &\leqslant \|\mu - \nu\|_{\text{TV}} + \left|\exp\left(-t/\mathbb{E}_\nu H(\chi)\right) - \exp\left(-t/\mathbb{E}_\mu H(\chi)\right)\right|, \end{aligned} \tag{5.3.50}$$

since $\mathbb{P}_\mu \{H(\chi) > t\} = \mathbb{E}_\nu H(\chi)$. Now note

$$\begin{aligned}
\sup_t |\exp(-t/\mathbb{E}_\nu H(\chi)) - \exp(-t/\mathbb{E}_\mu H(\chi))| &= \frac{e}{e-1} \left| \frac{\mathbb{E}_\nu H(\chi)}{\mathbb{E}_\mu H(\chi)} - 1 \right| \\
&\leqslant \frac{e}{e-1} \frac{|\mathbb{E}_\nu H(\chi) - \mathbb{E}_\mu H(\chi)|}{|\mathbb{E}_\mu H(\chi)|} \\
&\leqslant \frac{e}{e-1} \frac{\|\mu - \nu\|_{\mathrm{TV}} \max_\alpha \mathbb{E}_\alpha H(\chi)}{|\mathbb{E}_\mu H(\chi)|} \\
&\leqslant C\|\mu - \nu\|_{\mathrm{TV}}, \qquad (5.3.51)
\end{aligned}$$

since we may take the maximum of α in the bound

$$\mathbb{E}_\alpha H(\chi) \leqslant t_{\mathrm{mix}} + \sum_\beta \mathbb{P}_\alpha \{x_{t_{\mathrm{mix}}} = \beta\} \mathbb{E}_\beta H(\chi) \leqslant t_{\mathrm{mix}} + \mathbb{E}_\mu H(\chi) + \frac{1}{4} \max_\beta \mathbb{E}_\beta H(\chi) \qquad (5.3.52)$$

and rearrange, then finally use the assumption $\Delta < 1$ to see $\mathbb{E}_\mu H(\chi) \geqslant 3t_{\mathrm{mix}}$.

Thus, it suffices to bound $\|\mu - \nu\|_{\mathrm{TV}}$. Define

$$\nu_k(\alpha) := \mathbb{P}_\mu \{x_{ks} = \alpha | H(\chi) \geqslant ks\}, \qquad (5.3.53)$$

for

$$s := (1 + \log(\mathbb{E}_\mu H(\chi)/t_{\mathrm{mix}}))\, t_{\mathrm{mix}} \qquad (5.3.54)$$

so that $\lim_{k\to\infty} \rho_k = \rho$. We have

$$\begin{aligned}
\|\nu_{k+1} - \mu\|_{\mathrm{TV}} &= \|\mathbb{P}_{\nu_k} \{x_s = \cdot | H(\chi) \geqslant s\} - \mu\|_{\mathrm{TV}} \\
&\leqslant \|\mathbb{P}_{\nu_k} \{x_s = \cdot | H(\chi) \geqslant s\} - \mathbb{P}_{\nu_k} \{x_s = \cdot\}\|_{\mathrm{TV}} + \|\mathbb{P}_{\nu_k} \{x_s = \cdot\} - \mu\|_{\mathrm{TV}} \\
&\leqslant \mathbb{P}_{\rho_k} \{H(\chi) < s\} + \|\mathbb{P}_{\nu_k} \{x_s = \cdot\} - \mu\|_{\mathrm{TV}}. \qquad (5.3.55)
\end{aligned}$$

We bound each term separately. For the right-hand term, we see for any initial distribution λ

$$\|\mathbb{P}_\lambda \{x_s = \cdot\} - \mu\|_{\mathrm{TV}} \leqslant e^{(s - t_{\mathrm{mix}})/t_{\mathrm{mix}}} = t_{\mathrm{mix}}/\mathbb{E}_\mu H(\chi) \qquad (5.3.56)$$

by the definition of s and t_{mix}. So applying Equation (5.3.56) with $\lambda = \nu_k$, we bound the right-hand term. For the left-hand term, first note Equation (5.3.56) implies

$$\mathbb{P}_\lambda \left\{H(\chi) > 2s\right\} \leqslant 1 - \mathbb{P}_\mu \left\{H(\chi) \leqslant s\right\} + t_{\mathrm{mix}}/\mathbb{E}_\mu H(\chi). \tag{5.3.57}$$

Then, iterating this argument, we find

$$\begin{aligned}
\mathbb{P}_\lambda \left\{H(\chi) > 2sk + 2s\right\} &\leqslant \mathbb{P}_\lambda \left\{H(\chi) > 2sk\right\} \sum_\alpha \mathbb{P}_\alpha \left\{H(\chi) > 2s\right\} \mathbb{P}_\lambda \left\{x_{2sk} = \alpha | H(\chi) > 2sk\right\} \\
&\leqslant \left(1 - \mathbb{P}_\mu \left\{H(\chi) \leqslant s\right\} + t_{\mathrm{mix}}/\mathbb{E}_\mu H(\chi)\right) \left(1 - \mathbb{P}_\mu \left\{H(\chi) \leqslant s\right\} + t_{\mathrm{mix}}/\mathbb{E}_\mu H(\chi)\right)^k \\
&\leqslant \left(1 - \mathbb{P}_\mu \left\{H(\chi) \leqslant s\right\} + t_{\mathrm{mix}}/\mathbb{E}_\mu H(\chi)\right)^{k+1}. \qquad (5.3.58)
\end{aligned}$$

Finally, summing over k and using the inequality in Equation (5.3.58), we get

$$\mathbb{E}_\mu H(\chi) \leqslant \sum_{k=0}^{\infty} 2s \mathbb{P}_\mu \left\{H(\chi) > 2ks\right\} \leqslant \frac{2s}{\mathbb{P}_\mu \left\{H(\chi) \leqslant s\right\} - t_{\mathrm{mix}}/\mathbb{E}_\mu H(\chi)}, \tag{5.3.59}$$

which rearranges to

$$\mathbb{P}_\mu \left\{H(\chi) \leqslant s\right\} \leqslant \frac{2s + t_{\mathrm{mix}}}{\mathbb{E}_\mu H(\chi)}. \tag{5.3.60}$$

Finally, using the bounds in Equation (5.3.57) and (5.3.60), we find

$$\|\nu_{k+1} - \mu\|_{\mathrm{TV}} \leqslant \frac{2s + t_{\mathrm{mix}}}{\mathbb{E}_\mu H(\chi)} + \frac{t_{\mathrm{mix}}}{\mathbb{E}_\mu H(\chi)} \leqslant \frac{C t_{\mathrm{mix}} \left(1 + \log^+ \left(\mathbb{E}_\mu H(\chi)/t_{\mathrm{mix}}\right)\right)}{\mathbb{E}_\mu H(\chi)}, \tag{5.3.61}$$

so taking the limit as $k \to \infty$ completes the proof. ■

Remark 5.15. Note that by Theorem 5.14, $\mathbb{P}_\mu \left\{H(\chi) > t\right\}$ is only order 1 as $n \to \infty$ if we set

$$t = \theta \mathbb{E}_\mu H(\chi), \tag{5.3.62}$$

when $\mathbb{E}_\mu H(\chi) \gg t_{\mathrm{mix}} \to \infty$. In which case, we see

$$\frac{H(\chi)}{\mathbb{E}_\mu H(\chi)} \xrightarrow{\text{dist.}} \operatorname{Exp}(1). \tag{5.3.63}$$

The following Corollary shows how Theorem 5.14 can be applied when we know both that $\mathbb{E}_\mu H(\chi)$ is lower bounded by the density of χ under the stationary measure, and that $(\Gamma, \mathbb{T})$ is rapidly mixing.

Corollary 5.16. *Let* $(\Gamma, \mathbb{T})$ *be an origin-fixation process with stationary distribution* μ. *Let* $\chi \subseteq \Gamma$ *a target set such that* $\mu(\chi) t_{mix}^2 = o(1)$. *Suppose that there is some* $\chi_1 \subseteq \chi$ *such that* $\mu(\chi \backslash \chi_1) = o(\mu(\chi))$ *and that*

$$\mathbb{T}(\alpha, \chi \backslash \{\alpha\}) = o(1) \tag{5.3.64}$$

for all $\alpha \in \chi_1$. *Then we have*

$$\sup_t \left| \mathbb{P}_\mu \left\{ H(\chi) > t \right\} - \exp\left(-t / \mathbb{E}_\mu H(\chi)\right) \right| \to 0. \tag{5.3.65}$$

Proof. First, note that by Theorem 5.12, we have

$$\frac{t_{\mathrm{mix}}}{\mathbb{E}_\mu H(\chi)} \lesssim \mathcal{O}\left(t_{\mathrm{mix}}^2 \mu(\chi)\right) = o(1). \tag{5.3.66}$$

Then, apply Theorem 5.14 to $H(\chi)$. Thus,

$$\sup_t \left| \mathbb{P}_\mu \left\{ H(\chi) > t \right\} - \exp\left(-t / \mathbb{E}_\mu H(\chi)\right) \right| \leqslant \frac{C t_{\mathrm{mix}} \left(1 + \log^+\left(\mathbb{E}_\mu H(\chi) / t_{\mathrm{mix}}\right)\right)}{\mathbb{E}_\mu H(\chi)} = o(1). \tag{5.3.67}$$

■

The following lemma shows that under mild assumptions, it is not necessary to start the origin-fixation process in its stationary distribution.

Lemma 5.17. *Define* s *as in Equation* (5.3.54). *Suppose for some* $\alpha \in \Gamma$ *that* $\mathbb{P}_\alpha \left\{ H(\chi) \leqslant s \right\} = o(1)$ *and that*

$$\frac{t_{mix} \left(1 + \log^+\left(\mathbb{E}_\mu H(\chi) / t_{mix}\right)\right)}{\mathbb{E}_\mu H(\chi)} = o(1), \tag{5.3.68}$$

then

$$\sup_t \left| \mathbb{P}_\alpha \left\{ H(\chi) > t \right\} - \exp\left(-t / \mathbb{E}_\mu H(\chi)\right) \right| \to 0. \tag{5.3.69}$$

Proof. First, note

$$\begin{aligned}
\|\mathbb{P}_\alpha\{x_s = \cdot | H(\chi) > s\} - \mu\|_{\mathrm{TV}} &\leqslant \|\mathbb{P}_\alpha\{x_s = \cdot | H(\chi) > s\} - \mathbb{P}_\alpha\{x_s = \cdot\}\|_{\mathrm{TV}} + \|\mathbb{P}_\alpha\{x_s = \cdot\} - \mu\|_{\mathrm{TV}} \\
&\leqslant \mathbb{P}_\alpha\{H(\chi) \leqslant s\} + \frac{t_{\mathrm{mix}}}{\mathbb{E}_\mu H(\chi)} \qquad (5.3.70)
\end{aligned}$$

by the definition of s and t_{mix}. Next, we have

$$\begin{aligned}
|\mathbb{P}_\alpha\{H(\chi) > t\} - \mathbb{P}_\mu\{H(\chi) > t\}| &\leqslant |\mathbb{P}_\alpha\{H(\chi) > t - s | H(\chi) > s\} - \mathbb{P}_\mu\{H(\chi) > t - s | H(\chi) > s\}| \\
&\quad + \mathbb{P}_\alpha\{H(\chi) \leqslant s\} + \mathbb{P}_\mu\{H(\chi) \leqslant s\} \\
&\leqslant \|\mathbb{P}_\alpha\{x_s = \cdot | H(\chi) > s\} - \mu\|_{\mathrm{TV}} + \|\mu - \mathbb{P}_\mu\{x_s = \cdot | H(\chi) > s\}\|_{\mathrm{TV}} \\
&\quad + \mathbb{P}_\alpha\{H(\chi) \leqslant s\} + \mathbb{P}_\mu\{H(\chi) \leqslant s\} \\
&\leqslant \mathbb{P}_\alpha\{H(\chi) \leqslant s\} + \frac{t_{\mathrm{mix}}}{\mathbb{E}_\mu H(\chi)} + \mathbb{P}_\mu\{H(\chi) \leqslant s\} \\
&\quad + \mathbb{P}_\alpha\{H(\chi) \leqslant s\} + \mathbb{P}_\mu\{H(\chi) \leqslant s\} \\
&\leqslant 2\mathbb{P}_\alpha\{H(\chi) \leqslant s\} + 2\frac{2s + t_{\mathrm{mix}}}{\mathbb{E}_\mu H(\chi)} \\
&= o(1) \qquad (5.3.71)
\end{aligned}$$

by (5.3.60). Taking the supremum over t in Equation (5.3.71) and applying Theorem 5.14 to see

$$\begin{aligned}
\sup_t |\mathbb{P}_\alpha\{H(\chi) > t\} - \exp(-t/\mathbb{E}_\mu H(\chi))| &\leqslant \sup_t |\mathbb{P}_\alpha\{H(\chi) > t\} - \mathbb{P}_\mu\{H(\chi) > t\}| \\
&\quad + \sup_t |\mathbb{P}_\mu\{H(\chi) > t\} - \exp(-t/\mathbb{E}_\mu H(\chi))| \\
&= o(1) + o(1) \qquad (5.3.72)
\end{aligned}$$

finishes the proof. ■

If individual evolutionary processes cannot find targets in polynomial time, then perhaps the success of evolution is based on the fact that many populations are searching independently and in parallel for a particular discovery [236]. We now prove a quick lemma on multiple independent searches, which says K independent searches in parallel is roughly equivalent to a single process searching for K times as long. This is not surprising when $H(\chi)$ has an exponential distribution.

Lemma 5.18. *Let* $(\Gamma, \mathbb{T})$ *be an origin-fixation process with stationary distribution* μ *and mixing time* t_{mix}*. Assume that*

$$\Delta := \frac{Ct_{mix}\left(1 + \log^{+}\left(\mathbb{E}_{\mu} H(\chi)/t_{mix}\right)\right)}{\mathbb{E}_{\mu} H(\chi)} < 1. \tag{5.3.73}$$

Let $H_1(\chi), \ldots, H_K(\chi)$ *be* K *discovery times derived from independent origin-fixation processes with* μ *as their initial condition, then*

$$\mathbb{P}_{\mu}\left\{\min_{i \in [\![K]\!]}\left\{\frac{H_i(\chi)}{\mathbb{E}_{\mu} H_1(\chi)}\right\} > t\right\} \sim e^{-Kt}. \tag{5.3.74}$$

Proof. Since the $H_i(\chi)$ are independent, we have

$$\mathbb{P}_{\mu}\left\{\min_{i \in [\![K]\!]}\left\{\frac{H_i(\chi)}{\mathbb{E}_{\mu} H_1(\chi)}\right\} > t\right\} = \mathbb{P}_{\mu}\left\{\frac{H_1(\chi)}{\mathbb{E}_{\mu} H_1(\chi)} > t\right\}^{K} = \left(e^{-t} + o(1)\right)^{K} = e^{-Kt} + o(1), \tag{5.3.75}$$

by Theorem 5.14. ■

5.3.6. Putting everything together. In this subsection, we give an example of how the results in Chapters 2 and 5 can be put together. Let $(\Gamma_n^{(p)}, \mathcal{M}^{(p)})$ the single point mutation process on the bond disorder hypercube (See Definition 2.26) and assume $p > p_{c,3}$. We know that the stationary distribution π is uniform and, from Theorem 2.27, that the mixing time of this mutation process is $\tilde{t}_{\text{mix}} = \mathcal{O}\left(n^2 \log(n)/\varepsilon\right)$ with probability $1 - o(1)$.

For each $\boldsymbol{\alpha} \in \Gamma_n^{(p)}$ assign a fitness

$$\mathcal{F}(\boldsymbol{\alpha}) = 1 + \frac{f_{\boldsymbol{\alpha}}}{N} \tag{5.3.76}$$

such that the $f_{\boldsymbol{\alpha}}$ are i.i.d. Gaussian $\mathcal{N}\left(0, \sigma^2\right)$ for $\sigma = \mathcal{O}(1)$. We need to bound two statistics related to these random fitnesses. First, we have

$$\mathbb{P}\left\{\min_{\boldsymbol{\alpha}} f_{\boldsymbol{\alpha}} < -\theta\right\} = 1 - \mathbb{P}\left\{\min_{\boldsymbol{\alpha}} f_{\boldsymbol{\alpha}} \geqslant -\theta\right\} = 1 - \mathbb{P}\left\{f_{\mathbf{0}} \geqslant -\theta\right\}^{\kappa^n} = 1 - \left(1 - \mathbb{P}\left\{f_{\mathbf{0}} < -\theta\right\}\right)^{\kappa^n}. \tag{5.3.77}$$

Now, we need a tail bound for the Gaussian:

$$\int_{\theta}^{\infty} \frac{e^{-\frac{x^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}} \mathrm{d}x \leqslant \int_{\theta}^{\infty} \frac{x}{\theta} \frac{e^{-\frac{x^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}} \mathrm{d}x = \frac{\exp\left(-\frac{\theta^2}{2\sigma^2}\right)}{\sqrt{2\pi x^2/\sigma^2}\theta}. \tag{5.3.78}$$

By Equation (5.3.78), we have

$$(1 - \mathbb{P}\{f_{\mathbf{0}} < -\theta\})^{\kappa^n} \geqslant \left(1 - \frac{\exp\left(-\theta^2/2\sigma^2\right)}{\theta\sqrt{2\pi/\sigma^2}}\right)^{\kappa^n} = \left(1 - \frac{1}{\kappa^n\sqrt{\pi n \log \kappa}}\right)^{\kappa^n} = \exp\left(-1/\sqrt{\pi n \log \kappa}\right) \tag{5.3.79}$$

for $\theta = \sqrt{2\sigma^2 n \log \kappa}$. Therefore,

$$\mathbb{P}\left\{\min_{\boldsymbol{\alpha}} f_{\boldsymbol{\alpha}} < -\sqrt{2\sigma^2 n \log \kappa}\right\} \to 0. \tag{5.3.80}$$

Note that in a similar way, we can show

$$\mathbb{P}\left\{\max_{\boldsymbol{\alpha}} f_{\boldsymbol{\alpha}} > \sqrt{2\sigma^2 n \log \kappa}\right\} \to 0. \tag{5.3.81}$$

For the second statistic, first observe that for all $x, y \in \mathbb{R}$ that

$$\frac{y - x}{e^{-x} - e^{-y}} \leqslant e^x(y - x). \tag{5.3.82}$$

Then, applying Equation (5.3.82), we find

$$\max_{\boldsymbol{\alpha},\boldsymbol{\beta}:\mathcal{D}(\boldsymbol{\alpha},\boldsymbol{\beta})=1} \frac{f_{\boldsymbol{\beta}} - f_{\boldsymbol{\alpha}}}{\exp\left(-f_{\boldsymbol{\alpha}}\right) - \exp\left(-f_{\boldsymbol{\beta}}\right)} = \max_{\boldsymbol{\alpha}} e^{f_{\boldsymbol{\alpha}}} \max_{\boldsymbol{\beta}:\mathcal{D}(\boldsymbol{\alpha},\boldsymbol{\beta})=1} \left(f_{\boldsymbol{\beta}} - f_{\boldsymbol{\alpha}}\right). \tag{5.3.83}$$

Now, we bound the right-hand side of Equation (5.3.83) into two parts. First if $f_{\boldsymbol{\alpha}} \leqslant \frac{1}{2}\log n$, we have

$$\max_{\boldsymbol{\alpha}:f_{\boldsymbol{\alpha}} \leqslant \frac{1}{2}\log n} e^{f_{\boldsymbol{\alpha}}} \max_{\boldsymbol{\beta}:\mathcal{D}(\boldsymbol{\alpha},\boldsymbol{\beta})=1} \left(f_{\boldsymbol{\beta}} - f_{\boldsymbol{\alpha}}\right) \leqslant \max_{\boldsymbol{\alpha},\boldsymbol{\beta}:f_{\boldsymbol{\alpha}} \leqslant \frac{1}{2}\log n} e^{f_{\boldsymbol{\alpha}}} f_{\boldsymbol{\beta}} \leqslant \sqrt{2\sigma^2 n^2 \log \kappa} \tag{5.3.84}$$

by Equation (5.3.81) with probability $1-o(1)$. Second, if $f_{\boldsymbol{\alpha}} > \frac{1}{2}\log n$, we see

$$\begin{aligned}
\mathbb{E}\left[\max_{\boldsymbol{\alpha}: f_{\boldsymbol{\alpha}} > \frac{1}{2}\log n} e^{f_{\boldsymbol{\alpha}}} \max_{\boldsymbol{\beta}:\mathcal{D}(\boldsymbol{\alpha},\boldsymbol{\beta})=1} (f_{\boldsymbol{\beta}} - f_{\boldsymbol{\alpha}})\right] &= \mathbb{E}\left[\max_{\boldsymbol{\alpha}: f_{\boldsymbol{\alpha}} > \frac{1}{2}\log n} e^{f_{\boldsymbol{\alpha}}} \mathbb{E}\left[\max_{\boldsymbol{\beta}:\mathcal{D}(\boldsymbol{\alpha},\boldsymbol{\beta})=1} f_{\boldsymbol{\beta}} - f_{\boldsymbol{\alpha}} | f_{\boldsymbol{\alpha}}\right]\right] \\
&\leqslant C\mathbb{E}\left[\max_{\boldsymbol{\alpha}: f_{\boldsymbol{\alpha}} > \frac{1}{2}\log n} e^{f_{\boldsymbol{\alpha}}} \mathbb{P}\left\{\max_{\boldsymbol{\beta}:\mathcal{D}(\boldsymbol{\alpha},\boldsymbol{\beta})=1} f_{\boldsymbol{\beta}} > f_{\boldsymbol{\alpha}} | f_{\boldsymbol{\alpha}}\right\}\right] \\
&\leqslant C\mathbb{E}\left[\max_{\boldsymbol{\alpha}: f_{\boldsymbol{\alpha}} > \frac{1}{2}\log n} e^{f_{\boldsymbol{\alpha}}} \left(1 - \left(1 - \frac{\exp\left(-\frac{f_{\boldsymbol{\alpha}}^2}{2\sigma^2}\right)}{f_{\boldsymbol{\alpha}}\sqrt{2\pi\sigma^2}}\right)^{(\kappa-1)n}\right)\right] \\
&\leqslant C\mathbb{E}\left[\max_{\boldsymbol{\alpha}: f_{\boldsymbol{\alpha}} > \frac{1}{2}\log n} \frac{(\kappa-1)n\exp\left(f_{\boldsymbol{\alpha}} - f_{\boldsymbol{\alpha}}^2/(2\sigma^2)\right)}{\sqrt{2\pi\sigma^2}f_{\boldsymbol{\alpha}}}\right] \\
&= o(1),
\end{aligned} \tag{5.3.85}$$

where we used the fact $(f_{\boldsymbol{\beta}} - f_{\boldsymbol{\alpha}} | f_{\boldsymbol{\alpha}}) \sim \mathcal{N}(-f_{\boldsymbol{\alpha}}, \sigma^2)$ and Equation (5.3.78). Therefore by Markov's inequality B.1, we see

$$\mathbb{P}\left\{\max_{\boldsymbol{\alpha}: f_{\boldsymbol{\alpha}} > \frac{1}{2}\log n} e^{f_{\boldsymbol{\alpha}}} \max_{\boldsymbol{\beta}:\mathcal{D}(\boldsymbol{\alpha},\boldsymbol{\beta})=1} (f_{\boldsymbol{\beta}} - f_{\boldsymbol{\alpha}}) > 1\right\} = o(1). \tag{5.3.86}$$

Putting together Equations (5.3.84) and (5.3.85), we have

$$\max_{\boldsymbol{\alpha},\boldsymbol{\beta}:\mathcal{D}(\boldsymbol{\alpha},\boldsymbol{\beta})=1} \frac{f_{\boldsymbol{\beta}} - f_{\boldsymbol{\alpha}}}{\exp(-f_{\boldsymbol{\alpha}}) - \exp(-f_{\boldsymbol{\beta}})} \leqslant \sqrt{2\sigma^2 n^2 \log\kappa} \tag{5.3.87}$$

with probability $1-o(1)$. Also denote $M_{\mathcal{F}} := \mathbb{E}e^{f_{\mathbf{0}}} = e^{\sigma^2/2} = \mathcal{O}(1)$ and

$$\mathcal{M}_{\mathcal{F},2} := \mathbb{E}\frac{f_{\boldsymbol{\beta}} - f_{\boldsymbol{\alpha}}}{\exp(-f_{\boldsymbol{\alpha}}) - \exp(-f_{\boldsymbol{\beta}})} = \mathcal{O}(1). \tag{5.3.88}$$

Suppose that we generate an origin-fixation process based on $(\Gamma_n^{(p)}, \mathcal{M}^{(p)}, \mathcal{F})$ and $\mathbb{D}$ given by the Moran process in a well-mixed population from Definition 3.13. Then by Lemma 5.5, the mixing time t_{mix} of this origin-fixation process can be bounded by

$$
\begin{aligned}
&\left(\max_{\boldsymbol{\alpha},\boldsymbol{\beta}:\mathcal{D}(\boldsymbol{\alpha},\boldsymbol{\beta})=1} \frac{\frac{f_{\boldsymbol{\beta}}-f_{\boldsymbol{\alpha}}}{\exp(-f_{\boldsymbol{\alpha}})-\exp\left(-f_{\boldsymbol{\beta}}\right)}}{\frac{1}{\kappa^n}\sum_{\boldsymbol{\alpha}'}\exp\left(f_{\boldsymbol{\alpha}'}\right)}\right)\left(\log\sum_{\boldsymbol{\alpha}'}\exp\left(f_{\boldsymbol{\alpha}'}\right)/\kappa^n - \min_{\boldsymbol{\alpha}}\log\left(\exp(f_{\boldsymbol{\alpha}})/\kappa^n\right)\right)\frac{n^2\log(n)}{\varepsilon}\\
\leqslant&\left(\max_{\boldsymbol{\alpha},\boldsymbol{\beta}:\mathcal{D}(\boldsymbol{\alpha},\boldsymbol{\beta})=1} \frac{\frac{f_{\boldsymbol{\beta}}-f_{\boldsymbol{\alpha}}}{\exp(-f_{\boldsymbol{\alpha}})-\exp\left(-f_{\boldsymbol{\beta}}\right)}}{M_{\mathcal{F}}}\right)\left(n\log\kappa+\log M_{\mathcal{F}}-\min_{\boldsymbol{\alpha}} f_{\boldsymbol{\alpha}}\right)\frac{n^2\log(n)}{\varepsilon}\\
\leqslant&\left(M_{\mathcal{F}}\sqrt{2\sigma^2n^2\log\kappa}\right)\left(n\log\kappa+\sqrt{2\sigma^2n\log\kappa}\right)\frac{n^2\log(n)}{\varepsilon}\\
=&\mathcal{O}\left(\frac{n^4\log n}{\varepsilon}\right)
\end{aligned}
\tag{5.3.89}
$$

with probability $1-o(1)$ by Equations (5.3.80) and (5.3.87).

