LARGE DEVIATIONS APPROACH FOR STOCHASTIC GENETIC EVOLUTION

A Dissertation

Presented to

the Faculty of the Department of Mathematics

University of Houston

In Partial Fulfillment

of the Requirements for the Degree

Doctor of Philosophy

By

Aanchal Aggarwal

August 2013

LARGE DEVIATIONS APPROACH FOR STOCHASTIC GENETIC EVOLUTION

Aanchal Aggarwal

APPROVED:

Prof. Robert Azencott, Advisor

Dr. Iliya Timofeyev, Co-Advisor

Dr. Bernhard Bodmann

Prof. Ricardo Azevedo

Dean, College of Natural Sciences and Mathematics

Acknowledgements

I am extremely grateful to my advisor, Dr.Azencott, for his constant support, motivation, and guidance during my past 5 years as a graduate student in mathematics department at University of Houston. He always encouraged me, especially during the lesser productive weeks of research and gave me sound advice and ample meeting time to keep up the momentum of research.

My fascination with mathematics started when I was in middle school. Spending time working out additional problems, helping out my friends understand material was never a chore. I did my undergrad from one of the premier universities of India, where I got to learn basics of advanced mathematics under some exceptional professors who were very passionate about mathematics. After my undergrad following the advice of one of my professors Dr.Abha Dev Habib, I got an opportunity to pursue graduate studies at University of Houston through Mathematical Sciences Foundation (MSF). Here I met Drs. Amber Habib, Sanjeev Agrawal, Dinesh Singh, Geetha Venkatraman, and Radhamohan who further inspired me to continue my higher education in mathematics research. I owe a big debt of gratitude to all of them for inspiring and preparing me for life at UH.

I was introduced to new areas of mathematics at UH and all the professors were always there to help and never failed to go the extra mile.

I would also like to express thanks to all my incredible friends in Houston who have provided perfect balance to these long years of research and made this phase of my life memorable.

And last, but not least, I want to mention the unwavering support of my parents Sunil and Aruna Aggarwal. They continuously provided me motivation at every step through all the ups and downs of my research curve. I would like to make a special mention of my brother Akhil for keeping me sane with his amazing sense of humor.

This dissertation is dedicated to my grandmother Mrs. Nirmala Rani.

LARGE DEVIATIONS APPROACH FOR STOCHASTIC GENETIC EVOLUTION

An Abstract of a Dissertation
Presented to
the Faculty of the Department of Mathematics
University of Houston

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

By
Aanchal Aggarwal
August 2013

Abstract

Theoretical ecologists have long strived to explain how the persistence of populations depends on biotic and abiotic factors and have proposed various models to predict the long time behavior of biological populations. We are interested in modeling the effects of natural selection and adaptation in a bacterial population of *Escherichia coli*, one of the most intensively studied organisms on Earth.

A distinctive signature of living systems is Darwinian evolution, that is, a tendency to generate as well as self-select individual diversity. Mathematical models built to describe this natural dynamics of populations must be rooted in the microscopic, stochastic description of discrete individuals characterized by one or several adaptive traits and interacting with each other. The simplest models assume asexual reproduction and haploid genetics, where an offspring usually inherits the trait values of her progenitor, except when a mutation causes the offspring to take a mutation step to new and different trait values and selection follows from ecological interactions among individuals.

In this dissertation we borrow results from large deviation theory to predict the most likely evolutionary trajectories for genetic traits in a given bacterial population leading from known initial multi-species frequencies to terminal domination by mutants with highest fitness. To compute the most likely evolution path, we seek the trajectory with minimal large deviations cost among all genetic evolution trajectories. The goal thus reached is to compute the most likely evolutionary steps which brought an actually observed terminal overwhelming dominance by a new mutant.

Contents

List of Figures

List of Tables

CHAPTER 0

Introduction

0.1 Broad outline

Large deviation theory deals with the decay rates for probabilities of increasingly unlikely events and is intimately related to the calculus of variations and Hamilton-Jacobi equations, as became clear after Wentzell and Freidlin's work on small diffusion [123, 124]. This dissertation applies large deviation theory to predict the most likely evolutionary trajectories for genetic traits in a given bacterial population leading from known initial multi-species frequencies to terminal domination by mutants with highest fitness.

Often used probabilistic growth models for asexual bacterial populations with roughly fixed sizes involve competing genotypes with various “fitnesses” controlling their respective growth rates. Such stochastic models are natural to analyze laboratory experiments studying the genetic evolution of bacteria *Escherichia coli*. In many of these experiments, periodic dilutions alternate with free growth periods of roughly constant duration.

Due to large cell populations sizes N ranging between 10^5 and 10^8, large deviations approximations are a natural tool to evaluate the probabilities of arbitrary population trajectories in the space of population histograms. This approach leads to computing the minimal “large deviations” cost among all these genetic evolution trajectories.

The goal of this dissertation is to actually compute the most likely evolutionary steps which brought an actually observed terminal overwhelming dominance by a new mutant.

We formalize and study a stochastic model involving successive cycles where a deterministic growth phase with random mutations is followed by the random selection of a population sub-sample of a fixed size N.

For this stochastic process and for large population size N, we fix the number of competing genotypes g and we introduce the convex space $HIST \subset R^g$ of all population histograms. We then apply large deviation concepts and tools to compute the Rate Functional $RF(tr) \geq 0$ associated to any evolution trajectory $tr = \{H_1 H_2 H_3 ... H_n\}$ where $H_j \in HIST$ for all j.

An equation linking H_n, H_{n+1}, H_{n+2} and characterizing the trajectories tr which are "local" minimizers for the large deviations cost $RF(tr)$ is then developed.

We apply this equation in reverse time to develop a reverse shooting algorithm dedicated to the computation of most likely trajectories starting from any given initial histogram H_{init} and ending with the fixation of a target histogram H_{tar}.

We implement this algorithm numerically to compute the most likely population trajectories to population fixation in the context of populations involving 4 genotypes, using concrete *E. coli* bacterial evolution models with parameters derived from the experimental results of T. Cooper (UH Biology) [129].

We also present a numerical application of these approaches to the notions of repeatability of evolution and clonal interference in a biological context analyzed by R. Azevedo et al. [94].

0.2 Detailed Outline of Chapters in the Dissertation

We introduce techniques of large deviation theory to predict and estimate probabilities of evolutionary trajectories for a large class of stochastic population dynamics.

Chapter 1 outlines the biological background of our work.

Chapter 2 introduces our Locked Box Model for stochastic population dynamics. We consider "cell" populations where the number g of competing potential genotypes is fixed. Every "day", the current population (of roughly constant size N) undergoes a deterministic growth phase driven by fixed growth factors F_j specific to each genotype G_j. The logarithmic growth factor of a genotype is directly linked to its fitness coefficient.

These growth phases stop when the population reaches a saturation size $N_{sat} >> N$, and then the saturated population undergoes multiple random Poisson mutations, occurring independently and at fixed mutation rates.

Then the selection process is implemented by random selection of a sub-sample of size N, which becomes the initial population for the next daily cycle.

We introduce the simplex $HIST \subset [0,1]^g$ of all possible population histograms, and we compute the transition probabilities of the Markov chain $\{H_1, H_2, ...H_n\}$ of successive population histograms.

In chapter 3 we introduce general concepts and a few basic results from large deviation theory. We present a survey of the main published applications of large deviation to stochastic population dynamics.

In chapter 4 we present, for N large, an explicitly detailed construction of the large deviation rate functionals for the one-step transition probabilities $P(H_{n+1} = G \mid H_n = H)$ of the Markov chain modeling our stochastic processes in the space $HIST$ of population histograms. We then compute the full rate functional $RF(tr)$ for arbitrary random trajectories

$$tr = \{H_1, H_2, \dots H_n\}$$

dynamically evolving in the space of histograms.

The residual terms in the large deviation approximations for logarithms of transition probabilities are also computed.

In chapter 5, we develop an adequate fast numerical approximation of the one step rate functional for transition probabilities which was derived in chapter 4. To achieve this, we need to derive first-order approximations for the optimal number of mutations which minimize the one-step transition rate functional. Theoretical and numerical evidence for local convexity of the one-step rate functional is also provided to reinforce the validity of the one-step numerical approximations just mentioned.