Chose a target set $\chi=\chi_{q,d}$ as in Definition (5.9) with $d=\mathcal{O}(1)$ and $q=o(n^d)$. Recall $I_{\boldsymbol{\alpha}}=\mathbf{1}\left(\boldsymbol{\alpha}\in C_{\chi_d^{(p)}}\right)$. Then we see

$$
\mathcal{Q}(\chi,\chi^c)=\kappa^{-n}\mathbb{T}(\partial\chi,\partial\chi^c)\leqslant\kappa^{-n}\sum_{\boldsymbol{\alpha}\in\Gamma_n^{(p)}} I_{\boldsymbol{\alpha}}\sum_{\boldsymbol{\beta}:\mathcal{D}(\boldsymbol{\beta},\boldsymbol{\alpha})=d}\;\sum_{\boldsymbol{\beta}:\mathcal{D}(\boldsymbol{\beta},\boldsymbol{\alpha}')=d+1}\mathbb{T}(\boldsymbol{\beta},\boldsymbol{\beta}'). \tag{5.3.90}
$$

Let $x_{(\boldsymbol{\beta},\boldsymbol{\beta}')}$ be the edge disorder indicators from Definition 2.26, then

$$
\begin{aligned}
\mathcal{Q}(\chi,\chi^c)&\lesssim\kappa^{-n}\sum_{\boldsymbol{\alpha}\in\Gamma_n^{(p)}} I_{\boldsymbol{\alpha}}\sum_{\boldsymbol{\beta}:\mathcal{D}(\boldsymbol{\beta},\boldsymbol{\alpha})=d}\;\sum_{\boldsymbol{\beta}:\mathcal{D}(\boldsymbol{\beta},\boldsymbol{\alpha}')=d+1}\frac{f_{\boldsymbol{\beta}'}-f_{\boldsymbol{\beta}}}{\exp\left(-f_{\boldsymbol{\beta}}\right)-\exp\left(-f_{\boldsymbol{\beta}'}\right)}\frac{\varepsilon x_{(\boldsymbol{\beta},\boldsymbol{\beta}')}}{n(\kappa-1)}\\
&\leqslant Cp\mathcal{M}_{\mathcal{F},2}\frac{\varepsilon\binom{n}{d}(\kappa-1)^{d+1}(n-d)}{(\kappa-1)n}\frac{1}{\kappa^n}\sum_{\boldsymbol{\alpha}\in\Gamma_n^{(p)}} I_{\boldsymbol{\alpha}},
\end{aligned}
\tag{5.3.91}
$$

by Lemma B.3, since

$$
\mathbb{E}\left[\frac{f_{\boldsymbol{\beta}'}-f_{\boldsymbol{\beta}}}{\exp\left(-f_{\boldsymbol{\beta}}\right)-\exp\left(-f_{\boldsymbol{\beta}'}\right)}x_{(\boldsymbol{\beta},\boldsymbol{\beta}')}\right]=p\mathcal{M}_{\mathcal{F},2}. \tag{5.3.92}
$$

and

$$
\sum_{\boldsymbol{\beta}:\mathcal{D}(\boldsymbol{\beta},\boldsymbol{\alpha})=d}\;\sum_{\boldsymbol{\beta}:\mathcal{D}(\boldsymbol{\beta},\boldsymbol{\alpha}')=d+1}1=\binom{n}{d}(\kappa-1)^{d+1}(n-d). \tag{5.3.93}
$$

Finally, a similar argument for the sum $\sum_{\boldsymbol{\alpha}\in\Gamma_n^{(p)}} I_{\boldsymbol{\alpha}}$ shows

$$
\mathcal{Q}(\chi,\chi^c)\lesssim\mathcal{O}\left(\varepsilon qn^d\right). \tag{5.3.94}
$$

Thus, applying Lemma 5.11, we find

$$\Omega\left(\frac{1}{\varepsilon q n^d}\right) \lesssim \mathbb{E}_\mu H(\chi) \lesssim \mathcal{O}\left(\frac{n^4 \log n}{\varepsilon q n^d}\right). \tag{5.3.95}$$

Finally, by Theorem 5.14, we have

$$\frac{H(\chi)}{\mathbb{E}_\mu H(\chi)} \xrightarrow{\text{dist.}} \operatorname{Exp}(1) \tag{5.3.96}$$

when

$$q = o\left(\frac{1}{n^{d+4}\log n}\right). \tag{5.3.97}$$

We put the whole discussion above into a Theorem.

Theorem 5.19. *Let* $(\Gamma_n^{(p)}, \mathbb{T})$ *be the origin-fixation process and* $\chi_{q,d}$ *the target set defined above with* $d = \mathcal{O}(1)$ *and* $q = o\left(\frac{1}{n^{d+4}\log n}\right)$, *then*

$$\Omega\left(\frac{1}{\varepsilon q n^d}\right) = \mathbb{E}_\mu H(\chi) = \mathcal{O}\left(\frac{n^4 \log n}{\varepsilon q n^d}\right) \tag{5.3.98}$$

and

$$\frac{H(\chi)}{\mathbb{E}_\mu H(\chi)} \xrightarrow{\textit{dist.}} \operatorname{Exp}(1). \tag{5.3.99}$$

5.4. Regeneration process

There are at least two interpretations of Theorem 5.19. One is to simply accept that this is how evolution makes novel discoveries. Either the specific functionality is very abundant, or there is not a specific functionality and the totality of functionalities that might increase fitness is very abundant. If we accept this interpretation, we must also accept that random sampling would make these discoveries just as quickly and is no less predictable than evolution.

Another interpretation is to view Theorem 5.19 as a negative result. It implies that some subsets are inaccessible to evolution in biological timescales, that is, timescales that are polynomial in n. Specifically, any target set χ such that $\mu(\chi) = c^n$ for some $c < 1$. We saw that simply adding more independent searches (unless it is an exponential number) cannot overcome this infeasibility in Lemma 5.18.

It is possible that we need to revise our assumption that mutation is completely unstructured and independent of phenotypic functionality. Are there specific mechanisms that structure mutation to be relevant to fitness? After all, mutation and fitness are both derived from the mechanism of copying. We now outline one such mechanism that we can analyze mathematically in great generality. We call the mechanism the *regeneration process*. The basic idea is that evolution can solve a new problem efficiently, if it has solved a similar problem already [14]. There is lots of experimental evidence to suggest that this is a reasonable hypothesis [128, 237–239].

Suppose gene duplication or genotype rearrangement can continuously give rise to some starting genotype α. Then the regeneration process, behaves exactly like a normal origin-fixation process, except after a small amount of time (which is give preciously by the mixing time) that we refer to as a period, the process reverts to its initial condition. Note that while we think of this process occurring sequentially, one could imagine several search happening in parallel with the same initial condition. The reason for choosing the mixing time as the period for the regeneration process, is that as we have seen (see (5.3.70) for example) after t_{mix} steps the process behaves similarly to random, independent sampling from μ in terms of the discovery time.

This process becomes interesting when the probability of hitting the target set before the mixing time given that α is the initial condition is large. Even though in a single period the process is unlikely to discover the target set and the discovery time of a single search can still be very large, repeatedly exploiting the initial condition's "closeness" to χ can lead to fast discovery times for the regeneration process. Here, to be as general as possible, we measure "closeness" by

$$\mathbb{P}_\alpha \left\{ H(\chi) \leqslant t_{\text{mix}} \right\}. \tag{5.4.1}$$

However, in most cases this probability can be bounded using assumptions about the distance between α and χ in genotype space—that is, we might assume that

$$\min_\beta \mathcal{D}\left(\alpha, \beta\right) = \mathcal{O}\left(1\right). \tag{5.4.2}$$

As we mentioned, the biological motivation here is that novel functionality that is similar to some preexisting functionality might be close in genotype space. There is good evidence that gene duplication is involved in the emergence of novel genes and our model supports this hypothesis [240, 241].

Note that for the regeneration process to efficiently discover χ, both of these components are necessary: the process must start close to χ and be able to regenerate to this initial condition. Removing either of these assumption results in infeasible discovery times. Note that no selection is necessary here.

The regeneration process formalizes the role of several existing ideas. First, it ties in with the proposal that gene duplications and genome rearrangements are major events leading to the emergence of new genes [128]. Second, evolution can be seen as a tinkerer playing around with small modifications of existing sequences rather than creating entirely new ones [14]. Third, the process is related to Gillespie's suggestion that the starting sequence for an evolutionary search must have high fitness [49]. In our theory, proximity in fitness value is replaced by proximity in genotype space. Another way to view this is that random sampling can be useful if the sampling is done in the right context—that is, a context where the rest of the genotype is structured to make the changes obtained by sampling meaningful for the phenotype. Again, this idea has experimental support [242, 243]. However, our results show that proximity alone is insufficient to break the exponential barrier, and only when combined with the process of regeneration do we see feasible discovery times with high probability.

Definition 5.20. *Let* $(\Gamma, \mathbb{T})$ *be an origin-fixation process with mixing time* t_{mix}*. Let* $\left(x_t^{(k)}\right)_{t\geqslant 0}$ *be an independent copy of the origin-fixation process such that* $x_0^{(k)} = \alpha$ *for all* $k \in \mathbb{N}$*. Then define the regeneration process that regenerates to* $\alpha \in \Gamma$ *by*

$$y_{(k-1)t_{mix}+i} := x_i^{(k)} \tag{5.4.3}$$

for $i \in [\![0, t_{mix} - 1]\!]$ *and* $k \in \mathbb{N}$ *(which defines* y_t *for all* t*). For a target set* $\chi \subseteq \Gamma$*, we denote the discovery time of* χ *under the regeneration process by* $\tilde{H}(\chi)$*:*

$$\tilde{H}(\chi) := \min\left\{t : y_t \in \chi\right\}. \tag{5.4.4}$$

Note that instead of having the process regenerate to α exactly every t_{mix} time steps, we could use a Poissonian clock with rate $1/t_{\mathrm{mix}}$ in a similar way to Definition 2.32. Incorporating the regeneration mutation into $\mathcal{M}$, means the regeneration process is a true Markov chain, but we avoid this here as the results are similar.

Theorem 5.21. *Let* y_t *be a regeneration process that regenerates to* α *such that*

$$k_n := \frac{1}{\mathbb{P}_\alpha\left\{H(\chi) \leqslant t_{mix}\right\}}. \tag{5.4.5}$$

Then

$$\mathbb{P}_\alpha\left\{\tilde{H}(\chi) \geqslant k_n t_{mix} \log\log(n)\right\} = o(1). \tag{5.4.6}$$

Proof. The event $\tilde{H}(\chi) \geqslant k_n t_{\mathrm{mix}} \log\log(n)$ means that the regeneration process did not discover χ in $k_n \log\log(n)$

periods, and since each of these periods is independent, we have

$$\mathbb{P}_\alpha\left\{\tilde{H}(\chi) \geqslant k_n t_{\text{mix}} \log\log(n)\right\} = \left(\mathbb{P}_\alpha\left\{H(\chi) > t_{\text{mix}}\right\}\right)^{k_n \log\log n} = \left(1 - \frac{1}{k_n}\right)^{k_n \log\log n} \lesssim \frac{C}{\log n}. \tag{5.4.7}$$

■

Example 5.22. *Theorem 5.21 gives a simple mechanism to efficiently discover novel functionality under two key assumptions. Consider again the hypercube $\Gamma_n = [\![\kappa]\!]^n$ with $\mathcal{F}(\alpha) = 1$ for all α. Suppose $\mathcal{D}(\alpha, \beta) = d = \mathcal{O}(1)$ for some $\alpha, \beta \in \Gamma_n$. Then*

$$\mathbb{P}_\alpha\left\{H(\beta) \leqslant t_{mix}\right\} \geqslant \mathbb{P}_\alpha\left\{H(\beta) \leqslant d\right\} = \left(\frac{1}{n\kappa}\right)^d, \tag{5.4.8}$$

which is polynomial in n. Therefore, Theorem 5.21 implies

$$\mathbb{P}_\alpha\left\{\tilde{H}(\chi) \geqslant \log^2(n)\frac{\kappa^d n^d}{\varepsilon}\right\} = o(1). \tag{5.4.9}$$

Note specifically the contrast between expected discovery time that is exponential in n in Example 5.13 and time that is polynomial in n in Equation (5.4.9).

6

Randomly structured populations

The stage of evolution is the population of reproducing individuals. The structure of the population is known to affect the dynamics and outcome of evolutionary processes, but analytical results for generic random structures have been lacking. The most general result so far, the isothermal theorem (see Theorem 3.10), assumes the propensity for change in each position is exactly the same (see Definition 3.9), but realistic biological structures are always subject to variation and noise. In this chapter, we consider a finite population under constant selection whose structure is given by a variety of weighted, directed, random graphs; vertices represent individuals and edges interactions between individuals. By establishing a robustness result for the isothermal theorem and using large deviation estimates to understand the typical structure of random graphs, we prove that for a generalization of the Erdős-Rényi model the fixation probability of an invading mutant is approximately the same as that of a mutant of equal fitness in a well-mixed population with high probability. Simulations of perturbed lattices, small-world networks, and scale-free networks behave similarly. We conjecture that the fixation probability in a well-mixed population, $(1-f^{-1})/(1-f^{-n})$, is universal: for many random graph models, the fixation probability approaches the above function uniformly as the graphs become large.

In physics, a system exhibits universality when its macroscopic behavior is independent of the details of its microscopic interactions [156]. Many physical models are conjectured as universal and long programs have been carried out to establish this mathematically [157, 158]. However such universality conjectures have been lacking in biological models.

It is well known that population structure can affect the behavior of evolutionary processes under both constant selection [123,244–249], on which we focus here, and frequency dependent selection [6,169,189,191,250–259]. However, so far deterministic and highly organized population structures have received the most attention [260–264]; while some populations are accurately modeled in this way [265–269], often a random structure is far more appropriate to describe the irregularity of the real world [92, 270–272]. Random population structures have been considered numerically, but analytical results have been lacking [123, 169, 273].

Evolutionary graph theory is a standard model to understand the effects of population structure. Simple one-rooted population structures are able to suppress selection and reduce evolution to a standstill, while intricate, star-like structures can amplify the intensity of selection to all but guarantee the fixation of mutants with arbitrarily slight fitness advantages [4, 123, 274]. The former has been proposed as a model for understanding the necessity of hierarchical lineages of cells to reduce the likelihood of cancer initiation [275]. The isothermal theorem characterizes the population structures whose fixation probability is given exactly by ρ_{M} [123]. This is our first hint of universality but it was not the first time certain quantities were observed as independent of population structure. Maruyama introduced geographical population structure by separating reproduction, which occurs within sub-populations, and migration, which occurs between sub-populations, and found that the fixation probability was the same as that of a well-mixed population structure [172]. In the framework of evolutionary graph theory, Maruyama's model would correspond to a symmetric graph. In this sense his finding is a special case of the isothermal theorem.

However, the assumptions of the isothermal theorem sit on a knife edge—when any small perturbation is made to the graph, the assumptions no longer hold and the original isothermal theorem is silent. In particular, it cannot be applied to directed, random graphs. We address these shortcomings with the robust isothermal theorem 6.1, where we strengthen the forward direction of the isothermal theorem by proving a deterministic statement: we weaken the theorem's assumptions to be only approximately true for a graph G and show that the conclusion is still approximately true, that is, the fixation probability of a general graph ρ_G is approximately equal to ρ. We call this the robust isothermal theorem (RIT) [3].

Theorem 6.1 (Robust isothermal theorem). *Fix $0 \leqslant \varepsilon < 1$. Let $G = (\llbracket N \rrbracket, W)$ be a connected graph. If for all nonempty $S \subsetneq \llbracket N \rrbracket$ we have*

$$\left| \frac{W(S, S^c)}{W(S^c, S)} - 1 \right| \leqslant \varepsilon, \tag{6.0.1}$$

where $W(S, S^c)$ and $W(S^c, S)$ are the sums of the outgoing and ingoing edges respectively, then

$$\sup_{f>0} |\rho_{\mathrm{M}}(f) - \rho_G(f)| \leqslant \varepsilon. \tag{6.0.2}$$

The proof begins by ignoring spacial structure and considering only the number of mutants. Since ρ_G depends only on the ratio of the probability of increasing to the probability of decreasing the number of mutants for each subset, a bound on these ratios and a coupling argument establish that ρ_G is close to ρ_{M}. Finally, the mean value theorem and smoothness properties of ρ_{M} simplify the bound and yield the result. We remark that assumption (6.0.1) is necessary in the sense that there are graphs whose fixation probability is far from ρ_{M} but whose weighted adjacency matrix is arbitrarily close to being doubly stochastic (see Equation (3.3.21) for an example).

The proof verifies something essential for the process: as in physics, our laws should not depend on arbitrarily small quantities nor make disparate predictions for small perturbations of a system. The RIT generalizes the isothermal theorem in this sense; if an isothermal graph is perturbed with strength ε such that the assumption (6.0.1) holds, then its fixation probability is close to that of the original graph (Figure 6.1). There are many ways of rigorously perturbing a graph, so we do not make a precise definition of perturbation here. All we claim is that any perturbation which changes the assumptions of the RIT continuously can be controlled. The RIT has many useful applications and is our first ingredient to universality.

Robustness is essential for the analysis of random graphs. We say a random graph model exhibits universal mean-field behavior if its fixation probability behaves like ρ_{M} as the graph becomes large. That is, as the graphs become large their macroscopic properties, fixation probabilities, are independent of their microscopic structures, the distributions of individual edges. Note that while other have found that some random graph models frequently produce amplifiers, the size of the amplification effect decays to 0 as the graphs become larger [173]. Mathematically, we ask that the random variable $\sup_{f>0} |\rho_G(f) - \rho_{\mathrm{M}}(f)|$ converges in probability to 0, as N goes to infinity. To strength the convergence to almost sure convergence, information about the speed of the convergence is necessary to apply the Borel-Cantelli lemma. For finite values of N, we can require finer control over this convergence such that

$$\mathbb{P}\left\{\sup_{f>0}|\rho_G(f)-\rho_M(f)|\leqslant\delta(N)\right\}=1-\varepsilon(N),\tag{6.0.3}$$

where the functions $\delta(N)=o(1)$ and $\varepsilon(N)=o(1)$ can be specified. For the generalized Erdős-Rényi model [92] where edges are produced independently with fixed probability p (see Definition 6.13) we prove universality. In Section 6.3 we analyze the typical behavior of random graphs and show that with very high probability they satisfy the assumptions of the RIT, giving us the paper's main result:

THEOREM 6.2. *Let $(G_N)_{n\geqslant 1}$ be a family of random graphs where the directed edge weights are chosen independently according to some suitable distribution (the outgoing edges may be normalized to sum to 1 or not). Then there is a constant $C>0$, not dependent on N, such that the fixation probability of a randomly placed mutant of fitness $f>0$ satisfies*

$$|\rho_G(f)-\rho_{\mathrm{M}}(f)|\leqslant\frac{C\,(\log N)^{C+C\xi}}{\sqrt{N}}\tag{6.0.4}$$

uniformly in f with probability greater than $1-\exp(-\nu\,(\log N)^{1+\xi})$, for some positive constants ξ and ν.

The proof applies the RIT to random graphs, showing that with high probability they satisfy assumption (6.0.1) with ε approximately order $1/\sqrt{N}$. This relies on two main results. Using large deviation estimates, we show that the sum of the ingoing edge weights to each vertex (its temperature) are within approximately order $1/\sqrt{N}$ of 1 with high probability (Lemma 6.16) and that sum of the outgoing (and ingoing) edge weights of each subset are at least the same order as the size of the subset or its complement for some uniform constant with high probability (Lemma 6.17).

This theorem isolates the typical behavior of the Moran process on these random structures. It can be interpreted as stating that random processes generating population structures where vertices and edges are treated independently and interchangeably will almost always produce graphs with mean-field behavior. While such processes can generate graphs that do not have mean-field behavior (for example one-rooted or disconnected graphs), these graphs are generated with very low probability as the size of the graphs becomes large. Moreover, it improves upon diffusion approximation methods by explicitly controlling the error rates [276].

The result holds with high probability but sometimes this probability becomes close to 1 only as the graphs become large. The necessary graph size depends on the distribution that the random graph's edge weights are drawn from. In particular, it depends inversely on the parameter p from the generalized Erdős-Rényi model, which is the probability that there is an edge of some weight between two directed vertices. The smaller this parameter the more disordered and sparse the random graphs and the less uniform their vertices' temperatures, which all tend to decrease the control over the graph's closeness to isothermality, (6.0.1). Regardless, our choice of the parameters ξ and ν guarantees that the bound (6.3.54) decays to 0 and that it holds with probability approaching 1 as N becomes large.

We investigated the issues of convergence for small values of N numerically to illustrate our analytical result (Figure 6.2). For Erdős-Rényi random graphs (see Section 6.3 with the distribution chosen as Bernoulli), we generated 10 random graphs according to the procedure outlined in Definition 6.13 for fixed values of $0 < p < 1$. On each graph the Moran process was simulated 10^4 times for various values of $0 \leqslant f \leqslant 10$ to give the empirical fixation probability, that is, the proportion of times that the mutant fixed in the simulation. Degenerate graphs were not excluded from the simulations but rather than estimating their fixation probabilities, we calculated them exactly, so that 1-rooted graphs were given fixation probability $1/N$ and many-rooted and disconnected graphs were given fixation probability 0. Trivially, such 1-rooted graphs are suppressors—that is, the fixation probability of a mutant of fitness $0 < f < 1$ (and a mutant of fitness $f > 1$) is greater than (and less than respectively) the mutant's fixation probability in a well-mixed population—but suppressing graphs without these degenerate properties were also observed [4]. As the graphs become larger their fixation probabilities match ρ_{M} closely and degeneracy becomes highly improbable as predicted by our result.

In addition to the generalized Erdős-Rényi random graphs, we also considered the Watts-Strogatz model and the Barabási-Albert model. The Watts-Strogatz model [270] produces random graphs with small-world properties, that is, high clustering and short average path length. The model has three inputs: a parameter $0 \leqslant \beta \leqslant 1$, the graph size N, and the mean degree $2k$. Typically, the model produces random, undirected graphs, thus, to escape isothermality, it was modified slightly to produce weighted, directed graphs. We do this in the most natural way: we start with a directed $2k$-regular graph where each node is connected to its $2k$ nearest neighbors if the graph is arranged on a cycle (Figure 6.3), and then we rewire each edge to a new vertex chosen uniformly at random with probability β independently. Since the number of edges leaving each vertex is fixed at $2k$, the weight of each edge is exactly $1/(2k)$. Potentially, there can be multiple edges for one vertex to another, which we account for by summing the edge weights. The model may be viewed as an interpolation between an isothermal, $2k$-regular graph and an Erdős-Rényi graph by the parameter β.

Mean-field behavior was observed in the Watts-Strogatz model for all values of the input parameters we simulated (Figure 6.3). While mathematical proof of universality in the Watts-Strogatz model is still needed, there is hope that the techniques of this paper may be applied in this situation as the in-degrees of the vertices are concentrated around 1 for graphs with large degree $2k$.

Unlike the Erdős-Rényi and Watts-Strogatz models, scale-free networks are random graphs where the in-degrees of the vertices follow a power law. Normally, scale-free networks are undirected and unweighted. To produce weighted, directed scale-free networks, we modified the preferential attachment algorithm of Barabási-Albert [277]: we start with a connected cycle and then add directed edges of equal weight in sequence to a randomly selected vertex where the destination of each edge is selected proportional to the in-degree of the current vertices.

Surprisingly, even though there is a sense in which vertices are not treated interchangeably in the preferential attachment algorithm, mean-field behavior was observed in all simulations (Figure 6.4). This is in contrast with the results in Lieberman *et al.* where they observed some amplification in scale-free networks [123]. The scale-free property is emergent and only becomes apparent as the graph becomes large, thus this increases the running time of the Monte Carlo method for estimating the fixation probability. More simulations are required here for conclusive findings and again there are currently no mathematical results.

In summary, we have generalized the isothermal theorem to make it biologically realistic and to increase its technical applicability. The conclusion of the robust isothermal theorem now depends continuously on its assumptions. With this new tool, we have proved analytically that fixation probabilities in a generalized Erdős-Rényi model converge uniformly to the fixation probability of a well-mixed population. In our proof, we identify the reason for this convergence and bound its rate. Thus, we confirm observations from many simulations and give a method of approximation with a specified error. Furthermore, we conjecture that many random graph models exhibit this universal behavior. However, it is easy to construct simple examples of random graphs which do not, thus it still remains to determine the necessary assumptions on the random graph model for it to exhibit universal behavior.

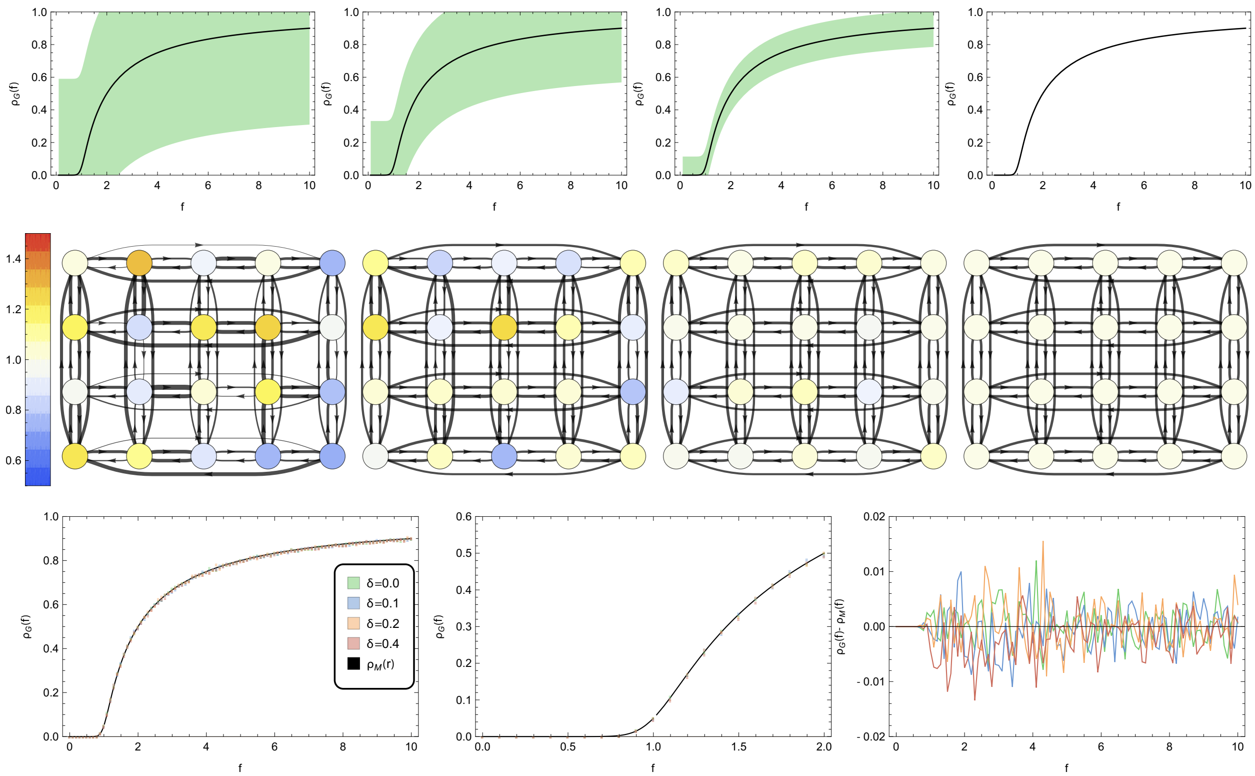

Figure 6.1: The robust isothermal theorem guarantees that the fixation probability of each approximately isothermal graph lies in the green region. Each edge of the 4×4, 2-dimensional torus is perturbed by a uniform random value from $[-\delta/2, \delta/2]$ where the total of the perturbations for one vertex are conditioned to sum to 0. As the perturbation strength decreases through $\delta \in [0.4, 0]$ and the graph approaches isothermality, the bound improves and converges uniformly to the solid black line, ρ_{M}. The figures of square lattices show how random perturbations shift the graphs from isothermality, as the perturbation strength decreases from left to right; we draw each graph with the directed edges' thickness proportional to their weight and the vertices' color given by the sum of the weights of edges pointing to them. In the bottom row empirical estimates of the fixation probabilities (small circles) are plotted against the values predicted by ρ_{M} (solid lines) and, despite the perturbations to the graphs, their fixation probabilities lie close to ρ_{M}.

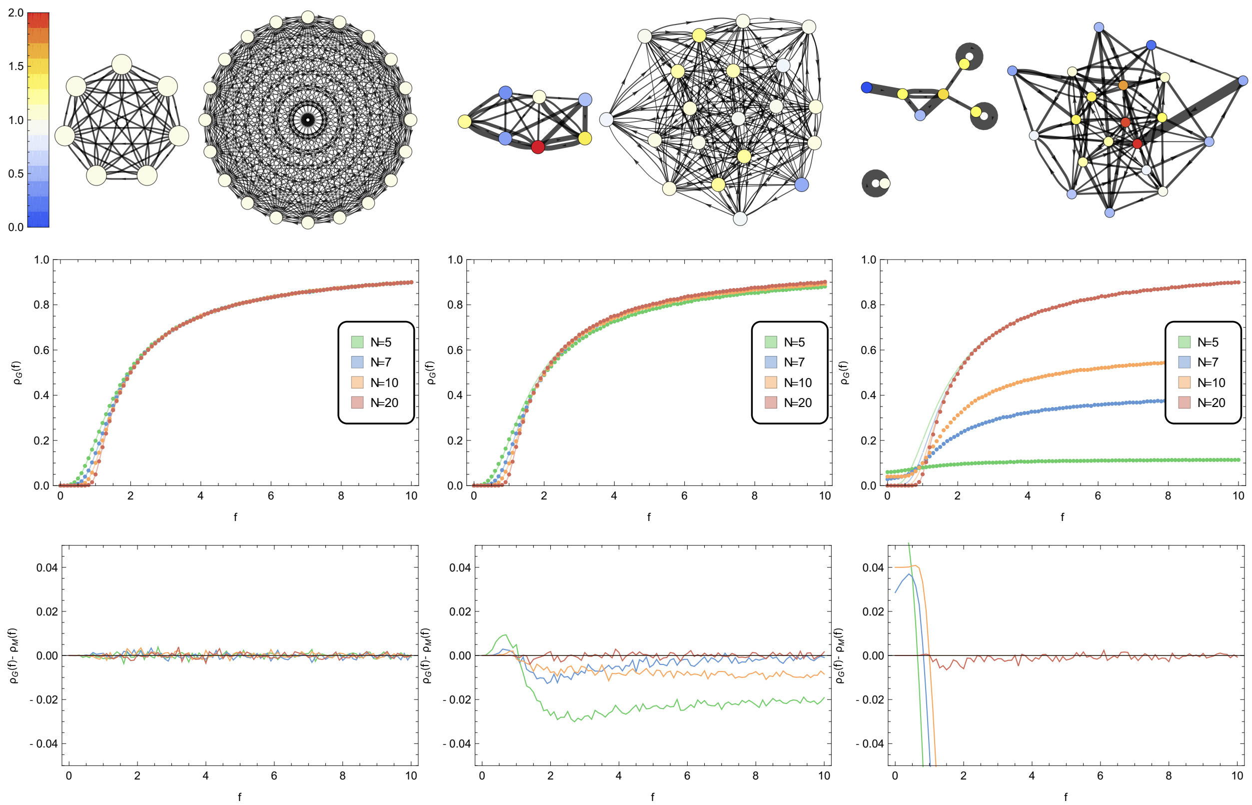

Figure 6.2: The fixation probability of the generalized Erdős-Rényi random graphs converge uniformly to ρ_{M}. The three columns from left to right correspond to Erdős-Rényi random graphs with decreasing connection probabilities $p = 1$, $p = 0.6$, and $p = 0.3$. The representative random graphs in the top row show both the increasing sparsity and disorder as p decreases and the elimination of degeneracy (rootedness and disconnectedness) and the increasing uniformity of temperature as the graph sizes increase. In the middle row empirical estimates of the fixation probabilities (small circles) are plotted against the values predicted by ρ_{M} (solid lines). When $p = 1$ the graphs are isothermal and thus correspond exactly to their predicted values which can be seen even more clearly in the bottom row, where the difference of the empirical fixation probabilities and their predicted values display as stochastic fluctuations about 0. For $p = 0.6$ and $p = 0.3$, the convergence of the empirical values to ρ_{M} as the graphs increases in size is apparent. Smaller graphs are typically suppressors as illustrated by the clear sign change at $r = 1$ in the difference of empirical and predicted values, whereas larger graphs fluctuate about 0. This phenomenon is not due only to the higher probability of obtaining degenerate graphs—simulations produced strongly connected, small suppressors. Moreover, the convergence is patently slower in N for smaller values of p.

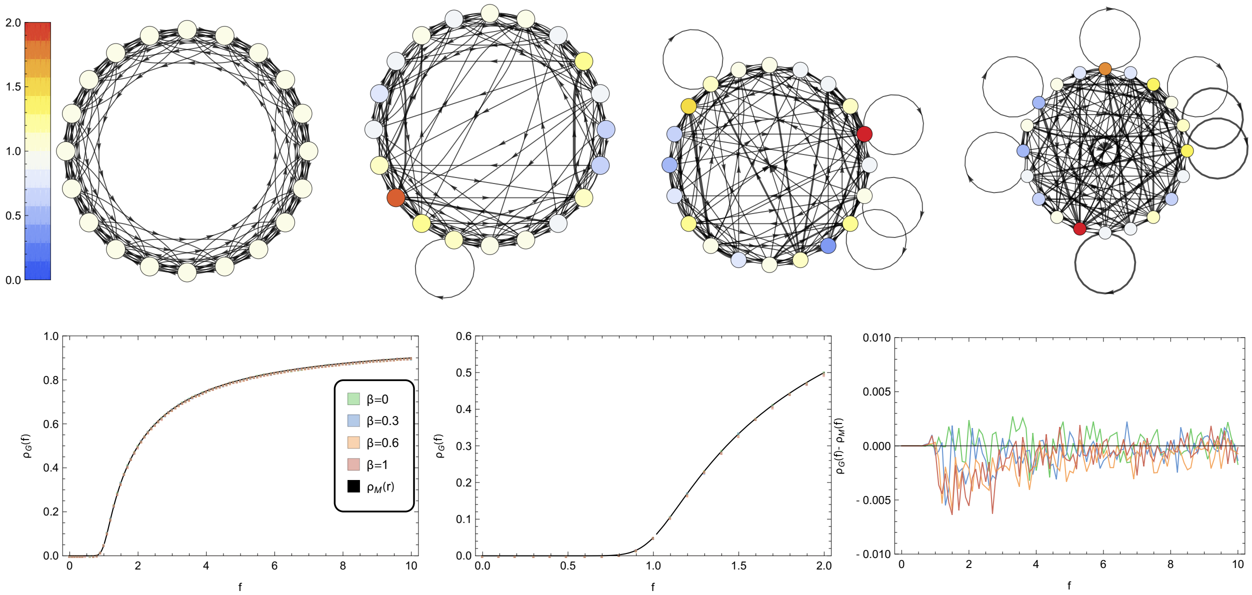

Figure 6.3: Small-world networks also show universal behavior. Representative Watts-Strogatz random graphs display increasing disorder as the rewiring probability β increases from 0 to 1, which may be viewed as an interpolation between an isothermal graph and an Erdős-Rényi random graph. For all values of β the correspondence to ρ_{M} is striking but mathematical proof is lacking.

6.1. Robust isothermal theorem

Theorem 6.3 (Robust isothermal theorem). *Fix $0 \leqslant \varepsilon < 1$. Let $G = ([\![N]\!], W)$ be a connected graph. If for all nonempty $S \subsetneq [\![N]\!]$ we have*

$$\left| \frac{W(S, S^c)}{W(S^c, S)} - 1 \right| \leqslant \varepsilon, \tag{6.1.1}$$

then

$$\sup_{f>0} |\rho_{\mathrm{M}}(f) - \rho_G(f)| \leqslant \varepsilon. \tag{6.1.2}$$

Proof. The state of the process is recorded with $S_t := \{i : \mathbf{x}_t(i) = \beta\}$, that is the subset of individuals who are of type β at time t. To briefly outline the proof, we begin by projecting the process from S_t to $|S_t|$. Next we consider the ratio of the probability of increasing the number of mutants to the probability of decreasing the number of mutants. By bounding this ratio we can use a coupling argument to establish that the fixation probability of the process is close to ρ_{M}. Finally, we use the mean value theorem and smoothness properties of ρ_{M} to simplify our bound and obtain the result.

Just as in the proof of the original isothermal theorem, we make the projection of the state space of all subsets

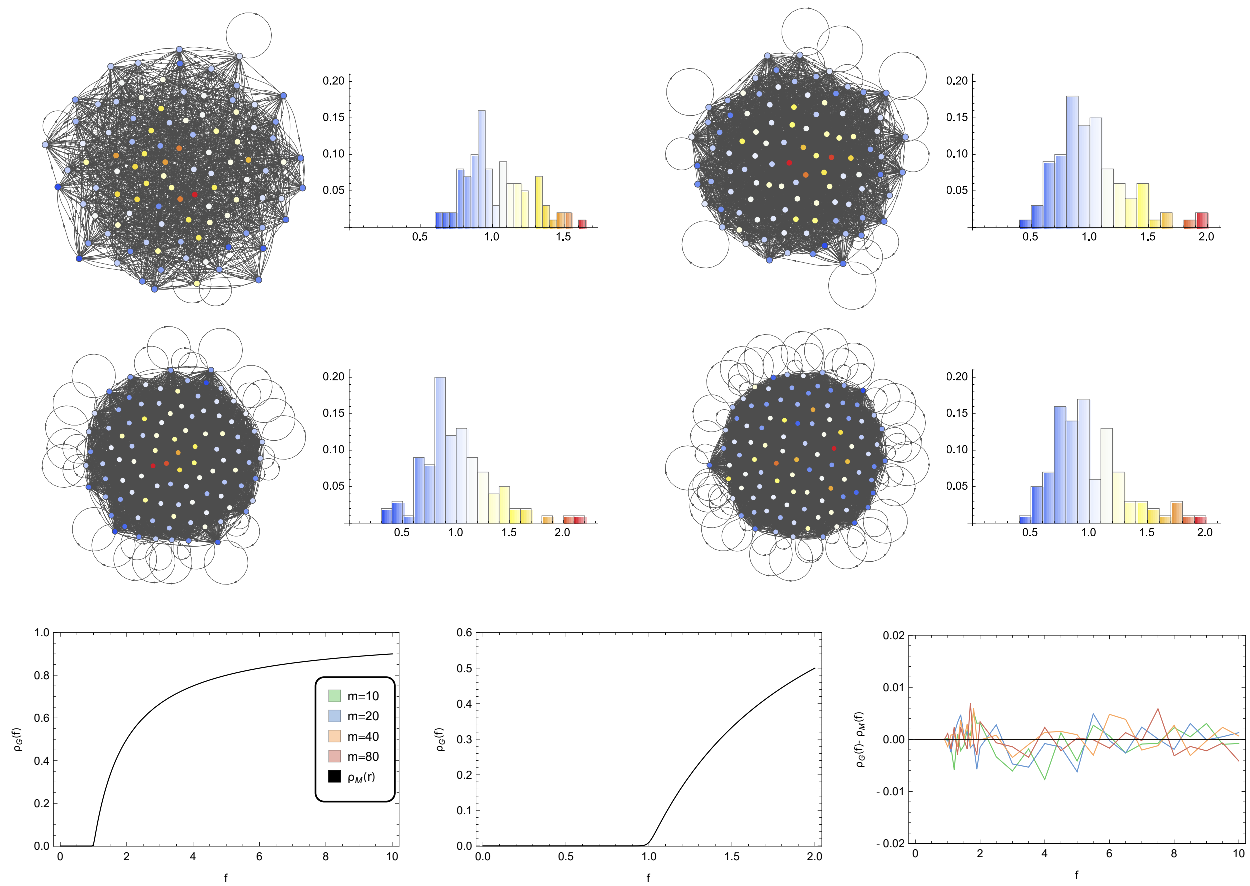

Figure 6.4: Simulations on graphs generated by preferential attachment yield fixation probabilities close to ρ_{M}. Several scale-free networks with varying out degrees, $m = 10$, $m = 20$, $m = 40$, and $m = 80$, were generated using a preferential attachment algorithm. Histograms of the sum of the weights of edges pointing to each vertex are plotted next to each graph, however, the small graphs size limits the resemblance to a power law. Given the comparatively large size of the graphs, only a restricted number of simulations were performed, but the simulations corresponded to ρ_{M} without a tendency to amplify or suppress. More extensive work is required.

of V, which records exactly which vertices are mutants, to the simpler state space $\{0, 1, \ldots, N\}$, which records only the number of mutants. The problem with making this projection in general is that the transition probabilities from one subset to another can depend on the structure of a subset not merely the number of mutants. However, it is clear that the only quantities which affect the fixation probability are the ratios of the probability of increasing the number of mutants to the probability of decreasing the number of mutants in a particular state S.

Define $p_+(S)$ and $p_-(S)$ as the probability that in the next step the number of mutants in the population increases and decreases by one respectively. Note these quantities do not sum to 1, as the number of mutants may remain constant. Thus

$$p_+(S) = \frac{W(S, S^c) f}{W(S, S^c) f + W(S^c, S)} \quad \text{and} \quad p_-(S) = \frac{W(S^c, S)}{W(S, S^c) f + W(S^c, S)}, \tag{6.1.3}$$

which gives, when the two equations are divided,

$$\frac{p_+(S)}{p_-(S)} = f \frac{W(S, S^c)}{W(S^c, S)}. \tag{6.1.4}$$

By assumption (6.1.1),

$$f(1 - \varepsilon) \leqslant \frac{p_+(S)}{p_-(S)} \leqslant f(1 + \varepsilon). \tag{6.1.5}$$

This states that the ratio of the probabilities of increasing to decreasing the number of mutants in any state S is approximately proportional to f.

If for some graph $G' = (\llbracket N \rrbracket, W')$ we have $p_+(S)/p_-(S) = f(1 \pm \varepsilon)$ for all $S \subseteq \llbracket N \rrbracket$, then by the standard result for fixation probabilities in birth-death processes, its fixation probability is given by

$$\rho_{G'}(f) = \rho_{\mathrm{M}}(f \pm f\varepsilon) = \frac{1 - (f \pm \varepsilon f)^{-1}}{1 - (f \pm \varepsilon f)^{-N}}. \tag{6.1.6}$$

From (6.1.5) and (6.1.6) we would like to conclude that

$$\rho_{\mathrm{M}}(f - f\varepsilon) = \frac{1 - (f - \varepsilon f)^{-1}}{1 - (f - \varepsilon f)^{-N}} \leqslant \rho_G(f) \leqslant \frac{1 - (f + \varepsilon f)^{-1}}{1 - (f + \varepsilon f)^{-N}} = \rho_{\mathrm{M}}(f + f\varepsilon). \tag{6.1.7}$$

The upper bound is given by taking the maximum allowed value for the probability of increasing the number of mutants relative to the probability of decreasing the number of mutants. For the lower bound we use the opposite.

This intuitive result can be proved with a coupling argument. We can couple the Moran process S_t of a mutant of fitness f on G with another process Y defined as follows: Y has state space $\{0, \dots, N\}$ (with 0 and N absorbing) and Y starts at 1. We couple Y to S_t as follows:

1. if $|S_t|$ decreases by 1, then Y must also decrease by 1;
2. if $|S_t|$ increases by 1, then independently Y increases by 1 with probability

$$\frac{p_+(S) + p_-(S)}{p_+(S)} \frac{f(1-\varepsilon)}{1 + f(1-\varepsilon)} \tag{6.1.8}$$

   (which is less than or equal to 1 by assumption (6.1.5)), else Y decreases by 1;
3. otherwise Y remains constant.

Note that marginally Y is a simple random walk on $\{0, \dots, n\}$ with forward bias $f(1-\varepsilon)$, since by the law of total probability, defining $Y^{\pm}$ as the event that the Y changes by ± 1 the next time it changes, we see

$$\begin{aligned}
\mathbb{P}\left[Y^+\right] &= \mathbb{P}\left[Y^+ \mid |X_{t+1}| = |X_t| + 1\right] \mathbb{P}\left[|X_{t+1}| = |X_t| + 1\right] \\
&= \frac{p_+(X_t) + p_-(X_t)}{p_+(X_t)} \frac{f(1-\varepsilon)}{1 + f(1-\varepsilon)} p_+(X_t) \\
&= \left(p_+(X_t) + p_-(X_t)\right) \frac{f(1-\varepsilon)}{1 + f(1-\varepsilon)}
\end{aligned} \tag{6.1.9}$$

and

$$\begin{aligned}
\mathbb{P}\left[Y^-\right] &= \mathbb{P}\left[Y^- \mid |X_{t+1}| = |X_t| + 1\right] \mathbb{P}\left[|X_{t+1}| = |X_t| + 1\right] \\
&\quad + \mathbb{P}\left[Y^- \mid |X_{t+1}| = |X_t| - 1\right] \mathbb{P}\left[|X_{t+1}| = |X_t| - 1\right] \\
&= \left(1 - \frac{p_+(X_t) + p_-(X_t)}{p_+(X_t)} \frac{f(1-\varepsilon)}{1 + f(1-\varepsilon)}\right) p_+(X_t) + p_-(X_t) \\
&= \frac{p_+(X_t) + p_-(X_t)}{1 + f(1-\varepsilon)}.
\end{aligned} \tag{6.1.10}$$

Thus the probability that Y reaches N before it reaches 0 is given by

$$\frac{1 - (f - \varepsilon f)^{-1}}{1 - (f - \varepsilon f)^{-N}}. \tag{6.1.11}$$

However, because the processes are coupled we have $Y \leqslant |S_t|$ and thus if $Y = N$, then the mutant has fixed in the process S_t. Equation (6.1.11) immediately implies the lower bound in (6.1.7). A similar coupling yields the upper bound. Thus

$$\rho_{\mathrm{M}}(f - f\varepsilon) - \rho_{\mathrm{M}}(f) \leqslant \rho_G(f) - \rho_{\mathrm{M}}(f) \leqslant \rho_{\mathrm{M}}(f + f\varepsilon) - \rho_{\mathrm{M}}(f). \tag{6.1.12}$$

By the mean value theorem,

$$\frac{\rho_{\mathrm{M}}(f + f\varepsilon) - \rho_{\mathrm{M}}(f)}{\varepsilon} \leqslant \frac{f\varepsilon}{\varepsilon} \sup_{f \leqslant x \leqslant f + f\varepsilon} \left|\rho_{\mathrm{M}}'(x)\right| = f \sup_{f \leqslant x \leqslant f + f\varepsilon} \left|\rho_{\mathrm{M}}'(x)\right| \tag{6.1.13}$$

and

$$\frac{\rho_{\mathrm{M}}(f) - \rho_{\mathrm{M}}(f - f\varepsilon)}{\varepsilon} \geqslant -f \sup_{f - f\varepsilon_N \leqslant x \leqslant f} \left|\rho_{\mathrm{M}}'(x)\right|. \tag{6.1.14}$$

Thus, it is sufficient to show for all $f > 0$

$$\sup_{N \geqslant 2} \left|\rho_{\mathrm{M}}'(f)\right| \leqslant f^{-1}. \tag{6.1.15}$$

We note that this is not an optimal bound, however, it suffices for our applications. Calculating, one finds

$$\rho_{\mathrm{M}}'(f) = \frac{f^{N-2}\left(f^N - Nf + N - 1\right)}{\left(f^N - 1\right)^2}. \tag{6.1.16}$$

First, when $f \geqslant 1$ we prove the stronger claim

$$\frac{f^{N-2}\left(f^N - Nf + N - 1\right)}{\left(f^N - 1\right)^2} \leqslant f^{-2}, \tag{6.1.17}$$

by noting the above is equivalent to

$$(f - 1)\left(Nf^N - \sum_{k=0}^{N-1} f^k\right) \geqslant 0, \tag{6.1.18}$$

which is true since $f \geqslant 1$. Similarly, one can prove

$$\frac{f^{N-2}\left(f^N - Nf + N - 1\right)}{\left(f^N - 1\right)^2} < 1 \tag{6.1.19}$$

when $f < 1$. Equations (6.1.17) and (6.1.19) imply $\sup_{N \geqslant 2} |\rho_{\mathrm{M}}'(f)| \leqslant f^{-1}$.

Therefore, we may conclude

$$\sup_{f > 0} |\rho_G(f) - \rho_{\mathrm{M}}(f)| \leqslant \varepsilon, \tag{6.1.20}$$

which completes the proof. ■

Remark 6.4. Theorem 6.3 is actually slightly stronger than stated, and thus we can draw a slightly stronger conclusion in Theorem 6.10. We may conclude that the fixation probability of a mutant of fitness f originating at a particular vertex [249] satisfies the bound in (6.1.2), for exactly the same reason as in the proof of the original isothermal theorem—the bound (6.1.1) is for all subsets S. Therefore, *a fortiori*, a mutant can be started with any probability vector on the vertices (not merely uniform) and its fixation probability will still satisfy (6.1.2). This observation is borne through simulations too (Figure 6.5).

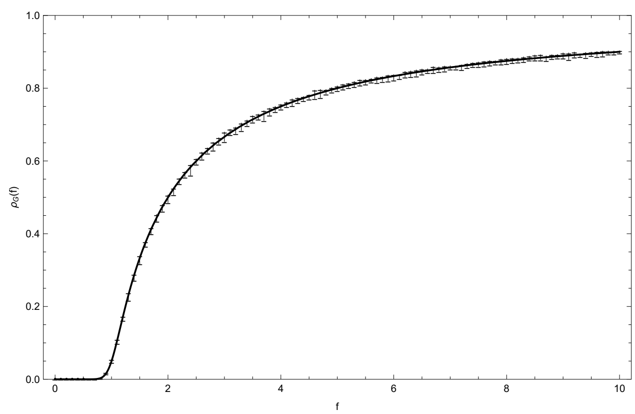

Figure 6.5: Fixation probability does not depend on starting location. We conducted trials where the fixation probability of a mutant of fitness f starting at vertex i was estimated with the Monte Carlo method of 10^4 samples for several values of $0 \leqslant f \leqslant 10$ on a Erdős-Rényi random graph. The fixation probability was similar regardless of starting vertex, and in particular, showed no correlation with vertex temperature. We plot the Moran fixation probability ρ_{M} and use the error bars to illustrate the minimum and maximum empirical fixation probabilities obtained from starting at any particular vertex.

6.2. Spectral isothermal theorem

Note that the assumptions of Theorem 6.3 require checking condition (6.1.1) for an exponential number of subsets. So trying to apply the theorem to particular graphs of moderate size is computationally challenging. However, there is a way to prove a version of the robust isothermal theorem whose assumptions can be checked in polynomial time in the size of the graph N.

Let $\|\cdot\|_2$ denote the Euclidean norm. Let u denote the uniform distribution and $\mathbf{1}_S$ be the indicator vector for $S \subseteq [\![N]\!]$, that is,

$$\mathbf{1}_S(i) := \mathbf{1}(i \in S). \tag{6.2.1}$$

Let G be a weighted, directed graph with N vertices and stochastic weight matrix W. For $j \in [\![N]\!]$, let $T(j) := W([\![N]\!], j)$ denote the temperature of the vertex j, that is, $\sum_i W(i,j)$. Define the constants

$$\lambda := \sup_{v \perp u} \frac{\|Wv\|_2}{\|v\|_2} \quad \text{and} \quad \varepsilon := \max_{j \in [\![N]\!]} |1 - W([\![N]\!], j)| \,. \tag{6.2.2}$$

We have $u^t W = T/N = (T(1), \ldots, T(N))/N$ and, since W is stochastic, we have $Wu = u$. Moreover, the Perron-Frobenius theorem implies $\lambda \leqslant 1$.

Theorem 6.5 (Spectral isothermal theorem). *Let $G = (V, W)$ be a connected graph. If there exist positive constants c_0 and c_1 such that $1 - \lambda - \varepsilon \geqslant c_0$ and $\varepsilon \leqslant c_0(1 - c_1)/2$, then there exists a positive constant C such that*

$$\sup_{f > 0} |\rho_{\mathrm{M}}(f) - \rho_G(f)| \leqslant C\varepsilon \,, \tag{6.2.3}$$

where $\rho_G(f)$ is the fixation probability of G and $\rho_{\mathrm{M}}(f)$ is the Moran fixation probability.

To prove Theorem 6.5, we want to apply the robust isothermal theorem. To do this we need to bound the quantity

$$\max_{S \subseteq [\![N]\!]} \left| \frac{W(S, S^c)}{W(S^c, S)} - 1 \right| \,, \tag{6.2.4}$$

where $W(S, S^c) := \sum_{i \in S, j \notin S} W(i,j)$. To do this we first prove several lemmata.

Lemma 6.6. *For all $S \subseteq [\![N]\!]$ such that $|S| = k$, we have*

$$\left| W(S^c, S) - \frac{k(N-k)}{N} \right| \leqslant \varepsilon \frac{k^{3/2}(N-k)^{1/2}}{N} + \lambda \frac{k(N-k)}{N} \,. \tag{6.2.5}$$

Proof. Fix $S \subseteq [\![N]\!]$, where $|S| = k$. Note that $W(S^c, S) = (\mathbf{1}_{S^c})^t \, W \mathbf{1}_S$. Now we decompose $\mathbf{1}_S$ into two components: $\mathbf{1}_S = au + \mathbf{1}_S^\perp$, where $\mathbf{1}_S^\perp$ is orthogonal to u. The coefficient a is $\langle \mathbf{1}_S \,, u \rangle / \langle u \,, u \rangle = k$. We do the same with $\mathbf{1}_{S^c}$. Thus,

$$\begin{aligned}
(\mathbf{1}_{S^c})^t \, W \mathbf{1}_S &= k(N-k) u^t W u + k u^t W \mathbf{1}_{S^c}^\perp + (N-k) \left(\mathbf{1}_S^\perp \right)^t W u + \left(\mathbf{1}_S^\perp \right)^t W \mathbf{1}_{S^c}^\perp \\
&= \frac{k(N-k)}{N} + \frac{k}{N} T \mathbf{1}_{S^c}^\perp + \left(\mathbf{1}_S^\perp \right)^t W \mathbf{1}_{S^c}^\perp .
\end{aligned} \tag{6.2.6}$$

Next, we have the following bounds. First, we see

$$k = \|\mathbf{1}_S\|_2^2 = \|ku\|_2^2 + \|\mathbf{1}_S^\perp\|_2^2 = k^2/N + \|\mathbf{1}_S^\perp\|_2^2 \tag{6.2.7}$$

and so $\|\mathbf{1}_S^\perp\|_2 = \sqrt{k(N-k)/N}$. Second, we see

$$\begin{aligned}
\left|\frac{k}{N} T \mathbf{1}_{S^c}^\perp\right| &\leqslant \frac{k}{N}\left|\left(Nu^t + \left(T - Nu^t\right)\right)\mathbf{1}_{S^c}^\perp\right| \\
&= \frac{k}{N}\left|\left(T - Nu^t\right)\mathbf{1}_{S_c}^\perp\right| \\
&\leqslant \frac{k}{N}\|T - Nu^t\|_2\|\mathbf{1}_{S^c}^\perp\|_2 \\
&\leqslant \frac{k}{N}\sqrt{\sum_i (t_i - 1)^2}\sqrt{k(N-k)/N} \\
&= \varepsilon\frac{k^{3/2}(N-k)^{1/2}}{N}\,.
\end{aligned} \tag{6.2.8}$$

Third, we see

$$\left|\left(\mathbf{1}_S^\perp\right)^t W\mathbf{1}_{S^c}^\perp\right| \leqslant \|\mathbf{1}_S^\perp\|_2\left\|W\mathbf{1}_{S^c}^\perp\right\|_2 \leqslant \lambda\|\mathbf{1}_S^\perp\|_2\|\mathbf{1}_{S^c}^\perp\|_2 \leqslant \lambda\frac{k(N-k)}{N}. \tag{6.2.9}$$

This completes the proof. ■

Corollary 6.7. *Assume that there is a positive constant c_0 such that $1 - \lambda - \varepsilon \geqslant c_0$. For all $S \subseteq [\![N]\!]$ such that $|S| = k \leqslant N/2$, we have*

$$W(S^c, S) \geqslant (1-\lambda)\,\frac{k(N-k)}{N} - \varepsilon\frac{k^{3/2}(N-k)^{1/2}}{N}\,. \tag{6.2.10}$$

Proof. By the reverse triangle inequality, we get

$$|W(S^c,S)| \geqslant \frac{k(N-k)}{N} - \left|W(S^c,S) - \frac{k(N-k)}{N}\right| \geqslant (1-\lambda)\,\frac{k(N-k)}{N} - \varepsilon\frac{k^{3/2}(N-k)^{1/2}}{N}\,, \tag{6.2.11}$$

by Lemma 6.6. Note the lower bound in (6.2.11) is positive by assumption. ■

Lemma 6.8. *For all $S \subseteq [\![N]\!]$ such that $|S| = k$, we have*

$$|W(S,S^c) - W(S^c,S)| \leqslant \varepsilon\min\{k, N-k\}\,. \tag{6.2.12}$$

PROOF. Fix $S \subseteq [\![N]\!]$, where $|S| = k$. Note

$$
\begin{aligned}
W(S, S^c) &= \sum_{i \in S} \sum_{j \notin S} W(i,j) = \sum_{i \in S} \left(1 - \sum_{j \in S} W(i,j)\right) \\
&= k - \sum_{j \in S} \left(W([\![N]\!], j) - \sum_{i \notin S} W(i,j)\right) \\
&= W(S^c, S) + \sum_{j \in S} (1 - W([\![N]\!], j))
\end{aligned} \tag{6.2.13}
$$

and

$$
\begin{aligned}
W(S^c, S) &= \sum_{i \notin S} \sum_{j \in S} W(i,j) \\
&= \sum_{i \notin S} \left(1 - \sum_{j \notin S} W(i,j)\right) \\
&= (N-k) - \sum_{j \notin S} \left(W([\![N]\!], j) - \sum_{i \in S} W(i,j)\right) \\
&= W(S, S^c) + \sum_{j \notin S} (1 - W([\![N]\!], j)) \ .
\end{aligned} \tag{6.2.14}
$$

Thus,

$$
|W(S,S^c) - W(S^c,S)| \leqslant \min\left\{ \left|\sum_{j \in S} (1 - W([\![N]\!], j))\right|, \left|\sum_{j \notin S} (1 - W([\![N]\!], j))\right| \right\} \leqslant \varepsilon \min\{k, N-k\}\,, \tag{6.2.15}
$$

by (6.2.2). ■

COROLLARY 6.9. *If there exist positive constants c_0 and c_1 such that $1 - \lambda - \varepsilon \geqslant c_0$ and $\varepsilon \leqslant c_0(1-c_1)/2$, then there exists a positive constant C such that*

$$
\max_{S \subseteq [\![N]\!]} \left| \frac{W(S,S^c)}{W(S^c,S)} - 1 \right| \leqslant C\varepsilon\,. \tag{6.2.16}
$$

PROOF. Note

$$
\left| \frac{W(S,S^c)}{W(S^c,S)} - 1 \right| = \frac{|W(S,S^c) - W(S^c,S)|}{W(S^c,S)}\,. \tag{6.2.17}
$$

First, suppose $k \leqslant N/2$. Thus, applying Corollary 6.7 and Lemma 6.8, we see

$$
\left| \frac{W(S,S^c)}{W(S^c,S)} - 1 \right| \leqslant \frac{\varepsilon N \min\{k, N-k\}}{(1-\lambda)\, k(N-k) - \varepsilon k^{3/2}(N-k)^{1/2}} \leqslant \frac{\varepsilon N}{(1-\lambda-\varepsilon)(N-k)} \leqslant 2c_0^{-1}\varepsilon \leqslant C\varepsilon\,. \tag{6.2.18}
$$

Now assume $k \geqslant N/2$. Applying Equation (6.2.18) to S^c, we see

$$\left|\frac{W(S^c, S)}{W(S, S^c)} - 1\right| = \left|\frac{W(S^c, (S^c)^c)}{W((S^c)^c, S^c)} - 1\right| \leqslant 2c_0^{-1}\varepsilon\,. \tag{6.2.19}$$

Let $x = W(S^c, S)/W(S, S^c)$. So,

$$\left|\frac{1}{x} - 1\right| \leqslant \frac{|1-x|}{|1-|1-x||} \leqslant \frac{2c_0^{-1}\varepsilon}{1-2c_0^{-1}\varepsilon} \leqslant C\varepsilon\,. \tag{6.2.20}$$

Thus, we let $C = \max\left\{2c_0^{-1}, 2/(c_0 - 2\varepsilon)\right\}$. ■

Proof of spectral isothermal theorem. The proof follows from Corollary 6.9 and the robust isothermal theorem. ■

The spectral isothermal theorem can be a useful bound on the fixation probability of random graphs, since their spectral gaps have been widely studied because of its relationship with mixing times [54, 67, 231]. Say we have a family of random graphs such that the spectral gap $1 - \lambda_2^{(N)} \to \gamma > 0$ and their vertex temperatures are such that $T(i) \sim T(j)$ for all i and j. The spectral isothermal theorem proves convergence in probability of the graph's fixation probability to ρ_{M}.