Chapter 6 develops explicit optimality conditions for rate minimizing evolutionary trajectories. These optimality conditions involve essentially an explicit equation linking any three successive steps H_n, H_{n+1}, H_{n+2} of any trajectory tr locally minimizing the large deviation "cost" $RF(tr)$.

We fix the target (i.e. terminal) histogram H_{tar} in our trajectory tr and use multiple penultimate histograms to define multiple reverse shooting directions, and generate in reverse time the corresponding multiple rate minimizing trajectories.

Here one needs to carefully analyze the various "near boundary" situations involving histograms having one or several coordinates equal to 0 or to 1. We develop an iterative algorithm to compute the trajectory with minimal cost starting from an initial histogram H_{init} and ending with a terminal histogram H_{tar}.

In chapter 7, we outline our numerical implementation for the reverse construction of cost minimizing trajectories outlined in chapter 6. We present a stage by stage implementation of our multi-stage algorithm generating rate minimizing trajectories.

We also develop a multi-scale approach to accelerate the preceding numerical implementation, by introducing successively finer discretization of the histogram space $HIST$.

Chapter 8 illustrates numerically and graphically the implementation of our techniques on concrete examples of genetic evolution processes. For the cases of 2, 3, and 4 genotypes, we present step by step construction of multi-stage rate minimizing trajectories leading to the fixation of arbitrary given genotypes, and list out the associated optimized costs $RF(tr_{opt})$.

In chapter 9, we present direct simulation of these evolutionary stochastic processes. Using the same parameters as in chapter 8 we generate probabilities of observing rare events in the population evolution. These empirical frequency results are displayed for 3 and for 4 genotypes as observed in simulation with 10^4 trajectories.

In chapter 10 our large deviations approach is applied to an evolutionary population model studied by R. Azevedo et al. [94], to analyze the effects of clonal interference, and the repeatability of evolution. We export our framework to this context and thus predict explicitly the most likely trajectories that lead the population to fixation.

Conclusions and future work are presented in chapter 11.

CHAPTER 1

Cell Populations : Genetic Evolution Models

1.1 Modeling Genetic Evolution

The first systematic presentation of evolution was put forth by the French scientist Jean Baptiste de Lamarck (1774 – 1829) in 1809. Lamarck described a mechanism known as "the inheritance of acquired characteristics" by which he believed evolution could occur.

However, Darwin's theory of evolution [28] by natural selection is one of the best substantiated theories in the history of science, supported by evidence from a wide variety of scientific disciplines, including paleontology, genetics, developmental biology and geology. Darwin coined the term natural selection to describe the process by which organisms with favorable variations survive and reproduce at a higher rate. An inherited variation that increases an organism's chance of survival in a particular environment is called an adaptation. According to Darwin an adaptation could spread throughout the entire species over many generations and evolution by natural selection would occur.

The physical and behavioral changes that make natural selection possible are mainly due to random mutations at the level of DNA and genes. Such changes are called "mutations". Mutations can be caused by chemical or radiation damage or errors in DNA replication. Mutations can even be deliberately induced in order to adapt to a rapidly changing environment. Most times, mutations are either harmful or neutral but in rare instances, a mutation might prove beneficial to the organism. If so, it will become more and more prevalent in the next generations and spread throughout the population. Thus, natural selection guides the evolutionary process, preserving and adding up the beneficial mutations and rejecting the bad ones.

Emergence and subsequent spread of beneficial mutations through natural selection leads to adaptive evolution. Various stochastic models have been developed and studied for this process which demonstrate the importance of stochastic events in evolving populations (Haldane (1927) [57]; Fisher (1930) [48]).

So a combination of "chance and necessity" governs the outcome of evolution (Monod (1971) [87]). Occurrence of a mutation in a particular individual at a certain time during the evolution process and its survival so as to be passed on to next generation characterizes the chance variable in the system. Necessity arises mainly through the action of natural selection and mutational biases. One of the main goals of evolutionary genetics is to understand the interplay between these factors and successfully predict evolving mutational trajectories taken by a population.

If a mutant with high selective advantage appears in a population, then it has a positive probability of survival, however large the population may be (Kimura (1983) [73]). A genetic advantage of one organism over its competitors that causes it to be favored in survival and reproduction rates over time is defined to be its selective advantage. Haldane (1927) [57] proved that for a constant sized population, the probability of a mutation with selective advantage s to survive random changes in allele frequency due to random sampling is approximately $2s$. These changes in allele frequency in a population due to random sampling are termed "Genetic Drift".

In population genetics, it is assumed that selection acts on individuals based on their phenotype and these phenotypes are determined by the individual's genotype. Thus, considering fitness as a property of an individual or as a property of a genotype is not an issue in population genetics (Rice (1961) [101]). The ability of an individual to both survive and reproduce in an environment is characterized as its fitness.

In a large asexual population, beneficial mutations compete with each other for fixation. In population genetics, fixation is the change in a gene pool from a situation where there exists at least two variants of a particular gene (allele) to a situation where only one of the alleles remains. Recently Wilke (2004) [125] showed that with the increase in population size, the rate of substitution approaches a constant which is equal to the mean effect of new beneficial mutations. He also shows that the mean effect of new beneficial mutations is smaller than the mean effect of new deleterious mutations and that the mean effect of fixed mutations grows logarithmically with the population size. Wilke derives a formula evaluating whether at a given population size, the beneficial mutations are expected to compete with each other or go to fixation.

Some of the main conclusions of the experimental work by Gerrish and Lenski in 1998 [51] include

1. The probability of fixation of a given beneficial mutation decreases with both population size and mutation rate.

2. As population size or mutation rate increase, adaptive substitutions result in larger fitness increases.

3. The rate of adaptation is an increasing, but decelerating, function of both population size and mutation rate.

4. Beneficial mutations that become transiently common but do not achieve fixation because of interfering beneficial mutations are relatively abundant.

More recent work using large population has shown that beneficial mutations can be very common in this setup (Joseph and Hall (2004) [67]; Desai et al.(2007) [33]; Desai and Fisher (2007) [32]). In such a case, many new mutations will occur before any of the mutation can fix, so there will be many different mutant lineages in the population concurrently. In asexual populations, these different mutant lineages interfere and not all can fix simultaneously (Perfeito et al. (2007) [96]; Gresham et al. (2008) [55]; Kao and Sherlock (2008) [68]). Work of Visser and Rozen (2008) [106] and Desai and Fisher (2007) [32], for instance, analyzes the dynamics of such multiple mutations and the interplay between multiple mutations and interference between clones.

"Clonal Interference" is the process by which different beneficial mutations arise in an asexual population and have to compete with one another for fixation (Gerrish and Lenski (1998) [51]). It can significantly affect the evolution dynamics ([90], [51], [107], [125], [70], [33], [95], [17], [16], [65]).

Beneficial mutations that occur in different lineages may be recombined into a single lineage in sexual populations (Peters and Otto (2003) [97]). However, in asexual populations, the clones that carry such alternative beneficial mutations compete with one another, and interfere with the expected progression of a given mutation to fixation. The idea that beneficial mutations must compete in asexual populations was originally proposed by Muller in 1932 [90]. Clonal interference is thus the phenomenon whereby the fate of the beneficial mutation is altered by the appearance of a superior alternative mutation (Atwood et al. (1951) [2]; Helling et al. (1987) [60]; Visser et al. (1999) [1]).

Such competition between beneficial mutations slows the spread of and may even eliminate the first mutation. Asexual populations adapt to their environment by the occurrence and subsequent rise in frequency of the beneficial mutations. Clonal interference ensures that those beneficial mutations that do achieve fixation are of large effect. So it is more likely for the population to follow trajectories beginning with mutations of large effect, since these can out-compete other mutations, even if they do not occur first [94].

Similar selective selection of beneficial mutations was also observed by Rozen et al. [107]. They confirmed that many beneficial mutations, mostly those of small effect are lost either due to (1) genetic drift or due to (2) competition among clones carrying different beneficial mutations, a phenomenon called the Hill-Robertson effect for sexual populations and clonal interference for asexual populations. Together, these two phenomena suggest that only those beneficial mutations of large fitness effect achieve fixation. This prediction was confirmed both empirically and theoretically by showing that fitness effects of fixed beneficial mutations follow a distribution whose mode is positive.