6.3. Random graphs

In this section we prove Theorem 6.10.

Theorem 6.10. *Let $(G_N)_{N\geqslant 1}$ be a family of random graphs as in Definition 6.13. Then there is a constant $C > 0$, not dependent on N, such that the fixation probability of a randomly placed mutant of fitness $f > 0$ satisfies*

$$|\rho_{G_N}(f) - \rho_{\mathrm{M}}(f)| \leqslant \frac{C\,(\log N)^{C+C\xi}}{\sqrt{N}} \tag{6.3.1}$$

uniformly in f with probability greater than

$$1 - \exp\left(-\nu\,(\log N)^{1+\xi}\right), \tag{6.3.2}$$

for positive constants ξ and ν.

To do this we need to apply Theorem 6.3 to our random graphs by showing that its assumptions hold with high probability. We do so in several steps. First, we define precisely generalized Erdős-Rényi random graphs in Definition 6.13 and outline the necessary assumptions on the distribution of the edge weights. After reviewing some notation, we introduce an event Ω, on which the graphs are well behaved, and show that Ω has high probability in Lemma 6.15. Then the general idea is to use large deviation estimates and concentration inequalities to show that with high probability the quantity (6.1.1) can be controlled. We bound both the numerator (Lemma 6.16) and denominator (Lemma 6.17) of

$$\frac{|W(S,S^c)-W(S^c,S)|}{|W(S,S^c)|}=\left|\frac{W(S,S^c)}{W(S^c,S)}-1\right| \tag{6.3.3}$$

for all S, then we put everything together to prove Theorem 6.10.

Remark 6.11 (Notation). We use the large constant $C>0$ and the small constant $c>0$, which do not depend on the size of the graph N but can depend on the distribution as outlined in Definition 6.13. We allow the constant C to increase or the constants c, ν, and ξ to decrease from line to line without noting it or introducing a new notation, sometimes we even absorb other constants such as p, p', and μ_1 without noting it; as is clear from the proof, this only happens a finite number of times, and thus we end with constants $C>0$, $c>0$, $\nu>0$, and $\xi>0$.

We also make use of standard order notation for functions, $o(\cdot)$, $\mathcal{O}(\cdot)$, and $\cdot \gg \cdot$, all of which are used with respect to N. Moreover, in some sums it is useful to exclude particular summands, e.g.

$$\sum_{\substack{1\leqslant j\leqslant N\\ j\notin S}}\cdot \equiv \sum_{j}^{(S)}\cdot \tag{6.3.4}$$

for $S\subset [\![N]\!]$. We abbreviate $(\{i\})$ and $(\{i,k\})$ as (i) and (i,k).

Remark 6.12 (High probability events). We say that an N-dependent event E holds with high probability if, for constants $\xi>0$ and $\nu>0$ which do not depend on N,

$$\mathbb{P}[E^c]\leqslant e^{-\nu(\log N)^{1+\xi}} \tag{6.3.5}$$

for $N \geqslant N_0(\nu, \xi)$. Moreover, we say an event E has high probability on another event E_0 if

$$\mathbb{P}[E_0 \cap E^c] \leqslant e^{-\nu(\log N)^{1+\xi}} \tag{6.3.6}$$

In particular, this has the property that the intersection of polynomially many (in N, say KN^K for some constant $K > 0$) events of high probability is also an event of high probability: by the union bound,

$$\mathbb{P}\left[\left(\bigcap_{i=1}^{KN^K} E_i\right)^c\right] = \mathbb{P}\left[\bigcup_{i=1}^{KN^K} E_i^c\right] \leqslant KN^K \cdot e^{-\nu(\log N)^{1+\xi}} = Ke^{K\log N - \nu(\log N)^{1+\xi}} \leqslant e^{-\nu(\log N)^{1+\xi}}, \tag{6.3.7}$$

with a possible increase in $N_0(\nu, \xi)$ and a decrease in the constants ν and ξ.

6.3.1. Proof of Theorem 6.10. Following the Erdős-Rényi model, we produce a weighted, directed graph as follows: Consider an $N \times N$ matrix X with zero for its diagonal entries and independent, identically distributed, nonnegative random variables for its off-diagonal entries. We now want to define a random, stochastic matrix W of weights. The natural definition for W is

$$W(i,j) := \frac{W(i,j)}{\sum_{k=1}^N W(i,k)}, \tag{6.3.8}$$

which is defined when at least one of the $W(i,1), \ldots, W(i,N)$ is nonzero; this happens almost surely in the limit as $N \to \infty$, when $\mathbb{P}\{X(i,j) > 0\} = p > 0$ is a constant:

$$\mathbb{P}\left\{X(i,1) = X(i,2) = \cdots = X(i,N) = 0\right\} = (1-p)^{N-1} \tag{6.3.9}$$

and by the union bound

$$\begin{aligned} \mathbb{P}\left[\bigvee_{i=1}^{n} X(i,1) = X(i,2) = \cdots = X(i,N) = 0\right] &\leqslant \sum_{i=1}^{n} \mathbb{P}\left\{X(i,1) = X(i,2) = \cdots = X(i,N) = 0\right\} \\ &= N(1-p)^{N-1} \to 0. \end{aligned} \tag{6.3.10}$$

However, the question is how to technically deal with the unusual event that all the entries of a row of X are zero, as there are several options. We make the following choice: for $1 \leqslant i \leqslant N$

$$W(i,j) := \begin{cases} 1 & \text{if } X(i,1) = X(i,2) = \cdots = X(i,N) = 0 \\ \\ 0 & \text{otherwise} \end{cases}, \tag{6.3.11}$$

and for all $1 \leqslant i, j \leqslant N$ and $i \neq j$

$$W(i,j) := \begin{cases} \frac{X(i,j)}{\sum_{k=1}^{N} X(i,k)} & \text{if } X(i,j) > 0 \\ \\ 0 & \text{if } X(i,j) = 0 \end{cases}. \tag{6.3.12}$$

This definition aligns with the definition in (6.3.8) with probability greater than $1 - N(1-p)^N$. Moreover, this definition has the advantage that the events that any non-loop edge weight is 0 are independent.

Definition 6.13 (Generalized Erdős-Rényi random graphs). *Let μ be a nonnegative distribution (not depending on N) with subexponential decay such that if $X \sim \mu$*

$$\mathbb{P}\{X > 0\} = p > 0 \text{ and } \mathbb{P}\{X \geqslant x\} \leqslant Ce^{-x^{1/C}} \tag{6.3.13}$$

for some positive constants p and C and all $x > 0$. We denote the mean and standard deviation of X by μ_1 and σ respectively. We generate a family of random graphs $G_N = ([\![N]\!], W_N)$ from μ by defining the weight matrices W_N according to (6.3.11) *and* (6.3.12)*, where $X(i,j)$ are independent and distributed according to μ for $i \neq j$.*

The subexponential decay is necessary to control the fluctuation of the graph's edge weights and imposes a bounded increase on the moments of μ. Let $X \sim \mu$, then simple calculations show,

$$\mu_k := \mathbb{E}X^k \leqslant Ck \int_0^\infty x^{k-1} \mathbb{P}\{X \geqslant x\} \, \mathrm{d}x \leqslant Ck \int_0^\infty x^{k-1} e^{-x^{1/C}} \, \mathrm{d}x = C^2 k \Gamma\left(C\left(1+k\right)\right) \leqslant (Ck)^{Ck}, \tag{6.3.14}$$

where the constant $C > 0$ depends on the constants in (6.3.13). Many distributions satisfy the subexponential assumption (6.3.13), for example any compactly supported distribution, the Gamma distribution, and the absolute value of a Gaussian distribution.

We now use the subexponential decay assumption to understand the typical behavior of the random variables $X(i,j)$.

Definition 6.14 (Good events Ω). *Let Ω be an n-dependent event such that the following hold:*

$$\Omega := \bigcap_{i=1}^{N} \left(\{X(i,i) = 0\} \cap \left\{ \left| \sum_{j}^{(i)} (X(i,j) - \mathbb{E}X(i,j)) \right| \leqslant \sigma \left(\log N\right)^{C+C\xi} \sqrt{N} \right\} \right) \tag{6.3.15}$$

$$\cap \bigcap_{i,j=1}^{N} \left\{ X(i,j) \leqslant C \left(\log N\right)^{C} \right\} \cap \{G_N \textit{ is connected}\}. \tag{6.3.16}$$

The conditions on Ω have natural interpretations. The first condition specifies that the normalization procedure outlined above has worked as intended and that we are not in the atypical case where the graph has a self-loop. The second condition specifies that the sums of N of the $X(i,j)$s are close to their expectation $(N-1)\mu_1$ and that they fluctuate about this value on the order of $\sqrt{N}$ as predicted by the central limit theorem. The third condition says that none of the $X(i,j)$ are too large and that typically they will all be less than $C\left(\log N\right)^{C}$. The last condition is self-explanatory, as the Moran process is not guaranteed to terminate on disconnected graphs.

Lemma 6.15. *The event Ω holds with high probability.*

Proof. By Remark 6.12, it suffices to show that each conjunct holds with high probability as there are only polynomially many choices for i and j. First fix i. By assumption (6.3.13) and the fact that $X(i,i) \neq 0$ only if $X(i,j) = 0$ for all $j \neq i$,

$$\mathbb{P}\left\{X(i,i) \neq 0\right\} = \mathbb{P}\left[X(i,j) = 0 \text{ for all } j \neq i\right] \leqslant (1-p)^{N-1} = e^{\log(1-p)(N-1)} \leqslant e^{-\nu(\log N)^{1+\xi}}, \tag{6.3.17}$$

since $0 < p < 1$ and $N - 1 \gg (\log N)^{1+\xi}$.

Now using the large deviation result, Lemma B.4, with $a_j = X(i,j) - \mathbb{E}X(i,j)$ and $A_j = 1$, we may verify the moment assumption (B.0.7): clearly $\mathbb{E}\left(X(i,j) - \mathbb{E}X(i,j)\right) = 0$ and $\mathbb{E}\left(X(i,j) - \mathbb{E}X(i,j)\right)^2 = \sigma^2$, then

$$\mathbb{E}\left|X(i,j) - \mathbb{E}X(i,j)\right|^k \leqslant (Ck)^{Ck} \tag{6.3.18}$$

by Equation (6.3.14). Thus we get

$$\mathbb{P}\left[\left|\sum_{i=1}^{N} a_i A_i\right| \geqslant \sigma \left(\log n\right)^{C+C\xi} \sqrt{N}\right] \leqslant e^{-\nu(\log N)^{1+\xi}}. \tag{6.3.19}$$

Now fix j too. Next, we use the subexponential decay assumption (6.3.13) with $x = C\left(\log N\right)^{C+1}$ to get

$$\mathbb{P}\left[X(i,j) > C\left(\log N\right)^{c^{-1}+1}\right] \leqslant C\exp\left(-\left(C\left(\log N\right)^{C+1}\right)^{1/C}\right) \leqslant Ce^{-C^{C^{-1}}(\log N)^{1+C^{-1}}} \leqslant e^{-\nu(\log N)^{1+\xi}}. \qquad (6.3.20)$$

Thus, $X(i,j) > C\left(\log N\right)^{C+1}$ holds with high probability since $C^{-1} > 0$ and $C^{C^{-1}} > 0$.

Finally, we show that the graph G is connected with high probability, i.e. that with high probability, the graph cannot be partitioned into two disjoint sets where there are no edges going from one subset to another. This follows from an argument similar to that contained in the proof of Lemma 6.17 but without the assumption that we are on the event Ω as we do not need a lower bound on the weights only that edges exist which they do with probability at least p. ■

Note that by definition W is stochastic. Define the sum of the jth column as

$$W(\llbracket N \rrbracket, j) := \sum_{i=1}^{N} W(i,j). \qquad (6.3.21)$$

Note that while the family $W(\llbracket N \rrbracket, j)$ is not independent, by symmetry, they are identically distributed. Hence

$$\mathbb{E}W(\llbracket N \rrbracket, 1) = \frac{1}{N}\sum_{j=1}^{N}\mathbb{E}W(\llbracket N \rrbracket, j) = \frac{1}{N}\mathbb{E}\sum_{j=1}^{N}\sum_{i=1}^{N} W(i,j) = 1. \qquad (6.3.22)$$

This tells us that in expectation W is doubly stochastic. The next lemma shows that with very high probability it is almost $N^{-1/2}$ close to being doubly stochastic, which is exactly the order of fluctuation we expect by the central limit theorem. The assumptions on the distribution μ and the event Ω guarantee that we can prove that the sum's fluctuations are of this order.

The idea of the proof is that for fixed j, the $W(i,j)$ are independent random variables and thus we can apply a LDE to bound the fluctuations of their sum. There are complications due to the normalization required by Definition 6.13 but on Ω these can be overcome by relating the sum $W(\llbracket N \rrbracket, j)$ to a simpler sum that may be controlled with Lemma B.4.

Lemma 6.16. *On Ω, there is a positive constant $C \equiv C_\mu$, not dependent on N, such that the following inequalities hold*

$$\left|W(\llbracket N \rrbracket, j) - 1\right| \leqslant \frac{C\left(\log N\right)^{C+C\xi}}{\sqrt{N}} \qquad (6.3.23)$$

for all $j \in [\![N]\!]$, with probability at least

$$1 - e^{-\nu(\log N)^{1+\xi}}. \tag{6.3.24}$$

Proof. Fix j. First we use the fact that $W(i,i) = X(i,i) = 0$ for $1 \leqslant i \leqslant N$ on Ω to see

$$W([\![N]\!], j) - 1 = \sum_i \left(W(i,j) - \mathbb{E}W(i,j)\right) = \sum_i^{(j)} \left(W(i,j) - \frac{1}{N-1}\right) + \mathcal{O}\left(N^2(1-p)^{-N+1}\right). \tag{6.3.25}$$

By the definition of $W(i,j)$, the above is equal to

$$\sum_i^{(j)} \left(\frac{X(i,j)}{\sum_k^{(i)} X(i,k)} - \frac{1}{N-1}\right) + \mathcal{O}\left(c_0^{-N}\right) = \sum_i^{(j)} \left(\frac{X(i,j) - \frac{1}{N-1}\sum_k^{(i)} X(i,k)}{\sum_k^{(i)} X(i,k)}\right) + \mathcal{O}\left(c_0^{-N}\right), \tag{6.3.26}$$

where $c_0 < 1$ is not dependent on N. Next, using the fact that on Ω

$$\frac{1}{N-1}\left|\sum_k^{(i)} \left(X(i,k) - \mathbb{E}X(i,k)\right)\right| \leqslant C\sigma \left(\log N\right)^{C+C\xi} \frac{1}{\sqrt{N}} \tag{6.3.27}$$

for all $1 \leqslant i \leqslant n$, we replace the average in the numerator of (6.3.26) with its expectation to find it equal to

$$\sum_i^{(j)} \left(\frac{X(i,j) - \mathbb{E}X(i,j)}{\sum_k^{(i)} X(i,k)} + \frac{C\sigma \left(\log N\right)^{C+C\xi}}{\sqrt{N}\sum_k^{(i)} X(i,k)}\right) + \mathcal{O}\left(c_0^{-N}\right). \tag{6.3.28}$$

Using (6.3.27) again, it is easy to see

$$\sum_k^{(i)} X(i,k) \geqslant (N-1)\mathbb{E}X(i,j) - C\sigma \left(\log N\right)^{C+C\xi} \sqrt{N}, \tag{6.3.29}$$

which gives an upper bound on the error term in (6.3.28) and we find the equation equal to

$$\sum_i^{(j)} \frac{X(i,j) - \mathbb{E}X(i,j)}{\sum_k^{(i)} X(i,k)} + \mathcal{O}\left(\frac{\left(\log N\right)^{2C+2C\xi}}{\sqrt{N}}\right) + \mathcal{O}\left(c_0^{-N}\right). \tag{6.3.30}$$

Next we compare these two expressions to find that the absolute value their difference can be expressed as

$$\left|\frac{X(i,j) - \mathbb{E}X(i,j)}{\sum_k^{(i)} X(i,k)} - \frac{X(i,j) - \mathbb{E}X(i,j)}{\sum_k^{(i,j)} X(i,k)}\right| = \frac{\left|X(i,j)\right|^2}{\left|\sum_k^{(i)} X(i,k) \cdot \left(\sum_k^{(i)} X(i,k) - X(i,j)\right)\right|}. \tag{6.3.31}$$

However, using that on Ω, for all $1 \leqslant i, j \leqslant N$, we have $X(i,j) \leqslant C\left(\log N\right)^C$ and using (6.3.27) as before, we may show the difference is bounded by

$$\mathcal{O}\left(\frac{(\log N)^{4C+2C\xi}}{N^2}\right). \tag{6.3.32}$$

We can then sum over these errors—one for each summand—to get a total error of $\mathcal{O}\left((\log N)^{4C+2C\xi}/N\right)$. Thus, (6.3.30) may be rewritten as

$$\sum_i^{(j)} \frac{X(i,j) - \mathbb{E}X(i,j)}{\sum_k^{(i,j)} X(i,k)} + \mathcal{O}\left(\frac{(\log N)^{2C+2C\xi}}{\sqrt{N}}\right), \tag{6.3.33}$$

since the other error terms are dominated by the remaining one.

Note that $X(i,j)$ does not appear in the summand's denominator and thus the denominator and numerator are independent. So we can use the large deviation estimate, Lemma B.4, with $a_i = X(i,j) - \mathbb{E}X(i,j)$ and $A_i^{-1} = \sum_k^{(i,j)} X(i,k)$. While the A_i are random, we may condition on their values and treat them as deterministic constants, then after we have used the LDE, we can bound them using the fact that we are on Ω. That is, on Ω

$$\left(\sum_k^{(i,j)} X(i,k)\right)^2 = (N-2)^2 \left(\mathbb{E}X(i,j)\right)^2 + \mathcal{O}\left((\log N)^{2C+2C\xi} N\sqrt{N}\right) \tag{6.3.34}$$

and so

$$\sum_k^{(j)} A_i^2 = \frac{1}{(N-2)\left(\mathbb{E}X(i,j)\right)^2} + \mathcal{O}\left((\log N)^{2C+2C\xi} \frac{1}{N\sqrt{N}}\right). \tag{6.3.35}$$

Thus the LDE gives us

$$\mathbb{P}\left[\left|\sum_i^{(j)} a_i A_i\right| \geqslant \frac{C(\log N)^{C+C\xi}}{\sqrt{N}}\right] \leqslant e^{-\nu(\log N)^{1+\xi}}, \tag{6.3.36}$$

which combined with (6.3.33)

$$\mathbb{P}\left[|W([\![N]\!], j) - 1| \geqslant \frac{C(\log N)^{2C+2C\xi}}{\sqrt{N}}\right] \leqslant e^{-\nu(\log N)^{1+\xi}}. \tag{6.3.37}$$

The properties of high probability and the fact that we have N choices for j completes the proof. ■

Next we prove a lower bound on sums of edge weights, $W(S, S^c)$ and $W(S^c, S)$ for all $\varnothing \neq S \subsetneq [\![N]\!]$. The proof relies on concentration inequalities for independent random variables and the simple fact that on Ω there is a constant $c > 0$ such that $W(i,j) \geqslant cN^{-1}\mathbf{1}(X(i,j) \geqslant c)$ for all $i, j \in [\![N]\!]$.

Lemma 6.17. *On Ω, for all $\varnothing \neq S \subsetneq [\![N]\!]$ and some small constant $c \equiv c_\mu > 0$, not dependent on N, we have the* *following bound*

$$|W(S,S^c)| = |W(S^c,S)| \geqslant c_\mu \min\{|S|, N-|S|\} \tag{6.3.38}$$

with probability greater than

$$1 - e^{-\nu(\log N)^{1+\xi}}. \tag{6.3.39}$$

PROOF. First note that as in the proof of Lemma 6.16, we can argue that on Ω the sum $\sum_k^{(i)} X(i,k) \leqslant CN$, see (6.3.29) for all $1 \leqslant i \leqslant N$. Moreover, by assumption on the distribution μ, we have $\mathbb{P}[X(i,j) > 0] = p > 0$ and thus there is a constant $c > 0$ such that $\mathbb{P}[X(i,j) \geqslant c] = p' > 0$. Therefore, on Ω

$$W(i,j) \geqslant cN^{-1}\mathbf{1}(X(i,j) \geqslant c). \tag{6.3.40}$$

However, for each $1 \leqslant i,j \leqslant N$ with $i \neq j$, define $\beta_{ij} := \mathbf{1}(X(i,j) \geqslant c)$ which are independent Bernoulli random variables such that

$$\mathbb{P}[\mathbf{1}(X(i,j) \geqslant c) = 1] = p' > 0, \tag{6.3.41}$$

since the $X(i,j)$ are independent.

Let $|S| = k$. By definition

$$W(S,S^c) = \sum_{i \notin S}\sum_{j \in S} W(i,j) = W(S^c,S). \tag{6.3.42}$$

Note that no diagonal terms are in these sums. Using (6.3.41),

$$\sum_{i \in S}\sum_{i \notin S} W(i,j) \geqslant \frac{c}{N}\sum_{i \in S}\sum_{i \notin S} \beta_{ij}. \tag{6.3.43}$$

Note that now this is a sum of $k(N-k)$ independent random variables. So for fixed $\varnothing \neq S \subsetneq [\![N]\!]$, by the Chernoff bound, Lemma B.3, the event

$$A_S := \left\{\sum_{i \in S}\sum_{i \notin S} \beta_{ij} \leqslant (1-1/2)p'k(N-k)\right\} \tag{6.3.44}$$

has probability less than

$$\exp\left(-\frac{1}{8}p'k(N-k)\right). \tag{6.3.45}$$

Thus, by the union bound

$$\begin{aligned}
\mathbb{P}\left[\bigcap_{\varnothing\neq S\subsetneq[\![N]\!]} A_S^c\right] &= 1-\mathbb{P}\left[\bigcup_{\varnothing\neq S\subsetneq[\![N]\!]} A_S\right] \\
&\geqslant 1-\sum_{\varnothing\neq S\subsetneq[\![N]\!]}\mathbb{P}\left[A_S\right] \\
&= 1-\sum_{k=1}^{N-1}\binom{N}{k}\mathbb{P}\left[A_S\right] \\
&= 1-\sum_{k=1}^{\lfloor\log N\rfloor}\binom{N}{k}\mathbb{P}\left[A_S\right]-\sum_{k=\lfloor\log N\rfloor+1}^{N-\lfloor\log N\rfloor-1}\binom{N}{k}\mathbb{P}\left[A_S\right]-\sum_{N-\lfloor\log N\rfloor}^{N-1}\binom{N}{k}\mathbb{P}\left[A_S\right] \\
&\geqslant 1-2N^{\log N}\exp\left(-p'N/8\right)-2^N\exp\left(-p'N\left(\log N\right)/8\right) \\
&\geqslant 1-2\exp\left(-\left(p'N/8-(\log N)^2\right)\right)-\exp\left(-\left(p'\log N/8-\log 2\right)N\right) \\
&\geqslant 1-\exp\left(-cp'N\right),
\end{aligned} \tag{6.3.46}$$

for some $c>0$. Finally, note

$$\frac{cp'}{2N}k(N-k)\geqslant\frac{cp'}{2}\min\left\{|S|,N-|S|\right\} \tag{6.3.47}$$

and

$$\exp(-cp'N)\leqslant e^{-\nu(\log N)^{1+\xi}} \tag{6.3.48}$$

for an appropriate choice of ξ and ν. ■

We now complete the proof of Theorem 6.10 by putting together the results of Section 6.1 and the lemmata from this section.

Proof of Theorem 6.10. Again let $|S|=k$. We check that the assumptions of Theorem 6.3 hold with high probability. Observe

$$\left|\frac{W(S^c,S)}{W(S,S^c)}-1\right|=\left|\frac{W(S^c,S)-W(S,S^c)}{W(S,S^c)}\right|=\frac{|W(S^c,S)-W(S,S^c)|}{|W(S,S^c)|}. \tag{6.3.49}$$

Expanding the numerator, we get

$$W(S^c,S)-W(S,S^c)=\sum_{i\in S}\sum_{j\notin S}W(i,j)-\sum_{i\notin S}\sum_{j\in S}W(i,j)=\sum_{i\in S}\sum_{j\in V}W(i,j)-\sum_{i\in V}\sum_{j\in S}W(i,j)=\sum_{j\in S}(1-W([\![N]\!],j)), \tag{6.3.50}$$

and similarly,

$$W(S^c, S) - W(S, S^c) = \sum_{i \in V} \sum_{j \notin S} W(i,j) - \sum_{i \notin S} \sum_{j \in V} W(i,j) = \sum_{j \notin S} (W(\llbracket N \rrbracket, j) - 1). \tag{6.3.51}$$

Thus Lemma 6.16 implies

$$|W(S^c, S) - W(S, S^c)| \leqslant \min\{k, N-k\} \cdot \frac{C(\log N)^{C+C\xi}}{\sqrt{N}}, \tag{6.3.52}$$

for all S with high probability on Ω.

Lemma 6.17 implies

$$|W(S, S^c)| \geqslant c \min\{|S|, N - |S|\} \tag{6.3.53}$$

for all S with high probability on Ω. Putting this together we see

$$\left| \frac{W(S^c, S)}{W(S, S^c)} - 1 \right| \leqslant \frac{C(\log N)^{C+C\xi}}{\sqrt{N}} \tag{6.3.54}$$

for all S with high probability on Ω. However, by Lemma 6.15, the event Ω holds with high probability itself and thus unconditionally (6.3.54) holds for all S with high probability.

Finally, applying Theorem 6.3, we get

$$\sup_{f>0} |\rho_{G_N}(f) - \rho_{\mathrm{M}}(f)| \leqslant \frac{C(\log N)^{C+C\xi}}{\sqrt{N}} \tag{6.3.55}$$

with high probability. ■

Remark 6.18 (On Theorem 6.10). The parameter p, the probability that an edge of some weight exists between two directed vertices, can be interpreted as a measure of the sparseness of the population structure. We can ask how few interactions on average can individuals in a population have with others and still yield populations with mean-field behavior? While p can be arbitrarily small, we have kept it constant and, in particular, not dependent on N. However, could p depend on N such that $p \to 0$ as $N \to \infty$ and still produce graphs which show mean-field behavior? An obvious lower bound on the rate of p's convergence to 0 is provided by the Erdős-Rényi model, which tells us that a graph is almost surely disconnected in the limit for $\sqrt{p} < (1-\varepsilon)(1/N)\log N$ for any $\varepsilon > 0$. This bound follows by noting that $(1-p)^2$ is the probability that there is no edge, in either direction, between two vertices and then applying the usual Erdős-Rényi threshold [92, 278]. There is much room between this lower bound and p constant—even whether such a sharp threshold for p exists is currently unclear. The issue is difficult to approach with naive simulations as the Moran process is not guaranteed to terminate on disconnected graphs.

6.4. Optimal fluctuations

In section 6.3, we proved a bound on the deviation of the fixation probability from the Moran fixation probability for a large class of random graphs. In this sense, those random graphs display mean-field behavior. For fixed values of $f > 1$, a natural question to ask is whether the bound of order $\sqrt{1/N}$ in (6.3.1) is optimal. Suppose

$$\mathbb{P}\left\{\sup_{f>0} |\rho_G(f) - \rho_M(f)| \leqslant \delta(N)\right\} = 1 - \varepsilon(N), \tag{6.4.1}$$

for some functions $\delta(N) = o(1)$ and $\varepsilon(N) = o(1)$, and that any other $\tilde{\delta}$ satisfying (6.4.1) is such that $\delta(N) = \Theta\left(\tilde{\delta}(N)\right)$, then we call δ the optimal fluctuation size of ρ_G. To address this question of optimal fluctuations, we consider the variance of the fixation probability of random graphs from the model defined in Section 6.3. Since $\rho_G \in [0, 1]$ and

$$\mathbb{E}\left[\rho_G - \rho_{\mathrm{M}}\right]^2 = \mathbb{E}\left[\rho_G - \mathbb{E}\rho_G\right]^2 + \left(\mathbb{E}\rho_G - \rho_{\mathrm{M}}\right)^2 \geqslant \mathbb{E}\left[\rho_G - \mathbb{E}\rho_G\right]^2 = \mathrm{Var}(\rho_G) \tag{6.4.2}$$

with equality if $\rho_{\mathrm{M}} = \mathbb{E}\rho_G$, the standard deviation provides a lower bound on the size of optimal fluctuations. For the standard deviation σ of ρ_G to give the correct order for the optimal fluctuations, we require

$$|\mathbb{E}\rho_G - \rho_{\mathrm{M}}| = o\left(\sigma(N)\right). \tag{6.4.3}$$

This appears to be correct based on simulation results, but we do not verify it mathematically here.

At this point is important to make a distinction that we hinted at in Remark 6.4. For the robust isothermal theorem, the initial condition of the Moran process did not affect our result, but in general it does affect the probability of fixation (as we saw in Section 3.3). In particular, the size of the optimal fluctuation shows different behavior when we consider a single mutant arising uniformly at random on the vertices of the graph and the mutant arising at a specific vertex. In our random graph model, the vertices of the graph are exchangeable, and thus the vertex we specify is not important.

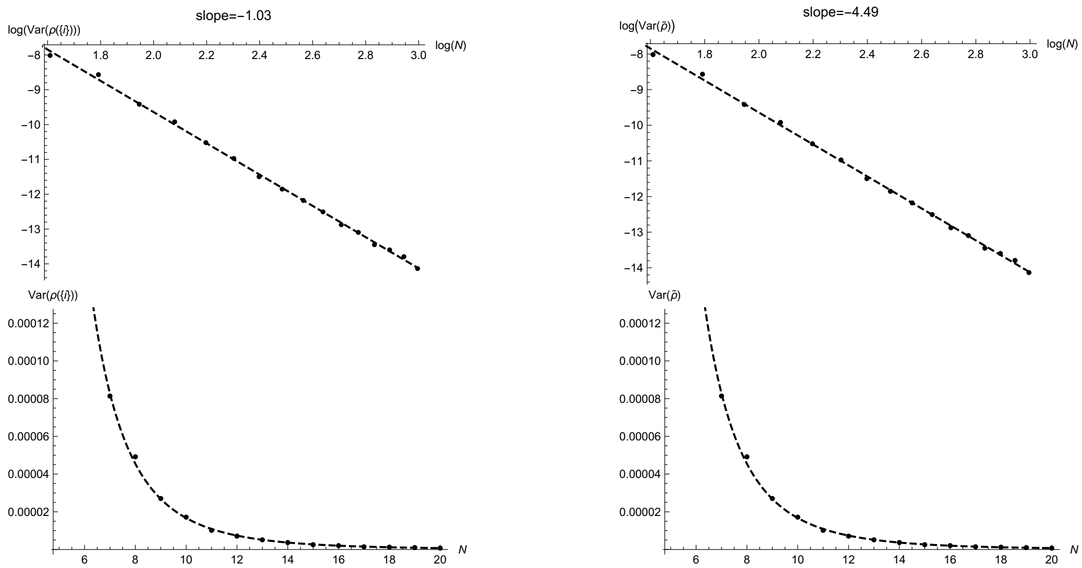

Figure 6.6: The fixation probabilities $\rho(\{i\})$ and $\tilde{\rho}$ were sampled 1,000 times each for $N \in \{5, 6, \ldots, 20\}$. The samples were obtained by randomly sampling the graph, then exactly calculating the fixation probability for that graph using matrix multiplication. Exactly calculating the fixation probability means that there is not an additional source of error introduced by Monte Carlo estimation of the fixation probability for a fixed graph. For each N, the sample was used to estimate the variance of the fixation probability. We then fit the estimated variance as a function of N. In particular, the log-log plots estimate c, where c is the exponent in the decay of the variance $\mathcal{O}\left(N^{-c}\right)$. We see that the simulation support Equation (6.4.36), but suggest that the variance is much smaller for uniform initialization than calculated in Equation (6.4.38). We set $f = 2$ throughout.

So let S_t be the Moran process on a graph G, where S_t denotes the subset of vertices occupied by mutants at time t and takes values in $2^{[\![N]\!]}$. Let $F := \{\exists t : S_t = V\}$ be the event of fixation. Define $\rho(S) := \mathbb{P}\left(F | X_0 = S\right)$ for the graph G. Let $O_i := \{X_0 = \{i\}\}$ and define

$$\tilde{\rho} := \frac{1}{N} \sum_{i=1}^{N} \rho\left(\{i\}\right) \tag{6.4.4}$$

and suppose $\mathbb{P}O_i = 1/N$, so that $\tilde{\rho} = \mathbb{P}F$.

The following lemma factorizes the fixation probability. The factorization has one term that is a rational function of the graphs entries and the fitness of the mutant, and a second term that is a complicated weighted average of fixation probabilities whose initial conditions contain more mutants.

Lemma 6.19. *Define*

$$Z_i := f(1 - W(i,i)) + (W([\![N]\!], i) - W(i,i)) \tag{6.4.5}$$

We have the following expressions for $\rho(\{i\})$ and $\tilde{\rho}$:

$$\rho(\{i\}) = \frac{f}{Z_i} \sum_j^{(i)} \rho(\{i,j\}) W(i,j), \tag{6.4.6}$$

and

$$\tilde{\rho} = \left(\frac{1}{N} \sum_{i=1}^{N} \frac{f(1 - W(i,i))}{Z_i} \right) \cdot \left(\sum_{i=1}^{N} \sum_{j=i+1}^{N} \frac{\frac{W(i,j)}{Z_i} + \frac{W(j,i)}{Z_j}}{\sum_k \frac{1-W(k,k)}{Z_k}} \rho\left(\{i,j\}\right) \right). \tag{6.4.7}$$

Proof. Denote the probability of increasing (decreasing) the number of mutants from state S by $p_{\pm}(S)$. Also, let $E := \{\max_t |S_t| > 1\}$ be the event that the mutants do not go extinct before increasing in number and define the (possibly infinite) stopping time $T := \inf\{t : |S_t| = 2\}$. Obviously, $F \subseteq E$.

Then, conditioning on E, we see

$$\rho(\{i\}) = \mathbb{P}\left(F | E \cap O_i\right) \mathbb{P}\left(E | O_i\right). \tag{6.4.8}$$

Note that if $E \cap O_i$ does occur, then there must be some $j \neq i$ such that the state $\{i,j\}$ is reached before any other state S such that $|S| = 2$. Define the event $E_{\{i,j\}} := \{X_T = \{i,j\}\}$. Note $E = \sqcup_i \sqcup_{j>i} E_{\{i,j\}}$ and $E \cap O_i = \sqcup_j^{(i)} O_i \cap E_{\{i,j\}}$. Thus, by the law of total probability, we have

$$\rho(\{i\}) = \mathbb{P}\left(E | O_i\right) \sum_j^{(i)} \mathbb{P}\left(F | O_i \cap E_{\{i,j\}}\right) \mathbb{P}\left(O_i \cap E_{\{i,j\}} | E \cap O_i\right). \tag{6.4.9}$$

Let $D := f|S| + N - |S|$, then note

$$\mathbb{P}\left(E | O_i\right) = \frac{p_+(\{i\})}{p_+(\{i\}) + p_-(\{i\})} = \frac{\frac{f}{D} \sum_j^{(i)} W(i,j)}{\frac{f}{D} \sum_j^{(i)} W(i,j) + \frac{1}{D} \sum_j^{(i)} W(j,i)} = \frac{f(1 - W(i,i))}{Z_i}, \tag{6.4.10}$$

$$\begin{aligned} \mathbb{P}\left(O_i \cap E_{\{i,j\}} | E \cap O_i\right) &= \mathbb{P}\left(E_{\{i,j\}} | E \cap O_i\right) \\ &= \frac{\mathbb{P}\left(E_{\{i,j\}} | O_i\right)}{\mathbb{P}\left(E | O_i\right)} \\ &= \frac{\frac{\frac{f}{D} W(i,j)}{\frac{f}{D} \sum_k^{(i)} W(i,k) + \frac{1}{D} \sum_k^{(i)} W(k,i)}}{\frac{\frac{f}{D} \sum_k^{(i)} W(i,k)}{\frac{f}{D} \sum_k^{(i)} W(i,k) + \frac{1}{D} \sum_k^{(i)} W(k,i)}} \\ &= \frac{W(i,j)}{1 - W(i,i)}, \end{aligned} \tag{6.4.11}$$

and

$$\mathbb{P}\left(F|O_i \cap E_{\{i,j\}}\right) = \mathbb{P}\left(F|E_{\{i,j\}}\right) = \rho(\{i,j\}), \tag{6.4.12}$$

by the strong Markov property. Therefore, Equation (6.4.9) becomes

$$\rho(\{i\}) = \frac{f}{Z_i}\sum_j^{(i)} \rho(\{i,j\})W(i,j), \tag{6.4.13}$$

and, summing over i, we get

$$\tilde{\rho} = \frac{1}{N}\sum_i \frac{f}{Z_i}\sum_j^{(i)} \rho(\{i,j\})W(i,j). \tag{6.4.14}$$

However, we can get a different expression for $\tilde{\rho}$:

$$\begin{aligned}
\mathbb{P}(F) &= \mathbb{P}(E)\,\mathbb{P}(F|E) \\
&= \left(\sum_{i=1}^{N} \mathbb{P}(E|O_i)\,\mathbb{P}(O_i)\right)\left(\sum_{i=1}^{N}\sum_{j=i+1}^{N} \mathbb{P}\left(F|E_{\{i,j\}}\right)\mathbb{P}\left(E_{\{i,j\}}|E\right)\right) \\
&= \left(\frac{1}{N}\sum_{i=1}^{N} \frac{f(1-W(i,i))}{Z_i}\right)\left(\sum_{i=1}^{N}\sum_{j=i+1}^{N} \mathbb{P}\left(E_{\{i,j\}}|E\right)\rho\left(\{i,j\}\right)\right)
\end{aligned} \tag{6.4.15}$$

Note that $E_{\{i,j\}} \subseteq O_i \cup O_j$, so

$$\begin{aligned}
\mathbb{P}\left(E_{\{i,j\}}|E\right) &= \mathbb{P}\left(E_{\{i,j\}}|E \cap O_i\right)\mathbb{P}(O_i|E) + \mathbb{P}\left(E_{\{i,j\}}|E \cap O_j\right)\mathbb{P}(O_j|E) \\
&= \mathbb{P}\left(E_{\{i,j\}}|E \cap O_i\right)\frac{\mathbb{P}(E|O_i)\,\mathbb{P}(O_i)}{\mathbb{P}(E)} + \mathbb{P}\left(E_{\{i,j\}}|E \cap O_j\right)\frac{\mathbb{P}(E|O_j)\,\mathbb{P}(O_j)}{\mathbb{P}(E)} \\
&= \mathbb{P}\left(E_{\{i,j\}}|E \cap O_i\right)\frac{\mathbb{P}(E|O_i)\,\mathbb{P}(O_i)}{\sum_k \mathbb{P}(E|O_k)\,\mathbb{P}(O_k)} + \mathbb{P}\left(E_{\{i,j\}}|E \cap O_j\right)\frac{\mathbb{P}(E|O_j)\,\mathbb{P}(O_j)}{\sum_k \mathbb{P}(E|O_k)\,\mathbb{P}(O_k)} \\
&= \frac{W(i,j)}{1-W(i,i)}\,\frac{\frac{f(1-W(i,i))}{Z_i}\frac{1}{N}}{\frac{1}{N}\sum_k \frac{f(1-W(k,k))}{Z_k}} + \frac{W(j,i)}{1-W(j,j)}\,\frac{\frac{f(1-W(j,j))}{Z_j}\frac{1}{N}}{\frac{1}{N}\sum_k \frac{f(1-W(k,k))}{Z_k}} \\
&= \frac{\frac{W(i,j)}{Z_i} + \frac{W(j,i)}{Z_j}}{\sum_k \frac{1-W(k,k)}{Z_k}}
\end{aligned} \tag{6.4.16}$$

Moreover, as a consistency check, we have

$$\sum_{i=1}^{N}\sum_{j=i+1}^{N} \frac{\frac{W(i,j)}{Z_i} + \frac{W(j,i)}{Z_j}}{\sum_k \frac{1-W(k,k)}{Z_k}} = \frac{1}{2}\sum_{i=1}^{N}\sum_j^{(i)} \frac{\frac{W(i,j)}{Z_i} + \frac{W(j,i)}{Z_j}}{\sum_k \frac{1-W(k,k)}{Z_k}} = 1. \tag{6.4.17}$$

■

However, the expressions we have found are unwieldy to analyze, so the following lemma simplifies these expressions.

Lemma 6.20. *We have the following expressions for ρ and $\tilde{\rho}$:*

$$\rho(\{i\}) = \left(\frac{f}{f+1} - \frac{f}{(f+1)^2}\varepsilon_i + \mathcal{O}\left(N^{-1}\right)\right)\sum_j^{(i)} \rho(\{i,j\})W(i,j), \tag{6.4.18}$$

and

$$\tilde{\rho} = \left(\frac{f}{1+f} + \frac{f}{(1+f)^3}\frac{1}{N}\sum_i \varepsilon_i^2 + \mathcal{O}\left(N^{-3/2}\right)\right) \tag{6.4.19}$$

$$\cdot\left(\sum_{i=1}^N \sum_{j=i+1}^N \left[\frac{W(i,j)+W(j,i)}{N} - \frac{W(i,j)\varepsilon_i + W(j,i)\varepsilon_j}{(f+1)N} + \mathcal{O}\left(N^{-3}\right)\right]\rho\left(\{i,j\}\right)\right) \tag{6.4.20}$$

with high probability.

Proof. Note we frequently use the facts that and $\mathbb{E}W(i,j) = 1/N$ and $W(i,j) = \mathcal{O}(1/N)$. Define $\varepsilon_i := W(\llbracket N \rrbracket, i) - 1 - (f+1)W(i,i)$. Note $\mathbb{E}\varepsilon_i = \Theta(1/N)$, $\mathrm{Var}\left(\varepsilon_i\right) = \Theta(1/N)$, and with high probability

$$\varepsilon_i = \mathcal{O}\left(1/\sqrt{N}\right). \tag{6.4.21}$$

By Taylor expansion,

$$\frac{f}{f(1-W(i,i)) + (W(\llbracket N \rrbracket, i) - W(i,i))} = \frac{f}{f+1} - \frac{f}{(f+1)^2}\varepsilon_i + \mathcal{O}\left(\varepsilon_i^2\right) \tag{6.4.22}$$

and so by Lemma 6.19, we see

$$\rho(\{i\}) = \left(\frac{f}{f+1} - \frac{f}{(f+1)^2}\varepsilon_i + \mathcal{O}\left(N^{-1}\right)\right)\sum_j^{(i)} \rho(\{i,j\})W(i,j). \tag{6.4.23}$$

Note

$$\frac{1}{N}\sum_{i=1}^N \varepsilon_i = -\frac{f+1}{N}\sum_{i=1}^N W(i,i) \tag{6.4.24}$$

Consider the first factor (6.4.7):

$$
\begin{aligned}
\frac{1}{N}\sum_{i=1}^{N}\frac{f(1-W(i,i))}{Z_i} &= \frac{1}{N}\sum_{i=1}^{N} f(1-W(i,i))\frac{1}{f+1+\varepsilon_i} \\
&= \frac{f}{1+f}\frac{1}{N}\sum_{i=1}^{N}(1-W(i,i))\left(1-\frac{\varepsilon_i}{1+f}+\frac{\varepsilon_i^2}{(1+f)^2}+O(\varepsilon_i)^3\right) \\
&= \frac{f}{1+f}+\frac{f}{(1+f)^3}\frac{1}{N}\sum_i \varepsilon_i^2+\frac{1}{N}\sum_i\left(O(\varepsilon_i)^3+O(W(i,i)\varepsilon_i)\right) \\
&= \frac{f}{1+f}+\frac{f}{(1+f)^3}\frac{1}{N}\sum_i \varepsilon_i^2+\mathcal{O}\left(N^{-3/2}\right). \qquad (6.4.25)
\end{aligned}
$$

Now consider the first factor (6.4.7):

$$
\begin{aligned}
&\frac{\frac{W(i,j)}{Z_i}+\frac{W(j,i)}{Z_j}}{\sum_k \frac{1-W(k,k)}{Z_k}} \\
&= \frac{\frac{W(i,j)+W(j,i)}{f+1}-\frac{W(i,j)\varepsilon_i+W(j,i)\varepsilon_j}{(f+1)^2}+\mathcal{O}\left(N^{-2}\right)}{\frac{N}{1+f}+\frac{1}{(1+f)^3}\sum_i\varepsilon_i^2+\mathcal{O}\left(N^{-1/2}\right)} \\
&= \frac{W(i,j)+W(j,i)-\frac{W(i,j)\varepsilon_i+W(j,i)\varepsilon_j}{f+1}+\mathcal{O}\left(N^{-2}\right)}{N+\frac{1}{(1+f)^2}\sum_i\varepsilon_i^2+\mathcal{O}\left(N^{-1/2}\right)} \\
&= \left(W(i,j)+W(j,i)-\frac{W(i,j)\varepsilon_i+W(j,i)\varepsilon_j}{f+1}+\mathcal{O}\left(N^{-2}\right)\right)\frac{1}{N}\left(1-\frac{1}{(1+f)^2N}\sum_i\varepsilon_i^2+\mathcal{O}\left(N^{-3/2}\right)\right) \\
&= \frac{W(i,j)+W(j,i)}{N}-\frac{W(i,j)\varepsilon_i+W(j,i)\varepsilon_j}{(f+1)N}+\mathcal{O}\left(N^{-3}\right). \qquad (6.4.26)
\end{aligned}
$$

Thus,

$$
\begin{aligned}
\tilde{\rho} = &\left(\frac{f}{1+f}+\frac{f}{(1+f)^3}\frac{1}{N}\sum_i\varepsilon_i^2+\mathcal{O}\left(N^{-3/2}\right)\right) \\
&\cdot\left(\sum_{i=1}^{N}\sum_{j=i+1}^{N}\left[\frac{W(i,j)+W(j,i)}{N}-\frac{W(i,j)\varepsilon_i+W(j,i)\varepsilon_j}{(f+1)N}+\mathcal{O}\left(N^{-3}\right)\right]\rho(\{i,j\})\right) \qquad (6.4.27)
\end{aligned}
$$

■

We now give a heuristic calculation to decrease the bound on the variances of $\rho(\{i\})$ and $\tilde{\rho}$. For two random variables X and Y, we have

$$
\mathrm{Var}(XY) = \mathrm{Cov}(X^2,Y^2) - \mathrm{Cov}(X,Y)^2 + \mathrm{Var}(X)\,\mathrm{Var}(Y) + \mathrm{Var}(X)\,(\mathbb{E}X)^2 + \mathrm{Var}(Y)\,(\mathbb{E}Y)^2 - 2\mathbb{E}X\mathbb{E}Y\,\mathrm{Cov}(X,Y). \qquad (6.4.28)
$$

By the Cauchy-Schwarz inequality

$$\operatorname{Cov}(X,Y) \leqslant \sqrt{\operatorname{Var}(X)\operatorname{Var}(Y)}. \tag{6.4.29}$$

Moreover, if we assume that X and Y are concentrated about their means, we see

$$\operatorname{Cov}(X^2,Y^2) \leqslant \sqrt{\operatorname{Var}(X)\operatorname{Var}(Y)}, \tag{6.4.30}$$

since the function $x \mapsto x^2$ is continuous. Thus,

$$\operatorname{Var}(XY) \leqslant \mathcal{O}\left(\max\left\{\operatorname{Var}(X), \operatorname{Var}(Y)\right\}\right) \tag{6.4.31}$$

for $\mathbb{E}X = \mathcal{O}(1)$, $\mathbb{E}Y = \mathcal{O}(1)$, $\operatorname{Var} X = o(1)$, $\operatorname{Var} Y = o(1)$, and both X and Y concentrated about their means.

Define F_1 and F_2 to be the two factors in the expression for $\rho(\{i\})$ in (6.4.18). Similarly, define $\tilde{F}_1$ and $\tilde{F}_2$ from (6.4.19). Define

$$\sigma^2(N) := \operatorname{Var}\left(\rho(\{i\})\right) \quad \text{and} \quad \tilde{\sigma}^2(N) := \operatorname{Var}\left(\tilde{\rho}\right), \tag{6.4.32}$$

where σ does not depend on i by the exchangeability of the random graph. Note also that by the robust isothermal theorem, $\mathbb{E}\rho(\{i\}) = \Theta(1)$ and $\mathbb{E}\tilde{\rho} = \Theta(1)$.

From now on we assume that the random variables $\rho(\{i\})$ and $\tilde{\rho}$ are concentrated about their means with high probability. Specifically,

$$|\rho(\{i\}) - \mathbb{E}\rho(\{i\})| \leqslant C\sigma(N) \tag{6.4.33}$$

and

$$|\tilde{\rho} - \mathbb{E}\tilde{\rho}| \leqslant C\tilde{\sigma}(N). \tag{6.4.34}$$

for some constant C not dependent on N with high probability.

Note that $\mathbb{E}F_1 = \Theta(1)$ and

$$\operatorname{Var}(F_1) = \Theta\left(\operatorname{Var}(\varepsilon_i)\right) + \mathcal{O}(N^{-1}) = \Theta(N^{-1}), \tag{6.4.35}$$

by Lemma 6.16. In fact, Lemma 6.16 implies F_1 is concentrated about its mean with fluctuations of order $\mathcal{O}(N^{-1/2})$. For the second factor, we see $\mathbb{E}F_2 = \Theta(1)$ by the robust isothermal theorem. For its variance, we assume $\operatorname{Var}(F_2) < \sigma^2(N)$ since it is a weighted average of fixation probabilities. We also assume it is concentrated about its mean.

Applying Equation (6.4.31) with $X = F_1$ and $Y = F_2$ (so that $XY = \rho(\{i\})$), we see

$$\sigma^2(N) = \mathcal{O}\left(\frac{1}{N}\right), \tag{6.4.36}$$

which agrees with the bound in the robust isothermal theorem.

Now, we see that $\mathbb{E}\tilde{F}_1 = \Theta(1)$ and

$$\mathrm{Var}(\tilde{F}_1) = \mathcal{O}\left(\mathrm{Var}\left(\frac{1}{N}\sum_i \varepsilon_i^2\right)\right) + \mathcal{O}\left(\frac{1}{N^3}\right) = \mathcal{O}\left(\frac{1}{N^3}\right), \tag{6.4.37}$$

since ε_i is asymptotically Gaussian $\mathcal{N}\left(\mathcal{O}\left(\frac{1}{N}\right), \mathcal{O}\left(\frac{1}{N}\right)\right)$. We should again expect $\tilde{F}_1$ to be concentrated about its mean. For the second factor, we see $\mathbb{E}\tilde{F}_2 = \Theta(1)$ by the robust isothermal theorem. For its variance, we assume $\mathrm{Var}(\tilde{F}_2) < \tilde{\sigma}^2(N)$ since it is a weighted average of fixation probabilities. We also assume it is concentrated about its mean. Applying Equation (6.4.31) with $X = \tilde{F}_1$ and $Y = \tilde{F}_2$ (so that $XY = \tilde{\rho}$), we see

$$\tilde{\sigma}^2(N) = \mathcal{O}\left(\frac{1}{N^3}\right), \tag{6.4.38}$$

which is much less than the bound from the isothermal theorem. Moreover, if we write

$$\tilde{\rho} = \frac{1}{N}\sum_i \rho\left(\{i\}\right) \tag{6.4.39}$$

and imagine that the $\rho\left(\{i\}\right)$ are independent, we would expect the variance of $\tilde{\rho}$ to be $\mathcal{O}\left(1/N^2\right)$. Our calculation above suggests it is much smaller. This is most likely due to the cancellation than happens in Equation (6.4.25).

A

Birth-death chains

A birth-death chain is a simple type of Markov chain. Let the state space be $[\![0,n]\!]$, then the process is a birth-death chain if all transitions from a state i are 0 except those to $i-1$, i, and $i+1$. We denote these transitions by p_i^+, $1-p_i^+-p_i^-$, and p_i^+ respectively. We assume the convention that $p_0^- = 0$. Note also that if $p_i^+ = 0$ for some i, then we can restrict the state space to $[\![0,i]\!]$. So from now on assume that $p_i^+ > 0$ for all $0 \leqslant i < n$ and $p_i^- > 0$ for all $0 < i \leqslant n$. We denote the value of the process at time t by x_t.