Beneficial mutations also play an essential role in bacterial adaptation. One prominent example of bacterial adaptation is antibiotic resistance. Tenaillon et al. [102] documented the selection and fixation of resistant mutations in populations of *Escherichia coli* that had never been exposed to antibiotics but instead evolved for 2000 generations at high temperature ($42.2°C$). They show that it is not always true that antibiotic resistance is selected by the presence of antibiotics because resistant mutations confer fitness costs in antibiotic free environments. They describe the resistance mutations that are not necessarily costly in the absence of antibiotics or compensatory mutations but are highly beneficial at high temperature and low glucose. Their fitness effects depend on the environment and the genetic background, providing glimpses into the prevalence of epistasis and pleiotropy. In genetics, epistasis is a phenomenon in which the expression of one gene depends on the presence of one or more 'modifier genes' whereas pleiotropy is the phenomenon in which a single gene contributes to multiple phenotypic traits.

Multiple models of evolutionary dynamics have been proposed and extensively studied. Imhof and his team [50] approached the dynamics in a finite population under the assumption of strong selection and weak mutations using game theory. They generalized the Moran process (Moran [88]; Ewens [45]) to include frequency dependent selection and mutation. In a classic Moran process mutations are not allowed and selection is constant. Game theory provides a means to study frequency dependent selection, where the fitness of a phenotype depends on the composition of population. Deterministic differential equations have been typically used to model game dynamics in evolutionary biology to describe the evolution of very large populations.

One of the most fundamental questions in population biology concerns the persistence of species and populations, or conversely their risk of extinction. Extinction risk is influenced by a myriad of factors, including interaction between species traits and various stochastic processes leading to fluctuations and declines in population size [82, 100, 61].

Conducting a population viability analysis involves the steps of choosing an appropriate model, fitting the model to data, and using the fitted model to predict the extinction risk. In 2004 Bahi and Michel [7] developed a new class of gene evolution models in which the nucleotide mutations are time dependent. These models allowed them to study nonlinear gene evolution by accelerating or decelerating the mutation rates at different evolutionary times.

VanderSluis et al. [43] studied the effects of duplicate genes on evolutionary trajectories. They show that duplicate gene pairs are highly imbalanced in their number of genetic interactions with other genes, a pattern which appears to result from asymmetric evolution, such that one duplicate evolves or degrades faster than the other and often becomes functionally or conditionally specialized. These differences in genetic interactions are predictive of differences in several other evolutionary and physiological properties of duplicate pairs.

Méléard and Roelly (2013) [84] modeled the effects of natural selection and adaptation for a multi-cell population. They model a two-level population dynamics, resulting from the interplay between individuals submitted to mutation and competition for resources and their composition in multi-type cells. They focus on the behavior of the individual and cell populations on the long time scale of evolution where phenotypic mutations can be fixed. By rigorously constructing the underlying mathematical model and proving its existence they obtain moment and martingale properties which are the key points to deducing approximations for large individual and cell populations.

Ovaskainen and Meerson [93] in their recent paper on stochastic models of population extinction have shown how predicted extinction risk depends on the structure and parameters of a stochastic population model.

Sylvie Méléard in her other paper with Villemonais [85] has presented the quasi-stationarity properties of models derived from ecology and population dynamics. The long time behavior of different stochastic population size processes when 0 is an absorbing point and is almost surely attained by the process is discussed.

1.2 Models and Experiments for *Escherichia coli*

Our primary focus is on models of evolutionary dynamics for simple bacterium populations. Bacterial populations seem ideal for studying beneficial mutations because these populations have large sizes and short generation times. Bacterial populations propagated in the laboratory over a relatively short period of time can undergo very large numbers of replications.

Herron and Doebeli [62] documented the genetic basis and the evolutionary dynamics of adaptive diversification in three replicate evolution experiments, in which competition for two carbon sources caused initially isogenic populations of the bacterium *Escherichia coli* to diversify into two coexisting ecotypes representing different physiological adaptations in the central carbohydrate metabolism. This process closely corresponds to the evolutionary dynamics seen in mathematical models of adaptive diversification due to frequency-dependent ecological interactions.

To justify diversity-stability hypothesis which states that ecosystem diversity is positively correlated with stability, Imhof [64] used diverged *Escherichia coli* cells and showed that the fitness of community members depends on the complexity (number of participants) of the system and concluded that system complexity provides a buffer against stochastic effects. For this study stability was defined as non stochastic and reproducible population dynamics.

Fong et al. [49] conducted laboratory experiments on *E. coli* K-12 $MG1655$ grown on either lactate or glycerol minimal media to address fundamental questions about adaptation to selection pressures. They investigated the reproducibility of

growth phenotypes and global gene expression states during adaptive evolution. The results from parallel, replicate adaptive evolution experiments showed that
(1) growth phenotypes at the endpoint of evolution are convergent and reproducible;
(2) endpoints of evolution have different underlying gene expression states;
(3) the evolutionary gene expression response involves a large number of compensatory expression changes and a smaller number of adaptively beneficial expression changes common across evolution strains.

The impact of stochasticity on gene expression is widely discussed. Whether stochastic gene expression is detrimental or beneficial is, however, still unclear. Hoek and Hogeweg [120] studied the effects of stochasticity from an evolutionary point of view in the lac operon of *E. coli*, using a detailed, quantitative model evolving through a mutation-selection process. The lac operon is a cluster of genes required for the transport and metabolism of lactose in *Escherichia coli*. They concluded that in a natural environment, the impact of stochastic gene expression on lac operon evolution is minor, but that this evolution responds with much stronger stochasticity when confronted with artificial inducers. By showing that high stochasticity increases the delay in lactose uptake in a variable environment, they prove that in this particular system stochasticity is detrimental.

The most popular and longest experiment on *E. coli* is being conducted by Lenski and his team. Experimental populations of *E. coli* bacteria have evolved for $20,000$ generations in a uniform environment. Twelve populations of these were founded in 1988 from a common ancestor (Lenski et al.(1991) [78] and Barrick and Lenski (2009) [8]).

These populations have evolved under the uniform environment with glucose as the density limiting resource (Lenski et al. (1991) [78]; Lenski et al. (1994) [79]; Lenski et al. (2000) [24]) that also contains citrate, which *E. coli* cannot use as a carbon source under toxic conditions. The populations adapted to this environment by the substitution of spontaneous beneficial mutations. One particularly significant adaption was the evolution of a strain of *E. coli* that was able to use citric acid as a carbon source in an aerobic environment [11].

1.3 Experimental Design: *Escherichia coli* Evolution Experiments

In this section we present the general setup of the experiments conducted by Tim Cooper's laboratory at the UH department of Biology and HK (Hegreness et al. (2006) [59]). We borrow the description of this experimental setup from the Ph.D. thesis of V.Sehgal [112].

$L - arabinose(Ara^{-})$ is mainly used as a culture medium in most of the experiments and the strain of *E. coli* is considered to be strictly asexual. An Ara^{+} mutant was isolated from this strain (Lenski (1988) [77]). The Ara^{-} and the Ara^{+} colonies form red and white colonies, respectively (Levin et al. (1977) [80] and Lenski et al. (1991) [78]) on the indicator medium. The arabinose marker has been shown to be effectively neutral under the culture conditions used in the present series of experiments (Lenski (1988) [77]). In HK experiments, yellow and cyan fluorescent protein markers were used. In both the experiments, the color markers used were neutral, having no detectable effect on the individual.

Both the TC (Tim Cooper et al. [129]) experiments, as well as the HK experiments, start with a number K of culture wells of replicate populations, each well containing an initial population of *E. coli* cells of size N

In each one of the first six TC experiments, the K replicate populations start with a distinct initial population composed of cells having identical genotypes.

One of the six initial genotypes is $Ara - 1$, the common ancestor of the 5 other initial genotypes. In the HK experiments, all populations were started from a single, different genotype $MC4100$ (Hegreness et al. (2006) [59]). In each well, the initial population was composed of a single genotype, except that half of the cells were of one marker type and the other half were of another. In these experiments, an arabinose marker was used (Lenski et al. (1991) [78]).

In this experimental stochastic population evolution, the populations of bacteria *Escherichia coli* evolve over generations with daily growth + dilution cycles. The number of mutations is assumed to have a Poisson distribution dependent on the size of the population. Stochastic distributions, such as the Poisson distribution, for mutants in each generation, allow for random fluctuations in population sizes. The whole evolutionary process is explained below in detailed steps.

1.3.1 Daily Growth

At the beginning of each daily growth period, the initial numbers of red and white cells in each culture well are equal to $N = 2$, where N is the size of the population in each one of the wells at the beginning. After the nutrients of a well have been exhausted, (which occurs after approximately 8 to 12 hours) the population growth stops. N_{sat} is the daily terminal cell population size in each well after nutrient exhaustion and N_{sat} is essentially fixed in this experimental context. Thus, during the daily growth period, number of cells in each one of the wells increases from the initial value N to N_{sat}, at the end of the daily growth period.

1.3.2 Daily Dilution

Every 24 hours, once the population in each well has reached size N_{sat}, a subpopulation of approximate fixed size N is randomly extracted from each one of the wells. There is an effective dilution of the culture by a fixed dilution factor, $D = N_{\text{sat}}/N$. The extracted cells are then transferred to a new well, containing fresh growth medium. This transfer step is repeated daily for all the K wells.

1.3.3 Estimating Marker Frequencies for Empirical Data

Once the daily “growth + dilution” cycle has been completed, and after transfer of the diluted population to a new well, another small random sample is extracted from the N cells in the new well. This complementary sub sampling is dedicated to daily estimation of color markers frequencies. These complementary samples of sizes ranging between 300 and 400 are extracted from each one of the new K populations, and transferred onto K culture plates, where the cells are allowed to grow again. On each cell plate, after a few days, the complementary subsample of 300 to 400 cells, can be inspected by a laboratory technician who determines by visual counting the frequencies of red and white cells.

1.3.4 Time Series Recording of the Experiment

The daily estimated color marker frequencies are recorded and indexed by the acquisition date t, encoded as the number of days since the start of the experiment.

Table 1.1 shows the structural design parameters for TC and HK experiments.

Parameter Name	Parameter	TC Experiment	HK Experiment
Number of Wells	K	11	72
Initial Size of Population	N	5×10^4	2.5×10^5
Saturation Size of Population	N_{sat}	10^7	8.25×10^8
Dilution Factor	D	200	3300

Table 1.1: Structural Design Parameters for TC and HK Experiments

In the HK experiments, the daily frequencies of the two cell marker types were recorded by direct fluorimetric measurements, which generate more accurate evaluations of the daily red and white cell frequencies.

The mutation rate is the average rate at which a given cell of one genotype may mutate into another genotype, per unit of time. The bacterial evolution experiments described above were designed and carried out to estimate the mutation rate and selective advantage of the newly arising beneficial mutations. A mutant arises with selective advantage s, and thus has an advantage of $(1+s)$ relative to the progenitor cell. The multiplicative growth factor per time interval is then given by F^{1+s} where F is the multiplicative growth factor per time interval of progenitor cells given by $F = N_{\text{sat}}/N = D$.

As described in detail in the joint paper of Tim Cooper with Sehgal et al. [129], the fitness of the evolved clone was calculated relative to the ancestor as the ratio of each strain's Malthusian parameter, estimated as $\log(f_{\text{sat}}/f_{\text{init}})$, where f_{sat} and f_{init} are the final and initial frequencies of one cell type, respectively. In other words the fitness of a genotype, g is defined as

$$Fit(g) = \log F_g = \log F^{1+s}.$$

The selective advantage of gene g over other ancestral gene anc is then given by $(Fit(g) - Fit(anc))/Fit(anc)$.

The selective advantage is essentially the basis for evolution by natural selection. It is the characteristic of an organism that enables it to survive and reproduce better than other organisms in a given environment.

New algorithms were developed in [129] to estimate these parameters directly from the observed red and white daily frequency data. These estimated mutation rates and selective advantages will be used in several of the numerical examples analyzed further on in the thesis

During population growth, and the daily "growth + dilution" cycles, mutant genotypes with various selective advantages $s > 0$ can occur with a small probability μ at each cell division. All individuals are asexual and, therefore, when a fitter genotype reaches "fixation" in a population, it will drive the ancestral genotype to extinction, and thus will eliminate it from the population. This will simultaneously cause the fixation of the color marker type in which it occurred. Because adaptive mutations can arise at many sites in the genome, it is usually impossible to experimentally follow their simultaneous dynamics directly. For this reason, one of the goal of these experiments is to infer the underlying genotypic dynamics from the changes in the frequencies of the two marker types.

For numerical results, we concentrate primarily on parameters estimated for the above described experiments carried out by T.C. [129] (hereafter will be called TC experiments), but also include a previously described experiment (Hegreness et al., [59], hereafter will be called HK experiment) as a point of reference and to demonstrate the effect of a realistic range of the experimental parameters on the application of our model. Reliable and robust estimates of these parameters can be found in [129].

CHAPTER 2

Stochastic Population Dynamics and Genotypic Composition

We consider models for the genetic evolution of asexual bacterial populations such as *E. Coli*, in contexts similar to the experiments of T. Cooper [129] and Hegreness [59], where populations undergo successive daily cycles of deterministic growth, random mutations, and random subsampling as formalized below.

2.1 Notations and Definitions

We introduce the following definitions and biological parameters for our model.

- The genotype is defined to be the genetic "signature" of an organism. It determines the hereditary potentials and limitations of an individual. We assume the presence of g distinct genotypes in our stochastic population model.

- H_n is defined to be the $g-$dimensional vector of population histogram at the beginning of day n where $H_n(j)$ is the frequency of genotype j in the population such that $\sum_{j=1}^{g} H_n(j) = 1$.

- $HIST$ is the space of all possible population histograms H with g distinct genotypes such that $\sum_{j=1}^{g} H(j) = 1$.

- Population trajectories spanning T generations can be viewed as discrete random Markov sequences $h = (H_n)_{0 \leq n \leq T}$ of population histograms $H_n \in HIST$. A sequence $X_1, X_2, ...$ of random variates is called Markov [69] if, for any n

$$\text{Prob}(X_n = x_n | X_1 = x_1, ..., X_{n-1} = x_{n-1}) = \text{Prob}(X_n = x_n | X_{n-1} = x_{n-1}).$$

- Let N be the fixed initial population size at the beginning of each daily growth period and N_{sat} be the saturation size reached by the population at the end of each growth period, just before random sampling.

- Define $\vec{m} = m(j),\ j = 1, ..., g$ to be the matrix of mutation rates in the model. The mutation rate $m(j)$ is the average rate at which a given cell may mutate into genotype j, per unit of time. Also mutation rates are of the order of 10^{-6} in our setup.

- Gene structure changes induced by single random mutations transform genotype j into genotype i where $i \neq j$. We also assume mutations to be independent, i.e., each individual cell may mutates at the end of one growth period. Real mutations in $E.\ Coli$ bacteria may actually occur only when a cell splits into two new cells, which happens essentially in continuous time throughout a daily growth period. Here we simplify these occurrence dates by assuming that they occur simultaneously at the end of the daily growth period. The mathematical analysis of this simplification is presented in [129], where one shows that the impact of this simplification is minor for large values of N.

- A mutant arises with selective advantage s, and thus has an advantage of $(1+s)$ relative to the progenitor cell. This means that the multiplicative growth factor per time interval is then given by F^{1+s} where F is the multiplicative growth factor per time interval of progenitor cells given by $F = N_{\text{sat}}/N = D$.

- Fitness defines the ability of an individual to both survive and reproduce in an environment and strongly influences its contribution to the gene pool in the next generation. The distribution of fitness reflects the selection coefficient relative to the fittest alleles in the population.

- Fitness of the evolved clone is calculated relative to the ancestor as the ratio of each strain's Malthusian parameter, estimated as $\log(f_{\text{sat}}/f_{\text{init}})$, where f_{sat} and f_{init} are the final and initial frequencies of one cell type, respectively [129]. In other words the fitness of a genotype, g is defined as

$$Fit(g) = \log F_g = \log F^{1+s}.$$

The typical value of growth factor in the TC experiments [129] on *E. coli* populations is about 200.

- For a population of size N with growth factors $F = (F_j)$ and population histogram H_n at the beginning of day n, the population size at the end of the nth daily growth period is $Size_n = N\langle F, H_n \rangle$, where $\langle , \rangle$ is the usual scalar product in $\mathbb{R}^g$.

2.2 Deterministic Daily Growth Periods

Locked Box models : competing genotypes in fixed size populations

For cell populations the main focus is on successive population samples pop_n of *fixed size* N extracted by periodic dilutions alternating with free growth periods of fixed duration τ. We simulate and analyze the evolution of finite sets Γ of species or cell genotypes competing within a fixed size population.