Theorem A.1. *A birth-death process is reversible with respect to the distribution*

$$\pi(i) = \prod_{j=1}^{i} \frac{p_{j-1}^+}{p_j^-} \Big/ \sum_{k=0}^{n} \prod_{j=1}^{k} \frac{p_{j-1}^+}{p_j^-}. \tag{A.0.1}$$

Proof. Note that the denominator in (A.0.1) is just a normalizing factor and we can simply show that the numerator is stationary. Calculating, we see

$$p_i^+ \prod_{j=1}^{i} \frac{p_{j-1}^+}{p_j^-} = p_{i+1}^- \prod_{j=1}^{i+1} \frac{p_{j-1}^+}{p_j^-}. \tag{A.0.2}$$

So, we immediately find that π is the stationary distribution. ■

Now we consider the mean hitting time of 0 for state i, defined as

$$t_i = \mathbb{E}\left[T|x_0 = i\right] = \mathbb{E}\left[\min\left\{t : x_t = 0\right\}|x_0 = i\right]. \tag{A.0.3}$$

Theorem A.2. *The expected hitting time of 0 from state i is*

$$\sum_{l=1}^{i}\sum_{k=0}^{n-1}\frac{\prod_{j=l+1}^{k} p_j^+}{\prod_{j=l+1}^{k+1} p_j^-}. \tag{A.0.4}$$

Proof. Note that the time for the chain to go from state i to state $i-1$ is the same as the time required for the chain to go from state 1 to 0, except with all the parameters shifted by $i-1$. So, denoting $T_i^j := \min\{t : x_t = i\}$ given $x_0 = j$, we can decompose T_0^j into a sum of similar random variables:

$$T_0^j = \sum_{l=1}^{j} T_{l-1}^l. \tag{A.0.5}$$

Consider T_0^1. In one step the process moves from state 1 to state 0 with probability p_1^-; it stays at state 1 with probability $1 - p_1^+ - p_1^-$; otherwise the process moves to state 2. Now, the expected number of steps for the process to return to state 1, is then $1/\pi\tilde(1)$, where $\tilde\pi$ is the stationary distribution of a birth-death chain with the state 0 removed:

$$\tilde\pi(1) = \left(1 - p_i^-\right) \Big/ \sum_{k=1}^{n-1}\prod_{j=2}^{k}\frac{p_{j-1}^+}{p_j^-} \tag{A.0.6}$$

which is easily found using Theorem A.1. Therefore,

$$\mathbb{E}T_0^1 = 1 + \left(1 - p_1^-\right)\left(\frac{1}{1 - p_i^-}\sum_{k=1}^{n-1}\prod_{j=2}^{k}\frac{p_{j-1}^+}{p_j^-} + \mathbb{E}T_0^1\right) \tag{A.0.7}$$

and so

$$\mathbb{E}T_0^1 = \sum_{k=0}^{n-1}\frac{\prod_{j=1}^{k} p_j^+}{\prod_{j=1}^{k+1} p_j^-}. \tag{A.0.8}$$

By the same argument, we find

$$\mathbb{E}T_l^{l+1} = \sum_{k=0}^{n-1}\frac{\prod_{j=l+1}^{k} p_j^+}{\prod_{j=l+1}^{k+1} p_j^-}. \tag{A.0.9}$$

Finally, applying (A.0.5), we obtain the result. ■

B

Concentration inequalities

In this section, we provide a brief review of large deviation estimates and concentration inequalities with a focus on those used above. A large deviation estimate (LDE) controls atypical behavior of sums of independent (or sometimes weakly dependent) random variables, whereas a concentration inequality controls the convergence of an average of independent (or sometimes weakly dependent) random variables to their mean. For a more in-depth review of LDEs, see for example [44, 278]. Many LDEs follow directly by applying Markov's inequality, so we state this now.

Theorem B.1 (Markov's inequality). *Let X be a nonnegative random variable and $t > 0$. Then*

$$\mathbb{P}\{X \geqslant t\} \leqslant \frac{\mathbb{E}X}{t}. \tag{B.0.1}$$

Proof. Define the indicator random variable $\mathbf{1}_{X \geqslant t}$. Then $t\mathbf{1}_{X \geqslant t} \leqslant X$, thus $\mathbb{E}[t\mathbf{1}_{X \geqslant t}] \leqslant \mathbb{E}X$. Therefore,

$$\mathbb{P}\{X \geqslant t\} = \mathbb{E}\mathbf{1}_{X \geqslant t} \leqslant \frac{\mathbb{E}X}{t}.$$

$\blacksquare$

This very simple result has lots of scope. The general idea is to define a nonnegative, increasing function f of some random variable X and note that Markov's inequality implies

$$\mathbb{P}\{X \geqslant t\} \leqslant \mathbb{P}\{f(X) \geqslant f(t)\} \leqslant \frac{\mathbb{E}f(X)}{f(t)}. \tag{B.0.2}$$

Normally, f is chosen as x^k or $e^{\lambda X}$ where k or λ is optimized to strengthen the inequality. If the random variable X is a sum of centered, independent random variables, $\sum_{i=1}^{N}(X_i - \mathbb{E}X_i)$, the function takes the form

$$\prod_{i=1}^{N} \exp\left(\lambda\left(X_i - \mathbb{E}X_i\right)\right). \tag{B.0.3}$$

In this way we get several inequalities.

Lemma B.2 (Hoeffding's inequality). *Suppose that $X_1, \ldots, X_N$ are i.i.d. Bernoulli random variables with parameter $p \in [0,1]$. Define $X := \sum_{i=1}^{N} X_i$. Then*

$$\mathbb{P}\left\{|X - \mathbb{E}X| \geqslant \delta\sqrt{N}\right\} \leqslant 2\exp\left(-2\delta^2\right) \tag{B.0.4}$$

for all $\delta > 0$.

Lemma B.2 states that X fluctuates about its expectation on the order of $\sqrt{N}$, and the probability of a fluctuant greater than $\delta\sqrt{N}$ decays exponential with $\delta > 0$. The next lemma bounds fluctuation of larger orders, and thus they occur even more infrequently. We shall only need a lower bound in this case:

Lemma B.3 (Multiplicative Chernoff bound). *Suppose that $X_1, \ldots, X_N$ are i.i.d. Bernoulli random variables with parameter $p \in [0,1]$. Define $X := \sum_{i=1}^{N} X_i$. Then*

$$\mathbb{P}\left\{X \leqslant (1-\varepsilon)\mathbb{E}X\right\} \leqslant \exp\left(-\frac{\varepsilon^2 p}{2}N\right) \tag{B.0.5}$$

and

$$\mathbb{P}\left\{X \geqslant (1+\varepsilon)\mathbb{E}X\right\} \leqslant \exp\left(-\frac{\varepsilon^2 p}{2}N\right). \tag{B.0.6}$$

We remark that far more general statements of Lemmas B.2 and B.3 are possible, but we state only the versions we use in Section 6.3.

Finally, we state a LDE for weighted sums of independent random variables with the following conditions on their moments:

$$\mathbb{E}X = 0, \quad \mathbb{E}|X|^2 = \sigma^2, \quad \text{and} \quad \mathbb{E}|X|^k \leqslant (Ck)^{Ck}, \tag{B.0.7}$$

for some positive constant $C > 0$ (not dependent on N or k) and for $k \geqslant 1$.

LEMMA B.4. *Suppose the independent random variables* $\left(a_i^{(N)}\right)_{i=1}^N$ *for* $N \in \mathbb{N}$ *satisfy* (B.0.7) *and that* $\left(A_i^{(N)}\right)_{i=1}^N$ *for* $N \in \mathbb{N}$ *are constants in* $\mathbb{R}$. *Then*

$$\mathbb{P}\left[\left|\sum_{i=1}^N a_i A_i\right| \geqslant \sigma (\log N)^{C+C\xi} \left(\sum_{i=1}^N |A_i|^2\right)^{1/2}\right] \leqslant e^{-\nu(\log N)^{1+\xi}}. \tag{B.0.8}$$

In words, we can bound the sum $\sum_{i=1}^N a_i A_i$ *on the same order as the norm of the coefficients with high probability.*

To prove this lemma we use a high-moment Markov inequality, so first we need a result bounding the higher moments of this sum.

LEMMA B.5. *Suppose the independent random variables* $\left(a_i^{(N)}\right)_{i=1}^N$ *for* $N \in \mathbb{N}$ *satisfy* (B.0.7) *and that* $\left(A_i^{(N)}\right)_{i=1}^N$ *for* $N \in \mathbb{N}$ *are constants in* $\mathbb{R}$. *Then*

$$\mathbb{E}\left|\sum_{i=1}^N a_i A_i\right|^k \leqslant (Ck)^{Ck} \left(\sum_{i=1}^N |A_i|^2\right)^{k/2} \tag{B.0.9}$$

PROOF. Without loss of generality let $\sigma = 1$. Let $A^2 := \sum_i |A_i|^2$, then by the classical Marcinkiewicz-Zygmund inequality [279] in the first line, we get

$$\begin{aligned}
\mathbb{E}\left|\sum_i a_i A_i\right|^k &\leqslant (Ck)^{k/2}\mathbb{E}\left|\left(\sum_i |A_i|^2 |a_i|^2\right)^{1/2}\right|^k \\
&= (Ck)^{k/2} A^k \mathbb{E}\left[\left(\sum_i \frac{|A_i|^2}{A^2} |a_i|^2\right)^{k/2}\right] \\
&\leqslant (Ck)^{k/2} A^k \mathbb{E}\left[\sum_i \frac{|A_i|^2}{A^2} |a_i|^k\right] \\
&= (Ck)^{k/2} A^k \sum_i \frac{|A_i|^2}{A^2} \mathbb{E}|a_i|^k \\
&\leqslant (Ck)^{Ck+k/2} A^k \\
&\leqslant (Ck)^{Ck} A^k,
\end{aligned} \tag{B.0.10}$$

where we have used Jensen's inequality in the third line and assumption (B.0.7) in line 5. ■

Proof of Lemma B.4. Without loss of generality let $\sigma = 1$. The proof is a simple application of Markov's inequality, Theorem B.1. Let $k = \nu(\log N)^{1+\xi}$, then by Lemma B.5, we get

$$\begin{aligned}
\mathbb{P}\left[\left|\sum_i a_i A_i\right| \geqslant (\log N)^{C+C\xi}\left(\sum_i |A_i|^2\right)^{1/2}\right] &= \mathbb{P}\left[\left|\sum_i a_i A_i\right|^k \geqslant (\log N)^{Ck+Ck\xi}\left(\sum_i |A_i|^2\right)^{k/2}\right] \\
&\leqslant \frac{\mathbb{E}\left|\sum_i a_i A_i\right|^k}{(\log N)^{Ck+Ck\xi}\left(\sum_i |A_i|^2\right)^{k/2}} \\
&\leqslant \left(\frac{Ck}{(\log N)^{1+\xi}}\right)^{Ck} \\
&= (C\nu)^{C\nu(\log N)^{1+\xi}} \\
&\leqslant e^{-\nu(\log N)^{1+\xi}},
\end{aligned} \tag{B.0.11}$$

for $\nu \leqslant e^{-1}$ small enough. ■

C

The bond percolated hypercube

In this section, we consider a lazy random walk on the bond percolation hypercube. This random Markov chain is defined differently than the process in Definition 2.26.

Definition C.1. *This process is a random Markov chain* $(\Gamma_n^{(p)}, \mathcal{M}^{(p)})$ *with state space* $\Gamma_n = [\![\kappa]\!]^n$ *and each* $\mathcal{M}^{(p)}(\boldsymbol{\alpha}, \boldsymbol{\beta})$ *a random variable defined as follows: for*

$$p \in \left(1 - \kappa^{-1/(\kappa-1)}, 1\right], \tag{C.0.1}$$

and for $e \in E$ *let* x_e *be i.i.d. Bern*(p)*, where* E *is the edge-set of the hypercube* Γ_n*. Then*

$$\mathcal{M}^{(p)}(\boldsymbol{\alpha}, \boldsymbol{\beta}) := \begin{cases} \frac{x_e}{2d_{\boldsymbol{\alpha}}} & \text{if } e = (\boldsymbol{\alpha}, \boldsymbol{\beta}) \\ \frac{1}{2} & \text{if } \boldsymbol{\alpha} = \boldsymbol{\beta} \\ 0 & \text{otherwise} \end{cases}, \tag{C.0.2}$$

where $d_{\boldsymbol{\alpha}} := \sum_{\boldsymbol{\beta}:(\boldsymbol{\alpha},\boldsymbol{\beta}) \in E} x_{(\boldsymbol{\alpha},\boldsymbol{\beta})}$.

As is discussed in Subsection 2.5.4, we know the process in Definition C.1 is irreducible, aperiodic, and reversible with probability $1 - o(1)$.

THEOREM C.2. *Let* $(\Gamma_n^{(p)}, \mathcal{M}^{(p)})$ *be the process defined in Definition C.1. Denote its stationary distribution by* π*, then*

$$\frac{1}{2\kappa^n \log^2 n} \leqslant \min_{\boldsymbol{\alpha}} \pi(\boldsymbol{\alpha}) \leqslant \max_{\boldsymbol{\alpha}} \pi(\boldsymbol{\alpha}) \leqslant \frac{\kappa}{\kappa^n} \tag{C.0.3}$$

with probability $1 - o(1)$.

PROOF. We can study the stationary distribution of a random walk on the hypercube using Equation (2.3.5). We find

$$\pi(\boldsymbol{\alpha}) = \frac{d_{\boldsymbol{\alpha}}}{\sum_{\boldsymbol{\beta}} d_{\boldsymbol{\beta}}}, \tag{C.0.4}$$

where

$$d_{\boldsymbol{\alpha}} := \sum_{\boldsymbol{\beta}:(\boldsymbol{\alpha},\boldsymbol{\beta})\in E} x_{(\boldsymbol{\alpha},\boldsymbol{\beta})}, \tag{C.0.5}$$

Note that Equation (C.0.5) tells us that $d_{\boldsymbol{\alpha}}$ is a sum of $(\kappa-1)n$ independent random variables that are each $\mathrm{Bern}(p)$. A trivial upper bound on each $d_{\boldsymbol{\alpha}}$ is $(\kappa-1)n$ and thus $\max_{\boldsymbol{\alpha}} d_{\boldsymbol{\alpha}} \leqslant (\kappa-1)n$. For a lower bound, let $k = Cn/(\log n)^2$, then

$$\mathbb{P}\left\{d_{\boldsymbol{\alpha}} \leqslant C\frac{n}{(\log n)^2}\right\} = \sum_{i=0}^{k} \binom{(\kappa-1)n}{i} p^i (1-p)^{(\kappa-1)n-i} \leqslant \mathcal{O}\left((1-p)^{(\kappa-1)n-k} n^k\right) \tag{C.0.6}$$

and using a union bound, we have

$$\mathbb{P}\left\{\min_{\boldsymbol{\alpha}} d_{\boldsymbol{\alpha}} \leqslant C\frac{n}{(\log n)^2}\right\} \leqslant \mathcal{O}\left(e^{n(\log\kappa + C/\log n + (\kappa-1)\log(1-p) - (C/\log n)^2 \log(1-p))}\right) \to 0 \tag{C.0.7}$$

since $\log\kappa < (1-\kappa)\log(1-p)$ for $p > p_c$. Note this improves on the bound in Equation (2.5.22). A similar concentration argument (see Lemma B.2) shows

$$\mathbb{P}\left\{\left|\sum_{\boldsymbol{\alpha}} d_{\boldsymbol{\alpha}} - p(\kappa-1)n\kappa^n\right| \geqslant \sqrt{C\log n}\sqrt{(\kappa-1)n\kappa^n}\right\} \leqslant 2\exp\left(-2C\log n\right) = 2n^{-2C} \to 0. \tag{C.0.8}$$

Thus, using Equations (C.0.7) and (C.0.8), we see

$$\frac{1}{2\kappa^n \log^2 n} \leqslant \min_{\boldsymbol{\alpha}} \pi(\boldsymbol{\alpha}) \leqslant \max_{\boldsymbol{\alpha}} \pi(\boldsymbol{\alpha}) \leqslant \frac{n}{p\kappa^n} \leqslant \frac{n}{\kappa^{n-1}}, \tag{C.0.9}$$

with probability $1 - o(1)$. ■

THEOREM C.3. *Let* $(\Gamma_n^{(p)}, \mathcal{M}^{(p)})$ *be the process defined in Definition C.1 and denote its relaxation time by* t_{rel}*. Suppose*

$$p > p_{c,3} := \sqrt[3]{1 - (1/\kappa)^{1/(\kappa-1)}} = \sqrt[3]{p_c} > p_c, \tag{C.0.10}$$

then

$$t_{rel} = \mathcal{O}\left(n^2 \log n\right) \tag{C.0.11}$$

with probability $1 - o(1)$.

PROOF. Let E be the edge set of the hypercube. Exactly the same argument as is given in the proof of Theorem 2.27 shows that for $p > p_{\mathrm{c},3}$, there is some path of length 3 between all pairs $\boldsymbol{\alpha}, \boldsymbol{\beta}$ such that $(\boldsymbol{\alpha}, \boldsymbol{\beta}) \in E$ with probability $1 - o(1)$. Call this event $\mathcal{E}$.

Now, we use the Comparison Theorem 2.23 applied to the lazy random walk on the regular hypercube, to bound the mixing time of the hypercube after percolation. We construct paths as follows conditional on the event $\mathcal{E}$: For any pair $\boldsymbol{\alpha}, \boldsymbol{\beta}$ such that $(\boldsymbol{\alpha}, \boldsymbol{\beta}) \in E$ where the edge $x_{(\boldsymbol{\alpha},\boldsymbol{\beta})} = 1$, we set $\phi_{\boldsymbol{\alpha},\boldsymbol{\beta}} = (\boldsymbol{\alpha}, \boldsymbol{\beta})$. For any pair $\boldsymbol{\alpha}, \boldsymbol{\beta}$ such that $(\boldsymbol{\alpha}, \boldsymbol{\beta}) \in E$ where the edge $x_{(\boldsymbol{\alpha},\boldsymbol{\beta})} = 0$, we choose $\phi_{\boldsymbol{\alpha},\boldsymbol{\beta}}$ arbitrarily from paths of the form (2.5.28), of which one is guaranteed to exist on the event $\mathcal{E}$.

Consider an arbitrary edge e in Γ, there are $3(n-1)(\kappa-1)$ paths of the form (2.5.28) that pass through e. To see this note that e can either form the beginning, middle, or end of a path, and after e is fixed there are exactly $(\kappa-1)(n-1)$ paths of this form. Alternatively, we may use the symmetry of the hypercube to note that the answer should not depended on the choice of e. Then a path of the form (2.5.28) is determined by its end points, of which there are $n(\kappa-1)\kappa^n$ possible choices, and then there are $(n-1)(\kappa-1)$ paths for each pair of end points. Moreover, each path contains 3 edges, so the each edge must be included in

$$\frac{3n(n-1)(\kappa-1)\kappa^n}{(\kappa-1)n\kappa^n} = 3(n-1)(\kappa-1) \tag{C.0.12}$$

paths.

Let $E' := \left\{(\boldsymbol{\alpha}, \boldsymbol{\beta}) : x_{(\boldsymbol{\alpha}, \boldsymbol{\beta})} = 1\right\}$. Then by Equation (C.0.8), we have

$$\begin{aligned} B &= \max_{(\alpha,\beta)\in E'} \left(\left(\pi(\boldsymbol{\alpha}) \frac{1}{2d_{\boldsymbol{\alpha}}} \right)^{-1} \sum_{\tilde{\alpha},\tilde{\beta}:(\alpha,\beta)\in\phi_{\tilde{\alpha}\tilde{\beta}}} \frac{1}{\kappa^n} \frac{1}{2(\kappa-1)n} \cdot 3 \right) \\ &\leqslant 9(\kappa-1)(n-1) \frac{\sum_{\boldsymbol{\beta}} d_{\boldsymbol{\beta}}}{(\kappa-1)n\kappa^n} \\ &\leqslant 9(\kappa-1)(n-1). \end{aligned} \tag{C.0.13}$$

Finally,

$$t_{\mathrm{rel}} \leqslant (9\kappa(\kappa-1)(n-1))\,(n \log n) \left(\max_{\alpha\in\Gamma} \frac{\pi(\alpha)}{\tilde{\pi}(\alpha)} \right) = \mathcal{O}\left(n^2 \log n\right). \tag{C.0.14}$$

■

D

Notation

4.1 Basic notation ($[\![N]\!]$, $[\![i,j]\!]$, $\mathbf{v}(i)$, $M(i,j)$, $M(S,S')$, $\mathbb{P}$, $\mathbb{E}$. $\|\cdot\|_2$, $\|\cdot\|_{\text{TV}}$, $\sum_j^{(i)}\cdot$, $\sum_{i\neq j}\cdot$). We abbreviate $[\![N]\!] := \{1,\ldots,N\} = [1,N]\cap\mathbb{N}$ and $[\![i,j]\!] := \{i,i+1,\ldots,j\} = [i,j]\cap\mathbb{N}$. We denote the ith coordinate of a vector $\mathbf{v}$ by $\mathbf{v}(i)$ and the i,j entry of a matrix M by $M(i,j)$. For two subsets $S,S'\subseteq[\![N]\!]$, we let

$$M(S,S') := \sum_{i\in S}\sum_{j\in S'} M(i,j). \tag{D.1.1}$$

Throughout $\mathbb{P}$ is a probability measure (annealed if there is a random environment) and $\mathbb{E}$ is the expectation associated with $\mathbb{P}$. Sometimes we use a subscript for $\mathbb{P}$ or $\mathbb{E}$ to specify the distribution for a random variable. Often the random variable is the initial condition of a stochastic process, but in cases where there is ambiguity $\mathbb{P}_{X\sim\nu}$ is used to specify that X is distributed according to μ. The Euclidean norm on vectors is denoted by $\|\mathbf{v}\|_2^2 := \sum_i \mathbf{v}(i)^2$ and the total variation norm on a function over Γ is denoted by $\|f\|_{\text{TV}} := \max_{S\subseteq\Gamma}|f(S)|$. We also use the following convention to abbreviate summations:

$$\sum_{j}^{(i)} f(j) := \sum_{j=1}^{N} f(j)\mathbf{1}(j \neq i) \tag{D.1.2}$$

for a summand $f(\cdot)$. Similarly,

$$\sum_{i \neq j} f(i,j) := \sum_{i=1}^{N} \sum_{j}^{(i)} f(i,j) \tag{D.1.3}$$

for a summand $f(\cdot,\cdot)$.

4.2 Genotype space (Γ, n, α, β, χ, $\mathcal{M}$, $\mathcal{F}$, f, π, $\mathcal{D}$). We use Γ to denote a set of genotypes and sometimes, for a sequence of these sets, we use $n \in \mathbb{N}$ to index them like $(\Gamma_n)_{n\in\mathbb{N}}$. The size of Γ is some function of n. Genotypes in Γ are denoted with $\alpha, \beta \in \Gamma$ and subsets are denoted with $\chi \subseteq \Gamma$. This means that χ is used for target sets in Chapter 5. We use $\mathcal{M} : \Gamma \times \Gamma \to [0,1]$ for a mutation kernel on Γ and $\mathcal{F} : \Gamma \to (0,\infty)$ as a fitness function on Γ. When there are only two values for fitness, the values 1 and f are typically used. Taken together $(\Gamma, \mathcal{M}, \mathcal{F})$ make up a genotype space in the sense of Definition 2.1 and $(\Gamma, \mathcal{M})$ form a mutation process in the sense of Definition 2.2. The stationary distribution of a mutation process is denoted by π. Finally, $\mathcal{D}$ is a distance metric on Γ, which is sometimes the graph distance of the transition matrix $\mathcal{M}$. In general, greek letters correspond to genotype space and calligraphic letters to functions on genotype space.

4.3 Synonyms for genotypes. Depending on the context several synonyms for a genotype are used. When the genotype corresponds to a state of a Markov chain *state* is often used. Similarly, *point* is used when the geometric properties of genotype spaces are being discussed, and *vertex* is used when the geometry is described by a graph. *Sequence* is used for genotypes in the hypercube (see Section 2.1). *Permutation* is used for genotypes in the symmetric group (see Section 2.2).

4.4 Populations (N, i, j, S, R, r, $\mathbb{D}$, W). The size of a population is denoted by $N \in \mathbb{N}$. Individuals in a population are indexed by these numbers, so specific individuals in the population are referred to with $i, j \in [\![N]\!]$. Subsets of the population are denoted by $S \subseteq [\![N]\!]$. We also use $R \subseteq [\![N]\!]$ as a subset of the population, but this is primarily in the context of a replacement set in an evolutionary process. With a replacement function for an evolutionary process $r : R \to [\![N]\!]$, (R, r) forms an update step of an evolutionary process. These updates are sampled from some distribution, which we denote by $\mathbb{D}$. For structured populations (see Section 3.3), we use W to denote the $N \times N$ weight matrix of the graph describing the population structure. Latin letters are used for populations.

4.5 Evolutionary dynamics (t, $\mathbf{x}_t$, $\mathbf{x}_t(i)$, x_t, ρ, μ). Time or discrete time steps are denoted by $t \in \{0, 1, 2 \ldots\}$. The state of an evolutionary process at time t is given by $\mathbf{x}_t \in \Gamma^N$, where bold is used to indicate that it is a vector. Coordinate i of the vector yields the genotype of individual i, so $\mathbf{x}_t(i) \in \Gamma$. Sometimes it is necessary to project $\mathbf{x}_t$ and we denote the projection with x_t. The fixation probability (see Chapter 3) for an evolutionary process is denoted by ρ throughout. The stationary distribution of a evolutionary process is denoted by μ.

4.6 Markov chains ($t_{\mathbf{mix}}$, $t_{\mathbf{rel}}$, γ_*, T). The mixing time of a Markov chain is denoted by t_{mix} and the relaxation time by $t_{\text{rel}} = \gamma_*^{-1}$, where γ is the spectral gap. Stopping times of stochastic processes are typically denoted using T.

Definition D.1 (Rapid mixing). *We say that a Markov chain* $(\Gamma, \mathcal{M})$ *mixes rapidly or is rapidly mixing if*

$$t_{mix} \leqslant p\left(\log |\Gamma|\right) \tag{D.6.1}$$

for some polynomial p.

4.7 Asymptotic notation ($\mathcal{O}(\cdot)$, $o(\cdot)$, $\Theta(\cdot)$, $\cdot \ll \cdot$). We use the familiar asymptotic notations throughout. We point out that the notation is either asymptotic with respect to $n \to \infty$ or $N \to \infty$ depending on the context. Often the context makes clear which limit is being used, but as a guideline Chapter 2 uses n, Chapters 3 and 6 use N, and Chapters 4 and 5 use both n and N.

Bibliography

[1] Martin A Nowak. *Evolutionary dynamics : exploring the equations of life.* Belknap Press of Harvard University Press, Cambridge, Massachusetts, 2006.

[2] Krishnendu Chatterjee, Andreas Pavlogiannis, Ben Adlam, and Martin A. Nowak. The Time Scale of Evolutionary Innovation. *PLoS Comput. Biol.*, 10(9), 2014.

[3] Ben Adlam and Martin A. Nowak. Universality of fixation probabilities in randomly structured populations. *Sci. Rep.*, 4, jul 2014.

[4] B Adlam, K Chatterjee, and M A Nowak. Amplifiers of selection. *Proc. R. Soc. London A Math. Phys. Eng. Sci.*, 471(2181), 2015.

[5] Alex McAvoy, Ben Adlam, Benjamin Allen, and Martin A Nowak. Stationary frequencies and mixing times for neutral drift processes with spatial structure. 2018.

[6] Andreas Pavlogiannis, Krishnendu Chatterjee, Ben Adlam, and Martin A. Nowak. Cellular cooperation with shift updating and repulsion. *Sci. Rep.*, 5, 2015.

[7] Charleston Noble, Ben Adlam, George M Church, Kevin M Esvelt, and Martin A Nowak. Current CRISPR gene drive systems are likely to be highly invasive in wild populations. *bioRxiv*, 2017.

[8] Steven Weinberg. Cosmology. *Ann. Phys. (N. Y).*, 54:612, 2008.

[9] Wikipedia. Timeline of the evolutionary history of life. *https://en.wikipedia.org/wiki/Timeline_of_the_ evolutionary_history_of_life*, 2018-05-04.

[10] Andrew H. Knoll and Martin A. Nowak. The timetable of evolution. *Sci. Adv.*, 3(5), 2017.

[11] Andreas Wagner. *Arrival of the fittest : how nature innovates.* Current, an imprint of Penguin Books, New York, New York, 2015.

[12] Leslie G. Valiant. Evolvability. *J. ACM*, 56(1):1–21, 2009.

[13] Warren John Ewens. *Mathematical Population Genetics, I. Theoretical introduction*, volume 27. Springer, 2004.

[14] F Jacob. Evolution and tinkering. *Science (80-.).*, 196(4295):1161–1166, 1977.

[15] H. Allen Orr. Book Review: No Free Lunch. *Bost. Rev.*, 2002.

[16] William A Dembski. *No free lunch: Why specified complexity cannot be purchased without intelligence.* Rowman & Littlefield, 2006.

[17] Santiago F Elena and Richard E Lenski. Microbial genetics: evolution experiments with microorganisms: the dynamics and genetic bases of adaptation. *Nat. Rev. Genet.*, 4(6):457, 2003.

[18] Jeffrey E Barrick, Dong Su Yu, Sung Ho Yoon, Haeyoung Jeong, Tae Kwang Oh, Dominique Schneider, Richard E Lenski, and Jihyun F Kim. Genome evolution and adaptation in a long-term experiment with Escherichia coli. *Nature*, 461(7268):1243, 2009.