At generation n, Locked Box dynamics start with a population pop_n of N cells $C_i\, ; 1 \leq i \leq N$ with respective genotypes $j(i)$ and deterministic growth factor $F_{j(i)}$.

We name this model as a Locked Box model to refer to that fact the population systems being studied here are assumed to be isolated in nature. This means we do now allow and emigrants or immigrants from neighboring population system into the population of interest.

Step 1: Growth.

During one pure deterministic growth step with no mutations, pop_n becomes a population Q and the number of cells with genotype j grows from $NH_n(j)$ to $NH_n(j)F_j$. Then, we have size of Q as $Size_n = N\langle F, H_n\rangle$. The frequency of cells with genotype j in the intermediary population Q is then $G_n(j) = F_j H_n(j)/\langle F, H_n\rangle$.

Hence a purely deterministic growth step transforms the current histogram H_n into a new histogram $G_n = \Phi(H_n)$, where the deterministic function $\Phi : HIST \to HIST$ is defined on the convex set $HIST$ by

$$\Phi_j(H) = F_j H(j)/ < F, H >, \quad \text{for all } \; H \in HIST. \tag{2.1}$$

This step is equivalent to the "Daily Growth" step described in 1.3.1 for TC experiments.

Step 2: Mutations.

Next, random Poisson distributed independent mutations, governed by the vector $\vec{m} = (m_j)$ of mutation rates, are implemented in the intermediary population Q to generate a (random) population POP$_{int}$. The global mutation rate $\bar{m} = \sum_{j=1}^{g} m_j$, is assumed to be very small. After these random mutations, and just before the next periodic dilution, the histogram of genotypes frequencies becomes a random histogram MG_n. This step is sometimes also called selection. We will study MG_n further below.

Step 3 : Random Sampling.

To generate the new population pop_{n+1}, one performs a "random sampling" of the population POP$_{int}$ by extracting from population POP$_{int}$ a random sample of size

N. This random sample is then called the population pop_{n+1}, and H_{n+1} is the genotypic histogram of pop_{n+1}. Given H_n, the random vector NH_{n+1} has then a *multinomial* distribution $Mult(N, MG_n) = \mu(N)$ with parameters N and MG_n. In long term laboratory experiments to study genotypic evolution of bacteria, this random sampling is implemented concretely by population dilutions at the end of daily growth periods.

This step is equivalent to “Daily Dilutions”, as described in 1.3.1 for TC experiments.

Hence the Markov chain dynamics $H_n \rightarrow H_{n+1}$ is implemented in 3 successive steps.

- The Growth step is a deterministic transformation $H_n \rightarrow G_n = \Phi(H_n)$.
- The Mutation step is a random perturbation $G_n \rightarrow MG_n$ of G_n by Poisson mutations, and the conditional expected value of MG_n is very close to G_n.
- The Selection step is the extraction of a random sample of fixed large size N which transforms MG_n into a random histogram H_{n+1} such that the conditional distribution of H_{n+1} given H_n is the multinomial distribution $Mult(N, MG_n)$.

2.3 Random Mutations Model

After changing the time unit, we can assume that the fixed duration of free growth periods is $\tau = 1$. For instance in laboratory experiments [129], τ is essentially equal to 1 day. In our numerical implementations, the number of genotypes will be inferior or equal to 4 to facilitate computations. But the theory developed below is valid for arbitrary numbers of genotypes.

At the beginning of day n, the population histogram is given by $h_n = (H_n(j))_{0 \leq j \leq g}$ so that $\sum_j^g H_n(j) = 1$.

After the nth free growth period for the population with growth factor $F = (F_j)$ the number of individuals with genotype j becomes $F_j H_n(j) N$, giving rise to an intermediary population of size

$$\mathrm{Size_n} = \mathrm{N}\langle \mathrm{F}, \mathrm{H_n} \rangle. \tag{2.2}$$

Now, let

$$S_n(j) = N F_j H_n(j) \tag{2.3}$$

denote the number of individuals of genotype j at the end of growth period of day n.

Next, random Poisson distributed independent mutations, governed by the vector $m = (m_j)$ of mutation rates, are implemented in the intermediary population, Q. The global mutation rate $\bar{m} = \sum_j m_j$, is assumed to be very small.

Alternatively, we also work with situation where only forward mutations are accepted (i.e.), mutants can only evolve into fitter genotype.

Thus, we focus on the following two cases separately :

1. Non-restricted mutations - Cells are allowed to mutate into cells of any other genotype right at the moment of their births. Only mutations which result in a change of genotype are considered here.
2. Restricted mutations - Cells are only allowed to mutate into cells with genotypes of higher fitness.

2.3.1 Non-restricted Mutations

Fix any time n. Recall that we consider the population existing at the end of the nth growth period and the number of individuals of genotype j present in this population is given by $F_j H_n(j) N$.

Let $X_{j,k}$ denote the number of mutants of genotype j changing into genotype k at time step n. We naturally assume that all the $X_{j,k}$ are independent and we impose $X_{j,j} = 0$. Due to the Poisson distribution assumption, we have [46]

$$\text{Prob}(X_{j,k} = R_{j,k}) = \exp(-m_k S_n(j)) \frac{(m_k S_n(j))^{R_{j,k}}}{R_{j,k}!}, \quad \forall j \neq k. \tag{2.4}$$

This implies

$$E(X_{j,k}) = m_k S_n(j), \quad \forall j \neq k. \tag{2.5}$$

This defines the new composition of the population after mutations. The new number of individuals of genotype j is given by

$$NF_jH_n(j) - \sum_{k=1}^{g} X_{j,k} + \sum_{k=1}^{g} X_{k,j}.$$

For the sub-population of genotype j, let the respective number of emigrants and immigrants be,

$$O_n(j) = \sum_{k=1}^{g} X_{j,k},$$

$$I_n(j) = \sum_{k=1}^{g} X_{k,j},$$

where $X_{j,j}$ is defined to be 0. Define the $g \times g$ matrix,

$$R = [R_{j,k}], \quad 1 \leq j, k \leq g.$$

where we systematically impose that the diagonal terms $R_{j,j}$ should be equal to zero. For instance, a 3×3 non restricted mutation matrix, R can be visualized as below

$$R = \begin{bmatrix} 0 & r_{1,2} & r_{1,3} \\ r_{2,1} & 0 & r_{2,3} \\ r_{3,1} & r_{3,2} & 0 \end{bmatrix}$$

Then for our case with g genotypes, for $j = k$ for instance, we see that

$$O_n(k) = \sum_{i|i \neq k} X_{k,i}$$

has a Poisson distribution with mean,

$$\mathrm{mean(O_n(k)) = (\sum_{i|i \neq k} m_i)S_n(k).}$$

Similarly, $I_n(k)$ will have a Poisson distribution with mean,

$$\mathrm{mean(I_n(k))} = \sum_{\mathrm{i|i\neq k}} \mathrm{m_i S_n(i)}.$$

Now, after mutations have been allowed the new population M_{int} has g groups of sizes $U_n(j)$, where

$$U_n = S_n - O_n + I_n.$$

Lemma 2.3.1. *The size of population* POP_{int} *is equal to* $\mathrm{Size_n}$.