[19] Daniel L Hartl and Andrew G Clark. *Principles of population genetics*. Sinauer Associate, 1998.

[20] Manfred Eigen. Selforganization of matter and the evolution of biological macromolecules. *Naturwissenschaften*, 58(10):465–523, 1971.

[21] Anthony Griffiths, Susan Wessler, Sean Carroll, Bernard Swynghedauw, and U.S. Department of Health and Human Services. Introduction to genetic analysis. *Biofutur*, April:707, 2007.

[22] Richard Dawkins. The Selfish Gene. *30th Anniv. Ed. a new Introd. by Author*, page 384, 1976.

[23] Alexander Rosenberg and Frederic Bouchard. Fitness. In Edward N Zalta, editor, *Stanford Encycl. Philos.* Metaphysics Research Lab, Stanford University, fall 2015 edition, 2015.

[24] Leslie Valiant. *Probably Approximately Correct: Nature's Algorithms for Learning and Prospering in a Complex World*. Basic Books (AZ), 2013.

[25] Benjamin Allen and Corina E. Tarnita. Measures of success in a class of evolutionary models with fixed population size and structure. *J. Math. Biol.*, 68(1-2):109–143, 2014.

[26] M Nei. *Molecular population genetics and evolution*, volume 40. Elsevier Science Publishing Co Inc., 1975.

[27] Motoo Kimura. *The Neutral Theory of Molecular Evolution*. Cambridge University Press, 1983.

[28] Motoo Kimura. Evolutionary rate at the molecular level. *Nature*, 217(5129):624–626, 1968.

[29] Manfred Eigen, John McCaskill, and Peter Schuster. Molecular quasi-species, 1988.

[30] Martin A Nowak. What is a quasispecies? *Trends in Ecology & Evolution*, 7(4):118–121, 1992.

[31] Christof K Biebricher and Manfred Eigen. What is a quasispecies? In *Quasispecies: Concept and Implications for Virology*, pages 1–31. Springer, 2006.

[32] Peter F Stadler. Fitness Landscapes. *J. Mol. Struct. Theochem*, 463(1-2):7–19, 2002.

[33] J Maynard Smith. The limitations of molecular evolution. *Sci. Speculates an Anthol. Partly-baked Ideas, Ed. by IJ Good. Basic Books, Inc., New York*, pages 252–256, 1962.

[34] John Maynard Smith. Natural selection and the concept of a protein space. *Nature*, 225(5232):563–564, 1970.

[35] E. Bornberg-Bauer. How are model protein structures distributed in sequence space? *Biophys. J.*, 73(5):2393–2403, 1997.

[36] E. Bornberg-Bauer and H. S. Chan. Modeling evolutionary landscapes: Mutational stability, topology, and superfunnels in sequence space. *Proc. Natl. Acad. Sci.*, 96(19):10689–10694, 1999.

[37] Mark A. DePristo, Daniel M. Weinreich, and Daniel L. Hartl. Missense meanderings in sequence space: A biophysical view of protein evolution, 2005.

[38] Andreas Wagner. *The origins of evolutionary innovations: a theory of transformative change in living systems.* Oxford University Press, Oxford ; New York, 2011.

[39] J D Hermes, S C Blacklow, and J R Knowles. Searching sequence space by definably random mutagenesis: improving the catalytic potency of an enzyme. *Proc. Natl. Acad. Sci. U. S. A.*, 87(2):696–700, 1990.

[40] D. Bartel and J. Szostak. Isolation of new ribozymes from a large pool of random sequences. *Science (80-.).*, 261(5127):1411–1418, 1993.

[41] Philip A Romero and Frances H Arnold. Exploring protein fitness landscapes by directed evolution. *Nature Reviews Molecular Cell Biology*, 10(12):866, 2009.

[42] Frances H. Arnold. The Library of Maynard-Smith: My Search for Meaning in the Protein Universe. *Microbe Mag.*, 6(7):316–318, 2011.

[43] Michael Conrad. The geometry of evolution. *BioSystems*, 24(1):61–81, 1990.

[44] Michel Ledoux. *The concentration of measure phenomenon*, volume v. 89 of *Mathematical surveys and monographs,*. American Mathematical Society, Providence, R.I., 2001.

[45] H. J. Muller. The relation of recombination to mutational advance. *Mutat. Res. - Fundam. Mol. Mech. Mutagen.*, 1(1):2–9, 1964.

[46] Stuart Kauffman and Simon Levin. Towards a general theory of adaptive walks on rugged landscapes. *J. Theor. Biol.*, 128(1):11–45, 1987.

[47] Daniel A. Levinthal. Adaptation on Rugged Landscapes. *Manage. Sci.*, 43(7):934–950, 1997.

[48] H Allen Orr. The genetic theory of adaptation: a brief history. *Nature Reviews Genetics*, 6(2):119, 2005.

[49] John H. Gillespie. Molecular Evolution Over the Mutational Landscape. *Evolution (N. Y).*, 38(5):1116–1129, 1984.

[50] John H. Gillespie. A simple stochastic gene substitution model. *Theor. Popul. Biol.*, 23(2):202–215, 1983.

[51] H. Allen Orr. The distribution of fitness effects among beneficial mutations in Fisher's geometric model of adaptation. *J. Theor. Biol.*, 238(2):279–285, 2006.

[52] Lawrence Hueston Harper. *Global methods for combinatorial isoperimetric problems*, volume 90. Cambridge University Press, 2004.

[53] Béla Bollobás. *Combinatorics: set systems, hypergraphs, families of vectors, and combinatorial probability.* Cambridge University Press, 1986.

[54] Shlomo Hoory, Nathan Linial, and Avi Wigderson. Expander graphs and their applications. *Bull. Am. Math. Soc.*, 43(4):439–561, 2006.

[55] P Erdös. On two problems of information theory. *Magy. Tud. Akad. Mat. Kut. Int. Közl*, 1963.

[56] Bernt Lindström. On a combinatory detection problem. I. *Magy. Tud. Akad. Mat. Kut. Int. Közlad.*, 9:195–207, 1964.

[57] José Cáceres, Carmen Hernando, Merce Mora, Ignacio M Pelayo, Maria L Puertas, Carlos Seara, and David R Wood. On the metric dimension of cartesian products of graphs. *SIAM J. Discret. Math.*, 21(2):423–441, 2007.

[58] G. E. Uhlenbeck and L. S. Ornstein. On the theory of the Brownian motion. *Phys. Rev.*, 36(5):823–841, 1930.

[59] Mark Kac. Random walk and the theory of Brownian motion. *Am. Math. Mon.*, 54(7):369–391, 1947.

[60] P. Diaconis, R.L. Graham, and J.A. Morrison. Asymptotic analysis of a random walk on a hypercube with many dimensions. *Random Struct. & Algorithms*, 1(1), 1990.

[61] Persi Diaconis and Ron Graham. An affine walk on the hypercube. *J. Comput. Appl. Math.*, 41(1-2):215–235, 1992.

[62] P. Diaconis. The cutoff phenomenon in finite Markov chains. *Proc. Natl. Acad.*, 93(4):1659–64, 1996.

[63] Persi. Diaconis. *Group representations in probability and statistics*, volume v. 11 of *Lecture notes-monograph series ;*. Institute of Mathematical Statistics, Hayward, Calif., 1988.

[64] José María, Carmen Segarra, and Alfredo Ruiz. Chromosomal homology and molecular organization of Muller's elements D and E in the Drosophila repleta species group. *Genetics*, 145(2):281–295, 1997.

[65] D Hartl and E Jones. *Genetics: Analysis of Genes and Genomes.* Jones & Bartlett Publ, 2000.

[66] Persi Diaconis and Mehrdad Shahshahani. Generating a random permutation with random transpositions. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 57(2):159–179, 1981.

[67] David A. Levin, Yuval Peres, and Elizabeth L. Wilmer. *Markov Chains and Mixing Times.* American Mathematical Soc., 2009.

[68] Rick Durrett. Shuffling Chromosomes. *J. Theor. Probab.*, 16(3):725–750, 2003.

[69] Nathanaël Berestycki and Rick Durrett. A phase transition in the random transposition random walk. *Probab. Theory Relat. Fields*, 136(2):203–233, 2006.

[70] Alexander Gamburd and Igor Pak. Expansion of product replacement graphs. *Combinatorica*, 26(4):411–429, 2006.

[71] Persi Diaconis. Some things we've learned (about Markov chain Monte Carlo). *Bernoulli*, 19(4):1294–1305, 2013.

[72] Anthony W. Knapp. *Basic Algebra.* Birkhäuser Basel, 2006.

[73] A Markoff. Extension of the law of large numbers to dependent events. *Bull. Soc. Phys. Math. Kazan*, 15:135–156, 1906.

[74] Henri Poincaré. *Calcul des probabilités.* Gauthier-Villars, 1912.

[75] Persi Diaconis, R L Graham, and William M Kantor. The mathematics of perfect shuffles. *Adv. Appl. Math.*, 4(2):175–196, 1983.

[76] Persi Diaconis. Mathematical developments from the analysis op riffle shuffling. In *Groups, Comb. Geom. Durham 2001*, pages 73–97. World Scientific, 2003.

[77] Nathanaël Berestycki. The hyperbolic geometry of random transpositions. *Ann. Probab.*, 34(2):429–467, 2006.

[78] Andrey Zharkikh. Estimation of evolutionary distances between nucleotide sequences. *J. Mol. Evol.*, 39(3):315–329, 1994.

[79] Federico Squartini and Peter F. Arndt. Quantifying the stationarity and time reversibility of the nucleotide substitution process. *Mol. Biol. Evol.*, 25(12):2525–2535, 2008.

[80] David M. Mccandlish. On the findability of genotypes. *Evolution (N. Y).*, 67(9):2592–2603, 2013.

[81] Stuart A Kauffman. The origins of order: Self-organization and selection in evolution. In *Spin Glas. Biol.*, pages 61–100. World Scientific, 1992.

[82] Gabriela Ochoa, Marco Tomassini, Sebástien Vérel, and Christian Darabos. A study of NK landscapes' basins and local optima networks. In *Proc. 10th Annu. Conf. Genet. Evol. Comput.*, pages 555–562. ACM, 2008.

[83] Leigh Van Valen. A new evolutionary law. *Evol Theory*, 1:1–30, 1973.

[84] George Von Dassow and Ed Munro. Modularity in animal development and evolution: Elements of a conceptual framework for EvoDevo, 1999.

[85] Günter P. Wagner, Mihaela Pavlicev, and James M. Cheverud. The road to modularity, 2007.

[86] David Aldous. Random walks on finite groups and rapidly mixing Markov chains. In *Séminaire Probab. XVII 1981/82*, pages 243–297. Springer, 1983.

[87] Nathanaël Berestycki. Lectures on mixing times. http://www.statslab.cam.ac.uk/ beresty/teach/Mixing/ mixing2.pdf, 2014.

[88] Fred Solomon. Random Walks in a Random Environment. *Ann. Probab.*, 3(1):1–31, 1975.

[89] Ya. G. Sinai. The limiting behaviour of a one-dimensional random walk in a random medium. *Theory Probab. its Appl.*, 27:256, 1982.

[90] Remco van der Hofstad. Stochastic processes on random graphs. In *Lect. notes 47th Summer Sch. Probab. Saint-Flour*, 2017.

[91] Eugene P. Wigner. Random Matrices in Physics. *SIAM Rev.*, 9(1):1–23, 1967.

[92] Paul Erdős and Alfréd Rényi. On the evolution of random graphs. *Publ. Math. Inst. Hungarian Acad. Sci.*, 5, 1960.

[93] Carl C. Z. Dou. *studies of random walks on groups and random graphs*. Ph.d. dissertation, Massachusetts Institute of Technology, 1992.

[94] Martin Hildebrand. Random walks on random simple graphs. *Random Struct. Algorithms*, 8(4):301–318, 1996.

[95] N. Fountoulakis and B. A. Reed. The evolution of the mixing rate of a simple random walk on the giant component of a random graph. *Random Struct. Algorithms*, 33(1):68–86, 2008.

[96] Itai Benjamini, Gady Kozma, and Nicholas Wormald. The mixing time of the giant component of a random graph. *Random Struct. Algorithms*, 45(3):383–407, 2014.

[97] Asaf Nachmias and Yuval Peres. Critical random graphs: Diameter and mixing time. *Ann. Probab.*, 36(4):1267–1286, 2008.

[98] Eyal Lubetzky and Allan Sly. Cutoff phenomena for random walks on random regular graphs. *Duke Math. J.*, 153(3):475–510, 2010.

[99] Paul Erdös and Joel Spencer. Evolution of the n-cube. *Comput. Math. with Appl.*, 5(1):33–39, 1979.

[100] Miklós Ajtai, János Komlós, and Endre Szemerédi. Largest random component of ak-cube. *Combinatorica*, 2(1):1–7, 1982.

[101] B Bollobás, Y Kohayakawa, and T Łuczak. The Evolution of Random Subgraphs of the Cube. *Random Struct. Algorithms*, 3(1):55–90, 1992.

[102] Remco van der Hofstad and Asaf Nachmias. Unlacing hypercube percolation: a survey. *Metrika*, 77(1):23–50, jan 2014.

[103] Markus. Heydenreich and Remco van der Hofstad. *Progress in High-Dimensional Percolation and Random Graphs*. Springer International Publishing, 1 edition, 2017.

[104] Itai Benjamini and Elchanan Mossel. On the mixing time of a simple random walk on the super critical percolation cluster. *Probab. Theory Relat. Fields*, 125(3):408–420, 2003.

[105] B. Morris and Yuval Peres. Evolving sets, mixing and heat kernel bounds. *Probab. Theory Relat. Fields*, 133(2):245–266, 2005.

[106] Remco Van Der Hofstad and Asaf Nachmias. Hypercube percolation. *J. Eur. Math. Soc.*, 19(3):725–814, 2017.

[107] Christian Reidys, Peter F Stadler, and Peter Schuster. Generic properties of combinatory maps: Neutral networks of RNA secondary structures. *Bull. Math. Biol.*, 59(2):339–397, mar 1997.

[108] Peter Schuster. Prediction of RNA secondary structures: From theory to models and real molecules. *Reports Prog. Phys.*, 69(5):1419–1477, 2006.

[109] João F. Matias Rodrigues and Andreas Wagner. Evolutionary plasticity and innovations in complex metabolic reaction networks. *PLoS Comput. Biol.*, 5(12), 2009.

[110] José Aguilar-Rodríguez, Joshua L. Payne, and Andreas Wagner. A thousand empirical adaptive landscapes and their navigability. *Nat. Ecol. Evol.*, 1(2), 2017.

[111] A. R. Davidson and R. T. Sauer. Folded proteins occur frequently in libraries of random amino acid sequences. *Proc. Natl. Acad. Sci.*, 91(6):2146–2150, 1994.

[112] Theodosius Dobzhansky and Theodosius Grigorievich Dobzhansky. *Genetics and the Origin of Species*, volume 11. Columbia university press, 1937.

[113] Sergey Gavrilets and Janko Gravner. Percolation on the fitness hypercube and the evolution of reproductive isolation. *J. Theor. Biol.*, 184(1):51–64, 1997.

[114] Sergey Gavrilets. Evolution and speciation on holey adaptive landscapes. *Trends in ecology & evolution*, 12(8):307–312, 1997.

[115] I. A. Campbel, R. Jullien, R. Botet, and J. M. Flesselles. Random walks on a hypercube and spin glass relaxation. *J. Phys. C Solid State Phys.*, 20(4):L47–L51, 1987.

[116] Peter Schuster, Walter Fontana, Peter F. Stadler, and Ivo L. Hofacker. From Sequences to Shapes and Back: A Case Study in RNA Secondary Structures. *Proc. R. Soc. B Biol. Sci.*, 255(1344):279–284, 1994.

[117] B. Morris and a. Sinclair. Random walks on truncated cubes and sampling 0-1 knapsack solutions. *40th Annu. Symp. Found. Comput. Sci. (Cat. No.99CB37039)*, 34(1):195–226, 1999.

[118] Christos H. Papadimitriou. *Computational Complexity*. Pearson, 1994.

[119] Barbara Drossel. Biological evolution and statistical physics. *Adv. Phys.*, 50(2):209–295, 2001.

[120] Sewall Wright. Evolution in mendelian populations. 1931. *Bull. Math. Biol.*, 52(1-2):241–295, 1990.

[121] Sewall Wright. Genic and Organismic Selection. *Evolution (N. Y).*, 34(5):825–843, 1980.

[122] Luis-Miguel Chevin, Russell Lande, and Georgina M Mace. Adaptation, plasticity, and extinction in a changing environment: towards a predictive theory. *PLoS Biol.*, 8(4):e1000357, 2010.

[123] Erez Lieberman, Christoph Hauert, and Martin A Nowak. Evolutionary dynamics on graphs. *Nature*, 433(Jan.):312–316, 2005.

[124] Kamran Kaveh, Alex McAvoy, and Martin A Nowak. The effect of spatial fitness heterogeneity on fixation probability. *arXiv Prepr. arXiv1709.03031*, 2017.

[125] J Maynard Smith and George R Price. The logic of animal conflict. *Nature*, 246(5427):15, 1973.

[126] Mohan Matthen and André Ariew. Two Ways of Thinking About Fitness and Natural Selection. *J. Philos.*, 99(2):55–83, 2002.

[127] André Ariew and R. C. Lewontin. The confusions of fitness, 2004.

[128] S Ohno. *Evolution by Gene Duplication*. Springer, 1970.

[129] A. Wagner. Redundant gene functions and natural selection. *J. Evol. Biol.*, 12(1):1–16, 1999.

[130] Spencer V Muse and Brandon S Gaut. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. *Mol. Biol. Evol.*, 11(5):715–724, 1994.

[131] N Goldman and Z Yang. A codon-based model of nucleotide substitution for protein- coding DNA sequences. *Mol. Biol. Evol.*, 11:725–736, 1994.

[132] Rasmus Nielsen and Ziheng Yang. Estimating the distribution of selection coefficients from phylogenetic data with applications to mitochondrial and viral DNA. *Mol. Biol. Evol.*, 20(8):1231–1239, 2003.

[133] Gilean A T McVean and Jorge Vieira. Inferring parameters of mutation, selection and demography from patterns of synonymous site evolution in Drosophila. *Genetics*, 157(1):245–257, 2001.

[134] Alexei V. Finkelstein. Implications of the random characteristics of protein sequences for their three-dimensional structure. *Curr. Opin. Struct. Biol.*, 4(3):422–428, 1994.

[135] Alan R. Davidson, Kevin J. Lumb, and Robert T. Sauer. Cooperatively folded proteins in random sequence libraries. *Nat. Struct. Biol.*, 2(10):856–864, 1995.

[136] A. D. Keefe and J. W. Szostak. Functional proteins from a random-sequence library. *Nature*, 410(6829):715–718, 2001.

[137] J F Reidhaar-Olson and R T Sauer. Functionally acceptable substitutions in two alpha-helical regions of lambda repressor. *Proteins*, 7(4):306–16, 1990.

[138] S. V. Taylor, K. U. Walter, P. Kast, and D. Hilvert. Searching sequence space for protein catalysts. *Proc. Natl. Acad. Sci.*, 98(19):10596–10601, 2001.

[139] Aderonke Babajide, Ivo L Hofacker, Manfred J Sippl, and Peter F Stadler. Neutral networks in protein space: A computational study based on knowledge-based potentials of mean force. *Fold. Des.*, 2(5):261–269, 1997.

[140] Peter Y. Chou and Gerald D. Fasman. Prediction of Protein Conformation. *Biochemistry*, 13(2):222–245, 1974.

[141] V a Simossis and J Heringa. Integrating protein secondary structure prediction and multiple sequence alignment. *Curr. Protein Pept. Sci.*, 5(4):249–66, 2004.

[142] Walter Pirovano and Jaap Heringa. Protein secondary structure prediction. In *Data Mining Techniques for the Life Sciences*, pages 327–348. Springer, 2010.

[143] I Y Koh, V A Eyrich, M A Marti-Renom, D Przybylski, M S Madhusudhan, N Eswar, O Grana, F Pazos, A Valencia, A Sali, and B Rost. EVA: Evaluation of protein structure prediction servers. *Nucleic Acids Res*, 31(13):3311–3315, 2003.

[144] Burkhard Rost. Review: Protein secondary structure prediction continues to rise, 2001.

[145] Carol A. Rohl, Charlie E M Strauss, Kira M S Misura, and David Baker. Protein Structure Prediction Using Rosetta. *Methods Enzymol.*, 383(2003):66–93, 2004.

[146] Rhiju Das and David Baker. Macromolecular Modeling with Rosetta. *Annu. Rev. Biochem.*, 77(1):363–382, 2008.

[147] Sergey Ovchinnikov, Hahnbeom Park, Neha Varghese, Po-Ssu Huang, Georgios A. Pavlopoulos, David E. Kim, Hetunandan Kamisetty, Nikos C. Kyrpides, and David Baker. Protein structure determination using metagenome sequence data. *Science (80-.).*, 355(6322):294–298, 2017.

[148] Walter Fontana, Peter F. Stadler, Erich G. Bornberg-Bauer, Thomas Griesmacher, Ivo L. Hofacker, Manfred Tacker, Pedro Tarazona, Edward D. Weinberger, and Peter Schuster. RNA folding and combinatory landscapes. *Phys. Rev. E*, 47(3):2083–2099, 1993.

[149] W Fontana, D a M Konings, P F Stadler, and P Schuster. Statistics of Rna Secondary Structures. *Biopolymers*, 33(9):1389–1404, 1993.

[150] P. Schuster, W. Fontana, P. F. Stadler, and I. L. Hofacker. From Sequences to Shapes and Back: A Case Study in RNA Secondary Structures. *Proc. R. Soc. B Biol. Sci.*, 255(1344):279–284, 1994.

[151] Peter Schuster. Prediction of RNA secondary structures: From theory to models and real molecules. *Reports Prog. Phys.*, 69(5):1419–1477, 2006.

[152] Martijn A Huynen. Exploring phenotype space through neutral evolution. *J. Mol. Evol.*, 43(3):165–169, 1996.

[153] Areejit Samal, João F. Matias Rodrigues, Jürgen Jost, Olivier C. Martin, and Andreas Wagner. Genotype networks in metabolic reaction spaces. *BMC Syst. Biol.*, 4, 2010.

[154] João F. Matias Rodrigues and Andreas Wagner. Genotype networks, innovation, and robustness in sulfur metabolism. *BMC Syst. Biol.*, 5, 2011.

[155] Johannes Berg, Stana Willmann, and Michael Lässig. Adaptive evolution of transcription factor binding sites. *BMC Evol. Biol.*, 4, 2004.

[156] Percy Deift. Universality for mathematical and physical systems. In *Proc. Int. Congr. Math.*, volume I, pages 125–152, Zürich, 2007. European Mathematical Society.

[157] László Erdős and Horng-Tzer Yau. Universality of local spectral statistics of random matrices. *Bull. Am. Math. Soc.*, 49(3):377–414, 2012.

[158] Alexei Borodin and Ivan Corwin. Macdonald processes. *Probab. Theory Relat. Fields*, 158(1-2):225–400, 2014.

[159] Christopher J Wills. A mechanism for rapid allopatric speciation. *Am. Nat.*, 111(979):603–605, 1977.

[160] M Nei. Mathematical models of speciation and genetic distance. *Popul. Genet. Ecol. Acad. Press. New York*, pages 723–766, 1976.

[161] S Gavrilets. A dynamical theory of speciation on holey adaptive landscapes. *Am. Nat.*, 154(1):1–22, 1999.

[162] Paul Ehrenfest and Tatjana Ehrenfest-Afanassjewa. *Über zwei bekannte Einwände gegen das Boltzmannsche H-Theorem.* Hirzel, 1907.

[163] P. A P Moran. Random processes in genetics. *Math. Proc. Cambridge Philos. Soc.*, 54(1):60–71, 1958.

[164] Patrick Alfred Pierce Moran. *The statistical processes of evolutionary theory.* Oxford University Press, Oxford, England, 1962.

[165] Arne Traulsen and Christoph Hauert. Stochastic Evolutionary Game Dynamics. In *Rev. Nonlinear Dyn. Complex.*, volume 2, pages 25–61. Wiley-VCH Verlag GmbH & Co. KGaA, 2010.

[166] Julian Keilson. Log-concavity and log-convexity in passage time densities of diffusion and birth-death processes. *J. Appl. Probab.*, 8(2):391–398, 1971.

[167] Julian Keilson. *Markov chain models—rarity and exponentiality*, volume 28. Springer Science & Business Media, 2012.

[168] James Allen Fill. The passage time distribution for a birth-and-death chain: Strong stationary duality gives a first stochastic proof. *J. Theor. Probab.*, 22(3):543–557, 2009.

[169] Hisashi Ohtsuki, Christoph Hauert, Erez Lieberman, and Martin A Nowak. A simple rule for the evolution of cooperation on graphs and social networks. *Nature*, 441:502–505, may 2006.

[170] Lorens A. Imhof and Martin A. Nowak. Evolutionary game dynamics in a Wright-Fisher process. *J. Math. Biol.*, 52(5):667–681, 2006.

[171] Ricky Der, Charles L. Epstein, and Joshua B. Plotkin. Generalized population models and the nature of genetic drift. *Theor. Popul. Biol.*, 80(2):80–99, 2011.

[172] Takeo Maruyama. A simple proof that certain quantities are independent of the geographical structure of population. *Theor. Popul. Biol.*, 5(2):148–154, 1974.

[173] Laura Hindersin and Arne Traulsen. Counterintuitive properties of the fixation time in network-structured populations. *J. R. Soc. Interface*, 11(99), aug 2014.

[174] Krishnendu Chatterjee, Rasmus Ibsen-Jensen, and Martin A Nowak. Faster Monte-Carlo Algorithms for Fixation Probability of the Moran Process on Undirected Graphs. *arXiv Prepr. arXiv1706.06931*, 2017.

[175] Bahram Houchmandzadeh and Marcel Vallade. The fixation probability of a beneficial mutation in a geographically structured population. *New J. Phys.*, 13:073020, jul 2011.

[176] T Monk, P Green, and M Paulin. Martingales and fixation probabilities of evolutionary graphs. *Proc. R. Soc. A Math. Phys. Eng. Sci.*, 470(2165):20130730, 2014.

[177] Marcus Frean, Paul B Rainey, and Arne Traulsen. The effect of population structure on the rate of evolution. *Proc. R. Soc. B*, 280(1762):20130211, 2013.

[178] Josep Díaz, Leslie Ann Goldberg, George B Mertzios, David Richerby, Maria Serna, and Paul G Spirakis. Approximating fixation probabilities in the generalized moran process. *Algorithmica*, 69(1):78–91, 2014.

[179] Rasmus Ibsen-Jensen, Krishnendu Chatterjee, and Martin A Nowak. Computational complexity of ecological and evolutionary spatial dynamics. *Proc. Natl. Acad. Sci.*, 112(51):15636–15641, 2015.

[180] Josep Díaz, Leslie Ann Goldberg, David Richerby, and Maria Serna. Absorption time of the Moran process. *Random Struct. Algorithms*, 49(1):137–159, 2016.

[181] Masatoshi Nei. *Mutation-driven evolution*. OUP Oxford, 2013.

[182] C. E. Tarnita, T. Antal, H. Ohtsuki, and M. A. Nowak. Evolutionary dynamics in set structured populations. *Proc. Natl. Acad. Sci.*, 106(21):8601–8604, 2009.

[183] A. Traulsen, C. Hauert, H. De Silva, M. A. Nowak, and K. Sigmund. Exploration dynamics in evolutionary games. *Proc. Natl. Acad. Sci.*, 106(3):709–712, 2009.

[184] L. Loewe and W. G. Hill. The population genetics of mutations: good, bad and indifferent. *Philos. Trans. R. Soc. B Biol. Sci.*, 365(1544):1153–1167, 2010.

[185] Christoph Hauert and Lorens A. Imhof. Evolutionary games in deme structured, finite populations. *J. Theor. Biol.*, 299:106–112, 2012.

[186] P. M. Altrock, A. Traulsen, and M. A. Nowak. Evolutionary games on cycles with strong selection. *Phys. Rev. E*, 95(2), 2017.

[187] T. Antal, H. Ohtsuki, J. Wakeley, P. D. Taylor, and M. A. Nowak. Evolution of cooperation by phenotypic similarity. *Proc. Natl. Acad. Sci.*, 106(21):8597–8600, 2009.