Proof :

$$\begin{aligned}
\mathrm{Size(POP_{int})} &= \sum_j U_n(j), \\
&= \sum_j (S_n(j) - O_n(j) + I_n(j)) \\
&= \sum_j S_n(j) - \sum_j O_n(j) + \sum_j I_n(j), \\
&= \mathrm{Size_n} - \sum_{\mathrm{j}} \sum_{\mathrm{k|k\neq j}} \mathrm{X_{j,k}} + \sum_{\mathrm{j}} \sum_{\mathrm{k|k\neq j}} \mathrm{X_{k,j}}, \\
&= \mathrm{Size_n}.
\end{aligned}$$

Corollary 2.3.2. *Thus the histogram MG_n of the population after growth and mutations is given by*

$$MG_n = U_n/\mathrm{Size_n}.$$

Lemma 2.3.3. *The conditional expectation of MG_n given H_n is given by*

$$E[MG_n \mid H_n] = (1 - \bar{m})\Phi(H_n) + m. \tag{2.6}$$

Proof : The number of individuals of genotype j after mutations is $U_n(j)$, where

$$U_n = S_n - O_n + I_n.$$

So, the expected number of individuals is,

$$\begin{aligned}
E[M_{int}(j) = U_n(j)|H_n = H] &= S_n(j) - E[O_n(j)|H_n = H] + E[I_n(j)|H_n = H], \\
&= S_n(j) - \sum_{k|k\neq j} E[X_{j,k}|H_n = H] + \sum_{k|k\neq j} E[X_{k,j}|H_n = H], \\
&= S_n(j) - S_n(j)\sum_k m_k + m_j \sum_k S_n(k), \\
&= S_n(j)(1-\bar{m}) + m_j \mathrm{Size_n}.
\end{aligned}$$

This implies that the conditional expectation of histogram MG_n given H_n is given by:

$$\begin{aligned}
E\left[MG_n(j)|H_n = H\right] &= \frac{S_n(j)}{\mathrm{Size_n}}(1-\bar{m}) + m_j, \\
&= (1-\bar{m})\Phi_j(H) + m_j
\end{aligned}$$

Thus,

$$E[MG_n \mid H_n] = (1-\bar{m})\Phi(H) + m.$$

Definition 2.3.4. *Call $X = [X_{j,k}]$ the random matrix of mutants at time step n . Fix any matrix of integers $R = [R_{j,k}]$ such that $R_{j,j} = 0$ and $R_{j,k} \geq 0$. Given $X = R$, the random histogram MG_n is the vector $p(R,H)$ defined by*

$$p(R,H)(j) = \frac{1}{\mathrm{Size_n}}[NF_jH(j) - \sum_{k|k\neq j} R_{j,k} + \sum_{k|k\neq j} R_{k,j}] \quad , \quad j = 1,2,...,g$$

2.3.2 Restricted Mutations

In reference to the restricted mutations discussed before, if we only allow mutations into genotypes with higher fitness and not into lower fitness genotypes, then

$$X_{j,k} = 0 \ \forall j \geq k$$

where we reorder the genotypes so that $F_1 < F_2 < ... < F_g$, where F_j is the growth factor for genotype j. For instance, a 3×3 restricted mutation matrix, R can be visualized as below

$$R = \begin{bmatrix} 0 & r_{1,2} & r_{1,3} \\ 0 & 0 & r_{2,3} \\ 0 & 0 & 0 \end{bmatrix}$$

Note that this does not alter any results regarding expectations and histograms proved before for general mutations. The only major difference is that now our mutation matrix is strictly upper triangular in nature. So, the vector $p(R, H)$ defined by

$$p(R,H)(j) = \frac{1}{\mathrm{Size_n}} \left[NF_jH(j) - \sum_{k|k>j} R_{j,k} + \sum_{k|k<j} R_{k,j} \right] \quad , \quad \forall j = 1, 2, ..., g$$

2.4 Random Sampling Model

Now that we have an intermediary population with random mutations, to generate the population at time $n+1$, we extract a random sample pop_{n+1} of size N from the population POP_{int}.

Let $V_n(j)$ be the number of individuals of genotype j present in population, pop_{n+1} at time $n+1$ after selection by random sampling. Then

$$\sum_j V_n(j) = N,$$

and the new population histogram becomes $H_{n+1}(j) = V_n(j)/N$.

Given $X = R$, the random vector V_n has a multinomial distribution $\mu(N)$ with g occurrences, which are given by the g coordinates of the vector $p(R, H)$.

Thus to generate the new population we have one deterministic step and two probabilistic steps. The concatenated probability of this sequence of 3 steps is given by the transition probability :

$$\begin{aligned}\Theta(H, G) &= \mathrm{Prob(H_{n+1} = G|H_n = H)}\\ &= \sum_R \prod_{j,k|j\neq k} \mathrm{Prob(X_{j,k} = R_{j,k}|H_n = H).\mu(N)[G]}. \qquad (2.7)\end{aligned}$$

In the case of g genotypes with general non restricted mutations the matrix R depends only on $g^2 - g$ parameters which are the off diagonal elements $R_{j,k}$, with j different of k. For restricted mutations R depends only on $(g^2 - g)/2$ parameters which are the upper triangular elements $R_{j,k}$, with $j < k$.

These elements have integer values verifying

$$0 \leq X_{j,k}, \;\; \text{and} \;\; \sum_{\mathrm{k}} \mathrm{X_{j,k} \leq S_n(j)}.$$

However, for $R_{j,k}$ we have

$$0 \leq R_{j,k} < \infty$$

Let

$$Q(R,H,N) = \prod_{j,k|j\neq k} \mathrm{Prob}(\mathrm{X_{j,k} = R_{j,k}|H_n = H}).$$

Thus, we have that

$$\begin{aligned} Q(R,H,N) &= \prod_{j,k|j\neq k} \mathrm{Prob}(\mathrm{X_{j,k} = R_{j,k}|H_n = H, X_{j,k} \leq S_n(j)}), \\ &= \prod_{j,k|j\neq k} \mathrm{Prob}(\mathrm{X_{j,k} = R_{j,k}|H_n = H, X_{j,k} \leq NF_jH_n(j)}), \\ &= \prod_{j,k|j\neq k} \mathrm{Prob}(\mathrm{X_{j,k} = R_{j,k}|X_{j,k} \leq NF_jH_j}). \end{aligned}$$

Now,

$$\begin{aligned} \mathrm{Prob}(\mathrm{X_{j,k} = R_{j,k}|X_{j,k} \leq NF_jH_j}) &= \frac{\mathrm{Prob}(\mathrm{X_{j,k} = R_{j,k}, X_{j,k} \leq NF_jH_j})}{\mathrm{Prob}(\mathrm{X_{j,k} \leq NF_jH_j})}, \\ &= \frac{\mathrm{Prob}(\mathrm{X_{j,k} = R_{j,k}, X_{j,k} \leq NF_jH_j})}{1 - \mathrm{Prob}(\mathrm{X_{j,k} > NF_jH_j})}. \end{aligned}$$

Recall here that $X_{j,k}$ is Poisson distributed with mean $m_kNF_jH_j$. Now in our computations, rate of mutations, m is assumed to be 10^{-6}, so $E[X_{j,k}] = 10^{-6} \times NF_jH_j$. We will show that

$$1 - \mathrm{Prob}(\mathrm{X_{j,k} > NF_jH_j}) \simeq 1$$

using lemma 2.4.1.

Lemma 2.4.1. *If X is a Poisson random variate with mean λ, then for $a > 0$*

$$\mathrm{Prob[X > a\lambda] < e^{\lambda(e^t - 1 - at)}},\ \forall \mathrm{t} > 0.$$

Proof: We know for $a > 0$

$$\mathrm{Prob[X > a\lambda] = Prob[e^{tX} > e^{ta\lambda}]},\ \forall \mathrm{t} > 0, \tag{2.8}$$

Now $e^{tX} > 0$, so using Markov inequality [46], we get for a positive random variable, $Z = e^{tX}$

$$\mathrm{Prob[Z \geq b] \leq \frac{E[Z]}{b}} \tag{2.9}$$

or

$$\mathrm{Prob[e^{tX} \geq b] \leq \frac{E[e^{tX}]}{b}} \tag{2.10}$$

where $b = e^{ta\lambda} > 0$. Hence,

$$\mathrm{Prob[X > a\lambda] = Prob[e^{tX} > e^{ta\lambda}] < \frac{E[e^{tX}]}{e^{ta\lambda}}.} \tag{2.11}$$

Now we know X has Poisson distribution, so $\mathrm{E[e^{tx}]}$ is the moment generating function given by [46]

$$\mathrm{E[e^{tx}] = e^{\lambda(e^t - 1)}}.$$

Substituting this into equation 2.11 we get our required result

$$\mathrm{Prob[X > a\lambda] < e^{\lambda(e^t - 1 - at)},\ \forall t > 0.} \tag{2.12}$$

Using lemma 2.4.1, $\mathrm{Prob(X_{j,k} > NF_jH_j)}$ is analyzed below. Assume $X = X_{j,k} \neq 0$, then $\lambda = m_k N F_j H_j$ and $a\lambda = NF_jH_j = 10^6\lambda$. This gives

$$\mathrm{Prob(X_{j,k} > NF_jH_j) < e^{\lambda(e^t - 1 - 10^6 t)},\ \forall t > 0.} \tag{2.13}$$

Now, $\lambda(e^t - 1 - 10^6 t)$ is a function which attains its minimum value at $t = \log 10^6$. Hence

$$\mathrm{Prob(X_{j,k} > NF_jH_j) < e^{\lambda(10^6 - 1 - 10^6 \log 10^6)}.} \tag{2.14}$$

or

$$\begin{aligned}\mathrm{Prob(X_{j,k} > NF_jH_j)} &< e^{10^6\lambda(-10^{-6}-(13.82)+1)},\\ &= e^{10^6\lambda(-10^{-6}-12.82)},\\ &= e^{NF_jH_j(-10^{-6}-12.82)}.\end{aligned}$$

If $H_j = 0$ then $X_{j,k} = 0$ since $X_{j,k}$ is number of mutants of genotype k emerging from genotype j and thus $\mathrm{Prob(X_{j,k} > NF_jH_j)} = 0$.

If instead $H_j \neq 0$, then minimum value of NF_jH_j is at least of the order of F since $H_j \geq 1/N$ and for our experiments we have assumed least growth factor value to be 200. So,

$$\mathrm{Prob(X_{j,k} > NF_jH_j)} < e^{200(-10^{-6}-12.82)} \simeq 0. \tag{2.15}$$

In other words,

$$1 - \mathrm{Prob(X_{j,k} > NF_jH_j)} \simeq 1$$

and hence

$$\begin{aligned}\mathrm{Prob(X_{j,k} = R_{j,k}|X_{j,k} \leq NF_jH_j)} &= \mathrm{Prob(X_{j,k} = R_{j,k}, X_{j,k} \leq NF_jH_j)} &(2.16)\\ &= \mathrm{Prob(X_{j,k} = R_{j,k})}. &(2.17)\end{aligned}$$

So,

$$Q(R, H, N) = \prod_{j,k|j\neq k} \mathrm{Prob(X_{j,k} = R_{j,k})} \tag{2.18}$$

The sum in the expression for probability $\Theta(H, G)$ in 2.7 denotes multiple sums, one over each non zero coordinate of R. Then the expression for probability $\Theta(H, G)$

in 2.7 simplifies to

$$\mathrm{Prob}(\mathrm{H}_{\mathrm{n}+1} = \mathrm{G}|\mathrm{H}_{\mathrm{n}} = \mathrm{H}) = \Theta(\mathrm{H}, \mathrm{G}) = \sum_{\mathrm{R}} \mathrm{Q}(\mathrm{R}, \mathrm{H}, \mathrm{N}).\mu(\mathrm{N})[\mathrm{G}]$$

where $Q(R, H, N)$ is given by equation 2.18.

CHAPTER 3

Large Deviation Approach to Stochastic Population Evolution

3.1 General Large Deviations Principles

Large deviation theory focuses on the asymptotic behavior of remote tails of sequences of probability distributions. The theory deals with the rates of decay of rare events probabilities as some natural parameter in the problem is allowed to vary.

The large deviation theory has its origin in the work of Boltzmann who brought probability ideas into thermodynamic theory in his effort to characterize energy and density fluctuations in physical systems. The first large deviations formula (in dimension 1) due to Cramér (1938) was widely extended to infinite dimensional vector spaces and trajectory spaces by the papers and books of Donsker and Varadhan [38, 39], Wentzell and Freidlin [123, 124], Bahadur and Zabell [6], Azencott [4, 3], Dembo and Zeitouni [29], Dupuis and Wang [40] and many others, to generate a new domain of probability theory.

The theory behind large deviations has been explored recently in detail by Varadhan [121], Azencott [4] in their respective books.

Let X_n be sequence of independent and identically distributed random variables(r.v.) having with mean 0 and finite second moment σ^2. Then, $\bar{X}_n = \frac{X_1+...+X_n}{n}$ tends almost surely to 0 as $n \to \infty$ by the law of large numbers. On the other hand, $\bar{Z}_n = \frac{X_1+...+X_n}{\sigma\sqrt{n}}$ has a limiting normal distribution according to the Central Limit Theorem. In particular

$$P(|\bar{X}_n| \geq \delta) \to 0, \quad n \to \infty, \tag{3.1}$$

and, for any interval A,

$$P(|\bar{Z}_n| \in A) \to \frac{1}{2\pi} \int_A e^{\frac{-x^2}{2\sigma^2}} \, \mathrm{d}x, \quad n \to \infty, \tag{3.2}$$

therefore

$$\frac{1}{n} \log P(|\bar{X}_n| \geq \delta) \to \frac{-\delta^2}{2\sigma^2}, \quad n \to \infty. \tag{3.3}$$

Expressions like the one derived in 3.3 present the type of probability estimates which are of primary interest in large deviation theory. Extensions to random variables taking values in infinite dimensional vector spaces have been studied in detail and presented below.

The *large deviation principle* (LDP) essentially characterizes the limiting behavior as $N \to \infty$, of a family of probability measures P_N in terms of a rate function.

Let Ω be a complete separable metric space, and P_N a family of probability measures on the Borel subsets of Ω.

Definition 3.1.1. *One says that $\{P_N\}$ obeys the LDP with a rate function $\lambda(\cdot)$ if there exists a function $\lambda(\cdot)$ from Ω into $[0, +\infty]$ satisfying*

(i) $0 \leq \lambda(\phi) \leq \infty \forall \phi \in \Omega$.

(ii) $\lambda(\cdot)$ is lower semi-continuous.

(iii) For each $L < \infty$, the set $\{\phi : \lambda(\phi) \leq L\}$ is a compact set in Ω.

(iv) For each closed set $F \subset \Omega$

$$\limsup_{N\to\infty} \frac{1}{N} \log P_N(F) \leq - \inf_{\phi \in F} \lambda(\phi).$$

(v) For each open set $A \subset \Omega$

$$\liminf_{N\to\infty} \frac{1}{N} \log P_N(A) \geq - \inf_{\phi \in A} \lambda(\phi).$$

For any borel set $A \subset \Omega$ define

$$\Lambda(A) = \inf_{\phi \in A} \lambda(\phi).$$

Whenever A verifies the conditions

$$\Lambda(A^o) = \Lambda(A) = \Lambda(\bar{A})$$

where A^o is interior and $\bar{A}$ is closure of A.

Then we have

$$\lim_{N\to\infty} \frac{1}{N} \log P_N(A) = -\inf_{\phi\in A} \lambda(\phi).$$

Definition 3.1.2. Cramér Transform and Legendre Duality on $\mathbb{R}$. *Let θ be a probability on $\mathbb{R}$. Let $\hat{\theta} : \mathbb{R} \to (0, +\infty]$ be its Laplace transform, defined by $\hat{\theta}(t) = \int_{\mathbb{R}} e^{tx} d\theta(x)$. Define the Cramèr transform $\lambda : \mathbb{R} \to [0, +\infty]$ of the measure θ by*

$$\lambda(x) = \sup_{t\in\mathbb{R}} [tx - \log \hat{\theta}(t)] \tag{3.4}$$

Then the function $\lambda(x)$ is convex and lower semi-continuous for $x \in \mathbb{R}$.

Theorem 3.1.3. Cramér Chernoff Theorem on $\mathbb{R}$. (Cramér [27], Chernoff [23]) *Let X_n be a sequence of independent real valued r.v. with the same probability distribution θ, and let $\bar{X}_n = (X1 + ... + Xn)/n$. Let λ be the Cramér transform of θ. Assume also that $\int |x| d\theta(x)$ is finite, and let $m = \int x d\theta(x)$. Then for all $a \in \mathbb{R}$,*

$$\lim_{n\to\infty} \frac{1}{n} \log P(|\bar{X}_n| \leq a) = -\lambda(a),\ a \leq m \tag{3.5}$$

$$\lim_{n\to\infty} \frac{1}{n} \log P(|\bar{X}_n| \leq a) = 0,\ m < a \tag{3.6}$$

$$\lim_{n\to\infty} \frac{1}{n} \log P(|\bar{X}_n| \geq a) = 0,\ a < m \tag{3.7}$$

$$\lim_{n\to\infty} \frac{1}{n} \log P(|\bar{X}_n| \geq a) = -\lambda(a),\ m \leq a. \tag{3.8}$$

So the sequence X_n satisfies the LDP with a rate function $\lambda(\cdot)$ given by 3.4.

The proof for above theorem can be found in most texts on this subject such as Varadhan [121] and Azencott [4].