[188] Tibor Antal, Arne Traulsen, Hisashi Ohtsuki, Corina E Tarnita, and Martin A Nowak. Mutation-selection equilibrium in games with multiple strategies. *J. Theor. Biol.*, 258(4):614–622, 2009.

[189] Corina E Tarnita, Hisashi Ohtsuki, Tibor Antal, Feng Fu, and Martin A Nowak. Strategy selection in structured populations. *J. Theor. Biol.*, 259(3):570–581, 2009.

[190] Martin Nowak, Corina E Tarnita, and Tibor Antal. Evolutionary dynamics in structured populations. *Phil.Trans.R.Soc.B*, 365(1537):19–30, 2010.

[191] Corina E Tarnita, Nicholas Wage, and Martin A Nowak. Multiple strategies in structured populations. *Proc. Natl. Acad. Sci.*, 108(6):2334–2337, 2011.

[192] Chaitanya S. Gokhale and Arne Traulsen. Strategy abundance in evolutionary many-player games with multiple strategies. *J. Theor. Biol.*, 283(1):180–191, 2011.

[193] Benjamin Allen, Arne Traulsen, Corina E. Tarnita, and Martin A. Nowak. How mutation affects evolutionary games on graphs. *J. Theor. Biol.*, 299:97–105, 2012.

[194] Bin Wu, Julián García, Christoph Hauert, and Arne Traulsen. Extrapolating Weak Selection in Evolutionary Games. *PLoS Comput. Biol.*, 9(12), 2013.

[195] Bin Wu, Arne Traulsen, and Chaitanya Gokhale. Dynamic Properties of Evolutionary Multi-player Games in Finite Populations. *Games*, 4(2):182–199, 2013.

[196] Pablo Catalán, Jesús M. Seoane, and Miguel A F Sanjuán. Mutation-selection equilibrium in finite populations playing a Hawk-Dove game. *Commun. Nonlinear Sci. Numer. Simul.*, 25(1-3):66–73, 2015.

[197] Yanling Zhang, Aizhi Liu, and Changyin Sun. Impact of migration on the multi-strategy selection in finite group-structured populations. *Sci. Rep.*, 6, 2016.

[198] Michael Lynch and William G. Hill. Phenotypic Evolution by Neutral Mutation. *Evolution (N. Y).*, 40(5):915, 1986.

[199] Bernard Derrida and Luca Peliti. Evolution in a flat fitness landscape. *Bull. Math. Biol.*, 53(3):355–382, 1991.

[200] Howard Ochman. Neutral Mutations and Neutral Substitutions in Bacterial Genomes. *Mol. Biol. Evol.*, 20(12):2091–2096, 2003.

[201] Masatoshi Nei. Selectionism and neutralism in molecular evolution. *Molecular biology and evolution*, 22(12):2318–2342, 2005.

[202] Jesse D. Bloom, Philip A. Romero, Zhongyi Lu, and Frances H. Arnold. Neutral genetic drift can alter promiscuous protein functions, potentially aiding functional evolution. *Biol. Direct*, 2, 2007.

[203] Rick Durrett. *Probability: Theory and Examples.* Cambridge University Press, 2010.

[204] Peter D. Taylor, Troy Day, and Geoff Wild. From inclusive fitness to fixation probability in homogeneous structured populations. *J. Theor. Biol.*, 249(1):101–110, 2007.

[205] F. Débarre. Fidelity of parent-offspring transmission and the evolution of social behavior in structured populations. *J. Theor. Biol.*, 420:26–35, 2017.

[206] Paul Fearnhead. The common ancestor at a nonneutral locus. *J. Appl. Probab.*, 39(1):38–54, 2002.

[207] Jesse E. Taylor. The common ancestor process for a wright-fisher diffusion. *Electron. J. Probab.*, 12:808–847, 2007.

[208] E. van Nimwegen, J. P. Crutchfield, and M. Huynen. Neutral evolution of mutational robustness. *Proc. Natl. Acad. Sci.*, 96(17):9716–9720, 1999.

[209] Benjamin Allen, Christine Sample, Yulia Dementieva, Ruben C. Medeiros, Christopher Paoletti, and Martin A. Nowak. The Molecular Clock of Neutral Evolution Can Be Accelerated or Slowed by Asymmetric Spatial Structure. *PLoS Comput. Biol.*, 11(2), 2015.

[210] Michael Doebeli, Christoph Hauert, and Timothy Killingback. The evolutionary origin of cooperators and defectors. *Science (80-.).*, 306(5697):859–862, 2004.

[211] Joe Yuichiro Wakano and Yoh Iwasa. Evolutionary branching in a finite population: Deterministic branching vs. stochastic branching. *Genetics*, 193(1):229–241, 2013.

[212] R Montenegro and P Tetali. Mathematical Aspects of Mixing Times in Markov Chains. *Found. Trends® Theor. Comput. Sci.*, 1(3):237–354, 2005.

[213] John Wakeley. Probability Theory for the Coalescent. *Coalescent Theory*, pages 17–39, 2000.

[214] J. Swetina. First and second moments and the mean hamming distance in a stochastic replication-mutation model for biological macromolecules. *J. Math. Biol.*, 27(4):463–483, 1989.

[215] Benjamin Allen and Martin A. Nowak. Games on graphs. *Eur. Math. Soc. Surv. Math. Sci.*, pages 113–151, 2014.

[216] Benjamin Allen, Gabor Lippner, Yu-Ting Chen, Babak Fotouhi, Naghmeh Momeni, Shing-Tung Yau, and Martin A. Nowak. Evolutionary dynamics on any population structure. *Nature*, 544(7649):227–230, 2017.

[217] R a Fisher. The Genetical Theory of Natural Selection. *Genetics*, 154:272, 1930.

[218] Paul G. Higgs and Bernard Derrida. Genetic distance and species formation in evolving populations. *J. Mol. Evol.*, 35(5):454–465, 1992.

[219] Thomas H Cormen, Charles E Leiserson, and Ronald L Rivest. *Introduction to Algorithms.* MIT Press, 2001.

[220] David M McCandlish and Arlin Stoltzfus. Modeling Evolution Using the Probability of Fixation: History and Implications. *Q. Rev. Biol.*, 89(3):225–252, sep 2014.

[221] Craig A. Fogle, James L. Nagle, and Michael M. Desai. Clonal interference, multiple mutations and adaptation in large asexual populations. *Genetics*, 180(4):2163–2173, 2008.

[222] Gregory I. Lang, David Botstein, and Michael M. Desai. Genetic variation and the fate of beneficial mutations in asexual populations. *Genetics*, 188(3):647–661, 2011.

[223] Gregory I. Lang, Daniel P. Rice, Mark J. Hickman, Erica Sodergren, George M. Weinstock, David Botstein, and Michael M. Desai. Pervasive genetic hitchhiking and clonal interference in forty evolving yeast populations. *Nature*, 500(7464):571–574, 2013.

[224] Katya Kosheleva and Michael M. Desai. The dynamics of genetic draft in rapidly adapting populations. *Genetics*, 195(3):1007–1025, 2013.

[225] David M McCandlish, Charles L Epstein, and Joshua B Plotkin. Formal properties of the probability of fixation: Identities, inequalities and approximations. *Theor. Popul. Biol.*, 99(Supplement C):98–113, 2015.

[226] Motoo Kimura and Takeo Maruyama. The substitutional load in a finite population. *Heredity (Edinb).*, 24(1):101, 1969.

[227] J. L. King and T. H. Jukes. Non-Darwinian Evolution. *Science (80-.).*, 164(3881):788–798, 1969.

[228] John H. Gillespie. Some Properties of Finite Populations Experiencing Strong Selection and Weak Mutation. *Am. Nat.*, 121(5):691, 1983.

[229] H. Allen Orr. The Population Genetics of Adaptation: The Distribution of Factors Fixed during Adaptive Evolution. *Evolution (N. Y).*, 52(4):935, 1998.

[230] Guy Sella and Aaron E Hirsh. The application of statistical physics to evolutionary biology. *Proc. Natl. Acad. Sci. U. S. A.*, 102(27):9541–6, 2005.

[231] David Aldous and Jim Fill. Reversible Markov chains and random walks on graphs, 2002.

[232] David J. Aldous. Markov chains with almost exponential hitting times. *Stoch. Process. their Appl.*, 13(3):305–310, 1982.

[233] Gérard Ben Arous and Véronique Gayrard. Elementary potential theory on the hypercube. *Electron. J. Probab.*, 13:1726–1807, 2008.

[234] Gérard Ben Arous and Jiří Černý. The arcsine law as a universal aging scheme for trap models. *Commun. Pure Appl. Math.*, 61(3):289–329, 2008.

[235] J Černý and V Gayrard. Hitting time of large subsets of the hypercube. *Random Struct. Algorithms*, pages 1–14, 2008.

[236] Martijn A Huynen, Peter F Stadler, and Walter Fontana. Smoothness within ruggedness: the role of neutrality in adaptation. *Proc. Natl. Acad. Sci.*, 93(1):397–401, 1996.

[237] Vernon M Ingram and Others. The hemoglobins in genetics and evolution. *hemoglobins Genet. Evol.*, 1963.

[238] Norman Harold Horowitz. Biochemical genetics of Neurospora. In *Adv. Genet.*, volume 3, pages 33–71. Elsevier, 1950.

[239] Cyrus Chothia. One thousand families for the molecular biologist. *Nature*, 357:543–544, 1992.

[240] Paul K Keese and Adrian Gibbs. Origins of genes:" big bang" or continuous creation? *Proc. Natl. Acad. Sci.*, 89(20):9489–9493, 1992.

[241] Bernard Dujon, David Sherman, Gilles Fischer, Pascal Durrens, Serge Casaregola, Ingrid Lafontaine, Jacky De Montigny, Christian Marck, Cécile Neuvéglise, Emmanuel Talla, and Others. Genome evolution in yeasts. *Nature*, 430(6995):35, 2004.

[242] Angelo Pavesi, Gkikas Magiorkinis, and David G Karlin. Viral proteins originated de novo by overprinting can be identified by codon usage: application to the "gene nursery" of Deltaretroviruses. *PLoS Comput. Biol.*, 9(8):e1003162, 2013.

[243] Zhilong Bao, Maureen A Clancy, Raquel F Carvalho, Kiona Elliott, and Kevin M Folta. Identification of novel growth regulators in plant populations expressing random peptides. *Plant Physiol.*, pages pp—-00577, 2017.

[244] Simon A Levin and Robert T Paine. Disturbance, patch formation, and community structure. *Proc. Natl. Acad. Sci.*, 71(7):2744–2747, 1974.

[245] Simon A Levin. Population dynamic models in heterogeneous environments. *Annu. Rev. Ecol. Syst.*, 7(1):287–310, 1976.

[246] Richard Durrett and Simon A Levin. Stochastic spatial models: a user's guide to ecological applications. *Philos. Trans. R. Soc. London. Ser. B Biol. Sci.*, 343(1305):329–350, 1994.

[247] Mark Broom and Jan Rychtář. An analysis of the fixation probability of a mutant on special classes of non-directed graphs. *Proc. R. Soc. A Math. Phys. Eng. Sci.*, 464(2098):2609–2627, oct 2008.

[248] Marcus Frean, Paul B Rainey, and Arne Traulsen. The effect of population structure on the rate of evolution. *Proc. R. Soc. B Biol. Sci.*, 280(1762):20130211, jul 2013.

[249] Tibor Antal, Sidney Redner, and Vishal Sood. Evolutionary dynamics on degree-heterogeneous graphs. *Phys. Rev. Lett.*, 96:188104, 2006.

[250] B Sinervo and CM Lively. The rock–paper–scissors game and the evolution of alternative male strategies. *Nature*, 380:240 – 243, 1996.

[251] Benjamin Kerr, Margaret a Riley, Marcus W Feldman, and Brendan J M Bohannan. Local dispersal promotes biodiversity in a real-life game of rock-paper-scissors. *Nature*, 418(6894):171–4, jul 2002.

[252] Yu-Ting Chen. Sharp benefit-to-cost rules for the evolution of cooperation on regular graphs. *Ann. Appl. Probab.*, 23(2):637–664, 2013.

[253] Mark Broom and Jan Rychtář. *Game-theoretical models in biology.* Chapman & Hall/CRC mathematical and computational biology series. Chapman and Hall/CRC Press, Taylor and Francis Group, Boca Raton, FL, 2013.

[254] Martin A Nowak and Robert M May. Evolutionary games and spatial chaos. *Nature*, 359(6398):826–829, 1992.

[255] György Szabó and Gabor Fath. Evolutionary games on graphs. *Phys. Rep.*, 446(4):97–216, 2007.

[256] Matjaž Perc. Coherence resonance in a spatial prisoner's dilemma game. *New J. Phys.*, 8(2):22, 2006.

[257] Matjaž Perc and Attila Szolnoki. Social diversity and promotion of cooperation in the spatial prisoner's dilemma game. *Phys. Rev. E*, 77(1):11904, 2008.

[258] Feng Fu, Martin A Nowak, and Christoph Hauert. Invasion and expansion of cooperators in lattice populations: Prisoner's dilemma vs. Snowdrift games. *J. Theor. Biol.*, 266(3):358–366, 2010.

[259] Jorge M Pacheco, Arne Traulsen, and Martin A Nowak. Coevolution of strategy and structure in complex networks with dynamical linking. *Phys. Rev. Lett.*, 97(25):258103, 2006.

[260] Karin Johst, Michael Doebeli, and Roland Brandl. Evolution of complex dynamics in spatially structured populations. *Proc. R. Soc. B Biol. Sci.*, 266(1424):1147–1154, 1999.

[261] Mark Broom, Jan Rychtář, and B Stadler. Evolutionary dynamics on small-order graphs. *J. Interdiscip. Math.*, 12:129–140, 2009.

[262] Jaroslav Ispolatov and Michael Doebeli. Diversification along environmental gradients in spatially structured populations. *Evol. Ecol. Res.*, 11:295–304, 2009.

[263] Josep Díaz, Leslie Ann Goldberg, George B. Mertzios, David Richerby, Maria Serna, J., and Paul G. Spirakis. On the fixation probability of superstars. *Proc. R. Soc. A Math. Phys. Eng. Sci.*, 469:20130193, 2013.

[264] Alastair Jamieson-Lane and Christoph Hauert. Fixation probabilities on superstars, revisited and revised. *arXiv Prepr. arXiv1312.6333*, 2013.

[265] Richard Durrett and Simon Levin. The Importance of Being Discrete (and Spatial). *Theor. Popul. Biol.*, 46:363–394, 1994.

[266] Michael P Hassell, Hugh N Comins, and Robert M May. Species coexistence and self-organizing spatial dynamics. *Nature*, 370:290–292, 1994.

[267] Paul B Rainey and Katrina Rainey. Evolution of cooperation and conflict in experimental bacterial populations. *Nature*, 425:72–74, 2003.

[268] Mickael Le Gac and Michael Doebeli. Environmental viscosity does not affect the evolution of cooperation during experimental evolution of colicigenic bacteria. *Evolution*, 64(2):522–33, feb 2010.

[269] Benjamin Allen, Jeff Gore, and Martin A Nowak. Spatial dilemmas of diffusible public goods. *Elife*, 2:e01169, jan 2013.

[270] Duncan J Watts and Steven H Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393:440–442, jun 1998.

[271] Alain Barrat and Martin Weigt. On the properties of small-world network models. *Eur. Phys. J. B*, 13(3):547–560, jan 2000.

[272] Mayuko Nakamaru and Simon A Levin. Spread of two linked social norms on complex interaction networks. *J. Theor. Biol.*, 230(1):57–64, sep 2004.

[273] Valmir C. Barbosa, Raul Donangelo, and Sergio R. Souza. Early appraisal of the fixation probability in directed networks. *Phys. Rev. E*, 82:046114, 2010.

[274] Andreas Pavlogiannis, Josef Tkadlec, Krishnendu Chatterjee, and Martin A Nowak. Strong Amplifiers of Natural Selection: Proofs. *arXiv Prepr. arXiv1802.02509*, 2018.

[275] Martin A Nowak, Franziska Michor, and Yoh Iwasa. The linear process of somatic evolution. *Proc. Natl. Acad. Sci.*, 100(25):14966–14969, 2003.

[276] Christoforos Hadjichrysanthou, Mark Broom, and Istvan Z Kiss. Approximating evolutionary dynamics on networks using a Neighbourhood Configuration model. *J. Theor. Biol.*, 312:13–21, jul 2012.

[277] Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *Science (80-.).*, 286(5439):509–512, oct 1999.

[278] Fan Rong King Chung and Linyuan Lu. *Complex graphs and networks*, volume no. 107 of *CBMS regional conference series in mathematics ;*. American Mathematical Society, Providence, RI, 2006.

[279] Daniel W Stroock. *Probability theory : an analytic view.* Cambridge University Press, Cambridge, UK, 2nd edition, 2011.

[280] Ben Adlam and Martin A. Nowak. Universality of fixation probabilities in randomly structured populations. *Sci. Rep.*, 4, 2014.

[281] Frances H. Arnold. The Library of Maynard-Smith: My Search for Meaning in the Protein Universe. *Microbe Mag.*, 6(7):316–318, 2011.

[282] Vikram Alva, Michael Remmert, Andreas Biegert, Andrei N. Lupas, and Johannes Söding. A galaxy of folds. *Protein Sci.*, 19(1):124–130, 2010.

[283] T Ohta. Population Size and Rate of Evolution. *J. Mol. Evol.*, 1(4):305–14, 1972.

[284] Nathanaël Berestycki. Recent progress in coalescent theory. *Ensaios Mat.*, 16(1):1–193, 2009.

[285] Diethard Tautz and Tomislav Domazet-Lošo. The evolutionary origin of orphan genes, 2011.

[286] Johannes Söding and Andrei N. Lupas. More than the sum of their parts: On the evolution of proteins from peptides, 2003.

[287] W Fontana and P Schuster. A computer model of evolutionary optimization. *Biophys Chem*, 26(2-3):123–147, 1987.

[288] Manfred Eigen and Peter Schuster. The Hypercycle. *Naturwissenschaften*, 65(1):7–41, 1978.

[289] J. William Schopf. The first billion years: When did life emerge? *Elements*, 2(4):229–233, 2006.

[290] Abigail C. Allwood, John P. Grotzinger, Andrew H. Knoll, Ian W. Burch, Mark S. Anderson, Max L. Coleman, and Isik Kanik. Controls on development and diversity of Early Archean stromatolites. *Proc. Natl. Acad. Sci. U. S. A.*, 106(24):9548–55, 2009.

[291] W. B. Whitman, D. C. Coleman, and W. J. Wiebe. Prokaryotes: The unseen majority. *Proc. Natl. Acad. Sci.*, 95(12):6578–6583, 1998.

[292] M. Legiewicz. Size, constant sequences, and optimal selection. *RNA*, 11(11):1701–1709, 2005.

[293] R. P. Worden. A speed limit for evolution. *J. Theor. Biol.*, 176(1):137–152, 1995.

[294] Su Chan Park, Damien Simon, and Joachim Krug. The speed of evolution in large asexual populations. *J. Stat. Phys.*, 138(1):381–410, 2010.

[295] Claus O. Wilke. The speed of adaptation in large asexual populations. *Genetics*, 167(4):2045–2053, 2004.

[296] W. J. Ewens. The probability of survival of a new mutant in a fluctuating environment. *Heredity (Edinb).*, 22(3):438–443, 1967.

[297] N. H. Barton. Linkage and the limits to natural selection. *Genetics*, 140(2):821–841, 1995.

[298] Tibor Antal and István Scheuring. Fixation of strategies for an evolutionary game in finite populations. *Bull. Math. Biol.*, 68(8):1923–1944, 2006.

[299] Paulo R A Campos. Fixation of beneficial mutations in the presence of epistatic interactions. *Bull. Math. Biol.*, 66(3):473–486, 2004.

[300] R C Griffiths and Simon Tavare. Ancestral Inference in Population Genetics, 1994.

[301] Claus O. Wilke. The speed of adaptation in large asexual populations. *Genetics*, 167(4):2045–2053, 2004.

[302] H. Allen Orr. The rate of adaptation in asexuals. *Genetics*, 155(2):961–968, 2000.

[303] Toby Johnson and Philip J. Gerrish. The fixation probability of a beneficial allele in a population dividing by binary fission. *Genetica*, 115(3):283–287, 2002.

[304] Philipp M. Altrock and Arne Traulsen. Fixation times in evolutionary games under weak selection. *New J. Phys.*, 11, 2009.

[305] Michael M. Desai, Daniel S. Fisher, and Andrew W. Murray. The Speed of Evolution and Maintenance of Variation in Asexual Populations. *Curr. Biol.*, 17(5):385–394, 2007.

[306] Motoo Kimura and Tomoko Ohta. The average number of generations until fixation of a mutant gene in a finite population. *Genetics*, 61(692):763–771, 1969.

[307] Man-Wah Cheung. Pairwise comparison dynamics for games with continuous strategy space. *J. Econ. Theory*, 153:344–375, 2014.

[308] Christian Hilbe, Krishnendu Chatterjee, and Martin A Nowak. Partners and rivals in direct reciprocity. *Nat. Hum. Behav.*, 2018.

[309] Lindi M Wahl and Martin A Nowak. The Continuous Prisoner's Dilemma: II. Linear Reactive Strategies with Noise. *J. Theor. Biol.*, 200(3):323–338, 1999.

[310] Lindi M Wahl and Martin A Nowak. The Continuous Prisoner's Dilemma: I. Linear Reactive Strategies. *J. Theor. Biol.*, 200(3):307–321, 1999.

[311] Timothy Killingback, Michael Doebeli, and Nancy Knowlton. Variable investment, the Continuous Prisoner's Dilemma, and the origin of cooperation. *Proc. R. Soc. London Ser. B, Contain. Pap. a Biol. character R. Soc. (Great Britain)*, 266(1430):1723–1728, 1999.

[312] U Dieckmann and M On Doebeli. On the origin of species by sympatric speciation. *Nature*, 400(22):354–357, 1999.

[313] Ulf Dieckmann and Johan A.J. Metz. Surprising evolutionary predictions from enhanced ecological realism. *Theor. Popul. Biol.*, 69(3):263–281, 2006.

[314] Timothy Killingback and Michael Doebeli. The Continuous Prisoner's Dilemma and the Evolution of Cooperation through Reciprocal Altruism with Variable Investment. *Am. Nat.*, 160(4):421–438, 2002.

[315] S. A.H. Geritz, É Kisdi, G. Meszéna, and J. A.J. Metz. Evolutionarily singular strategies and the adaptive growth and branching of the evolutionary tree. *Evol. Ecol.*, 12(1):35–57, 1998.

[316] Jayanta K. Ghosh. Probability Models for DNA Sequence Evolution, Second Edition by Richard Durrett. *Int. Stat. Rev.*, 77(2):304–304, 2009.

[317] Jörg Oechssler and Frank Riedel. Evolutionary dynamics on infinite strategy spaces. *Econ. Theory*, 17(1):141–162, 2001.

[318] Gilbert Roberts and Thomas N. Sherratt. Development of cooperative relationships through increasing investment. *Nature*, 394(6689):175–179, 1998.

[319] Matthijs Van Veelen and Peter Spreij. Evolution in games with a continuous action space. *Econ. Theory*, 39(3):355–376, 2009.

[320] John Cleveland and Azmy S. Ackleh. Evolutionary game theory on measure spaces: Well-posedness. *Nonlinear Anal. Real World Appl.*, 14(1):785–797, 2013.

[321] Man Wah Cheung. Imitative dynamics for games with continuous strategy space. *Games Econ. Behav.*, 99:206–223, 2016.

[322] Karl Sigmund. *The calculus of selfishness.* Princeton University Press, 2010.

[323] Martin Nowak and Karl Sigmund. The evolution of stochastic strategies in the Prisoner's Dilemma. *Acta Appl. Math.*, 20(3):247–265, sep 1990.

[324] Corina E Tarnita, Nicholas Wage, and Martin A Nowak. Multiple strategies in structured populations. *Proc. Natl. Acad. Sci. U. S. A.*, 108(6):2334–2337, 2011.

[325] Corina E. Tarnita, Tibor Antal, and Martin A. Nowak. Mutation-selection equilibrium in games with mixed strategies. *J. Theor. Biol.*, 261(1):50–57, 2009.

[326] Corina E. Tarnita, Hisashi Ohtsuki, Tibor Antal, Feng Fu, and Martin A. Nowak. Strategy selection in structured populations. *J. Theor. Biol.*, 259(3):570–581, 2009.

[327] M. Doebeli. *Adaptive diversification (MPB-48).* Princeton University Press, 2011.

[328] Martin Nowak. Stochastic strategies in the Prisoner's Dilemma. *Theor. Popul. Biol.*, 38(1):93–112, 1990.

[329] L. A. Imhof and M. A. Nowak. Stochastic evolutionary dynamics of direct reciprocity. *Proc. R. Soc. B Biol. Sci.*, 277(1680):463–468, 2010.

[330] Zhilong Bao, Maureen A Clancy, Raquel F. Carvalho, Kiona Elliott, and Kevin M Folta. Identification of Novel Growth Regulators in Plant Populations Expressing Random Peptides. *Plant Physiol.*, page pp.00577.2017, 2017.

[331] Themistoklis Melissourgos, Sotiris Nikoletseas, Christoforos Raptopoulos, and Paul Spirakis. Mutants and Residents with Different Connection Graphs in the Moran Process. oct 2017.

[332] Marc Kirschner and John Gerhart. Evolvability. *Proc. Natl. Acad. Sci.*, 95(15):8420–8427, 1998.

[333] Rodrigo S. Galhardo, P. J. Hastings, and Susan M. Rosenberg. Mutation as a stress response and the regulation of evolvability, jan 2007.

[334] Ole Peters and Alexander Adamou. The evolutionary advantage of cooperation. *Genetics*, 78(2), 2015.

[335] Walter Fontana and Peter Schuster. Continuity in evolution: On the nature of transitions. *Science (80-.).*, 280(5368):1451–1455, 1998.

[336] Walter Fontana. Modelling 'evo-devo' with RNA, dec 2002.

[337] Jeffrey A Fawcett, Steven Maere, and Yves Van de Peer. Plants with double genomes might have had a better chance to survive the Cretaceous-Tertiary extinction event. *Proc. Natl. Acad. Sci.*, 106(14):5737–5742, apr 2009.

[338] Ivana Bjedov, Olivier Tenaillon, Valeria Souza, Erick Denamur, Miroslav Radman, B. Gerard, Valeria Souza, Erick Denamur, Miroslav Radman, François Taddei, Ivan Matic, Bénédicte Gérard, Valeria Souza, Erick Denamur, Miroslav Radman, François Taddei, and Ivan Matic. Stress-induced mutagenesis in bacteria. *Science (80-.).*, 300(5624):1404–9, 2003.

[339] Avihu H Yona, Yair S Manor, Rebecca H Herbst, Gal H Romano, Amir Mitchell, Martin Kupiec, Yitzhak Pilpel, and Orna Dahan. Chromosomal duplication is a transient evolutionary solution to stress. *Proc. Natl. Acad. Sci.*, 109(51):21010–21015, dec 2012.

[340] NH Barton. The probability of fixation of a favoured allele in a subdivided population. *Genet. Res.*, pages 149–157, 1993.

[341] MC Whitlock. Fixation probability and time in subdivided populations. *Genetics*, 779(June):767–779, 2003.

[342] Martin A Nowak, Franziska Michor, and Yoh Iwasa. The linear process of somatic evolution. *Proc. Natl. Acad. Sci. U. S. A.*, 100(25):14966–14969, dec 2003.

[343] Natalia L Komarova, Anirvan Sengupta, and Martin A Nowak. Mutation–selection networks of cancer initiation: tumor suppressor genes and chromosomal instability. *J. Theor. Biol.*, 223(4):433–450, aug 2003.

[344] Z Patwa and Lindi M Wahl. The fixation probability of beneficial mutations. *J. R. Soc. Interface*, 5(28):1279–1289, nov 2008.

[345] Mark Broom, Christoforos Hadjichrysanthou, and Jan Rychtář. Evolutionary games on graphs and the speed of the evolutionary process. *Proc. R. Soc. A Math. Phys. Eng. Sci.*, 466(2117):1327–1346, dec 2009.

[346] Takeo Maruyama. A Markov process of gene frequency change in a geographically structured population. *Genetics*, 76(2):367–377, 1974.