We now present a result which characterizes the rate functional for empirical distributions in general.

Let Γ be a Polish topological space. Let π and ν be arbitrary probability measures on the Borel $\sigma-$algebra $B(\Gamma)$. Polish topological spaces are complete, metric spaces with a countable dense subset.

If ν is absolutely continuous with respect to π, the *Kullback information* $I_\pi(\nu)$ of ν with respect to π (see [4]; [76]) is defined by,

$$I_\pi(\nu) = \int_\Gamma \frac{d\nu}{d\pi}(x) \log\left(\frac{d\nu}{d\pi}(x)\right) d\pi(x). \tag{3.9}$$

and we set $I_\pi(\nu) = +\infty$ when ν is not absolutely continuous with respect to π. In particular, $I_\pi(\nu) = 0$ if and only if $\nu = \pi$. Recall that the non-negative function $I_\pi(\nu)$ is also called the relative entropy of ν with respect to π. Further information on properties of $I_\pi(\nu)$ can be found in [76].

Let $E = M(\Gamma)$ be the Frechet topological vector space of bounded Borel measures on Γ, endowed with the tight convergence topology. The random Dirac masses $X_n = \delta_{Y_n}$ can be viewed as independent random vectors with values in E having the same probability distribution μ. The probability μ is defined on $B(E)$ and its support is included in the convex set $M^1(\Gamma) \subset E$ of all probabilities on Γ [108].

Theorem 3.1.4. (see [38], [6], [4] and Sanov [111])
Let π be a probability on the Borel subsets of a Polish topological space Γ. Let Y_n be a sequence of independent random variables taking values in Γ, and having the same probability distribution π. For any $\nu \in M^1(\Gamma)$, let $I_\pi(\nu)$ be the Kullback information of ν with respect to π. Then Cramér transform λ of the probability μ has compact

level sets and is given by

$$\lambda(\nu) = I_\pi(\nu) \text{ for } \nu \in M^1(\Gamma) \text{ and } \lambda(\nu) = +\infty \text{ for } \nu \in [E - M^1(\Gamma)] \qquad (3.10)$$

For empirical distributions like multinomial distribution (random dilution in our case) the rate functional then is given by *Kullback information* function as described above. We derive the exact formulation for this rate functional in our set up in chapter 4. Similarly, the derivation of rate functional for Poisson probability (random mutations) is shown in detail in chapter 4.

Let $b(y) \in \mathbb{R}^n$ be a locally Lipschitz vector field defined for all $y \in \mathbb{R}^n$. One associates to b the dynamic system (D)

$$\frac{dy_t}{dt} = b(y_t), \quad (D).$$

Let $X_t : \Omega \to \mathbb{R}^k$ be a continuous Gaussian process defined on the time interval $J = [0, 1]$. The probability distribution of the trajectories of X_t over the time interval J is a Gaussian probability μ on that path space $C(\mathbb{R}^k)$. Let $\tilde{\lambda} : C(\mathbb{R}^k) \to [0, \infty]$ be the Cramér transform of the Gaussian measure μ.

Then for each small $\epsilon > 0$, consider the following stochastic dynamic system SDE^ϵ, where the random noise perturbing the underlying deterministic dynamic system D is modeled by X_t

$$\frac{dY_t}{dt} = b(Y_t) + \epsilon\sigma(Y_t)X_t, \quad (SDE^\epsilon)$$

with deterministic initial condition $Y_0 = x$ for some fixed $x \in \mathbb{R}^k$.

Our goal here is to analyze the behavior of the probability distribution of the trajectories of Y_t when $\epsilon \to 0$. More precisely, we will evaluate the probability $P^\epsilon(A)$ that the trajectories of $Y_t,\ 0 \leq t \leq 1$ belong to any given set A of paths starting at x but which are not solutions of the limit deterministic dynamic system (D).

For small ϵ, these events A are naturally "rare events" for the process Y_t and for "nice" sets A, we will have a large deviations principle, namely the probabilities $P^\epsilon(A)$ will tend to 0 with ϵ , at exponential speeds of order of $\exp(-\Lambda(A)/\epsilon^2)$.

Let Y_t^ϵ be the solution of the perturbed dynamic system SDE^ϵ. starting from x at time 0. Denote by $\mathcal{E}_x(\mathbb{R}^n)$ the space of possibly exploding paths starting at x and defined for $t \in J$.

Let $Y_t^\epsilon : \Omega \to \mathcal{E}_x(\mathbb{R}^n)$ be the random variable defined by the trajectories of Y_t^ϵ on time interval J. For any $g \in \mathcal{E}_x(\mathbb{R}^n)$, define the (possibly empty) set $B_x^{-1}(g)$ of all $f \in C(\mathbb{R}^k)$ such that g is the maximal solution on J of the differential equation

$$g'_t = b(g_t) + \epsilon\sigma(g_t)f_t \text{ with } \mathrm{g}_0 = \mathrm{x}.$$

For the perturbed dynamic system SDE^ϵ in 4.1, we then define the rate function $\lambda : \mathcal{E}_x(\mathbb{R}^n) \to [0, +\infty]$ by

$$\lambda(g) = \inf\{\tilde{\lambda}(f) | f \in B_g^{-1}\}.$$

On arbitrary subsets A of $\mathcal{E}_x(\mathbb{R}^n)$ define the "rate functional" $\Lambda(A)$ by

$$\Lambda(A) = \inf_{g \in A} \lambda(g).$$

This is the extension of general Wentzell-Freidlin theory [123] and more details along with proofs for existence of this rate functional can be found in the book by Azencott [4].

So essentially, large deviations theory focuses on computing

$$-\Lambda(A) = \lim_{N\to\infty} \frac{1}{N} \log[\mathrm{Prob}_N(A)]$$

via a rate functional $\lambda(\phi)$ where A is a rare event and ϕ is any trajectory of the process. Typically

$$\Lambda(A) = \inf_{\phi} \lambda(\phi)$$

where ϕ is a trajectory which realizes the event A, and in most cases this infimum is realized by a cost minimizing trajectory ϕ_m.

This broadly outlines our approach for estimating rate functionals (costs) of rare trajectories in the thesis for classes of population evolution processes.

3.2 Large Deviations Applied to Stochastic Population Dynamics

Numerous applications of large deviations have been explored in the context of Markov stochastic processes. For instance in economics, Noah Williams (2008) [126] in his paper on small noise asymptotics for a stochastic model derives a functional central limit theorem, a large deviation principle, and a moderate deviation principle. These are used to calculate analytically the asymptotic distribution of the capital stock, and to obtain bounds on the probability that the log of the capital stock will differ from its deterministic steady state level by a given amount. This latter result can be applied to characterize the probability and frequency of large business cycles.

Budhiraja and Ghosh [15] in 2005 studied the problem of asymptotically optimal control of a well known multi-class queuing network, referred to as the "criss-cross network", in heavy traffic. They consider exponential inter-arrival and service times, linear holding cost and an infinite horizon discounted cost criterion. Using the path wise solution of the Brownian control problem, they present an elementary and transparent treatment of the problem using large deviation ideas and obtain an asymptotically optimal scheduling policy which is of threshold type.

Tailleur and Lecomte [117] in 2008 used large deviation principles in thermodynamics. For the last ten years, physicists have been interested in large deviation functions mainly because they are good candidates to extend the concept of thermodynamic potentials to out of equilibrium situations and to dynamical observables. In 2011 Smith [115] used LDP to construct entropy functions that both express large deviations scaling of fluctuations, and describe system environment interactions, for discrete stochastic processes either at or away from equilibrium.

In 2009 Bresslof [12] analyzed a master equation formulation of stochastic neurodynamics for a network of synaptically coupled homogeneous neuronal populations. They showed how the path integral approach can be used to study large deviation or rare event statistics underlying escape from the basin of attraction of a stable fixed point of the mean-field dynamics in neural networks.

Liu [81] presented a new framework for finding the optimal transition paths of metastable stochastic chemical kinetic systems with large system size. The optimal transition paths were identified to be the most probable paths according to the large deviation theory of stochastic processes. Applications to gene regulatory networks such as the toggle switch model and the Lactose Operon Model in *Escherichia coli* are presented as numerical examples.

In the context of stochastic models for population evolution, large deviation theory has also been used to study evolutionary models with mutations and selection in the specific biological framework of adaptive dynamics.

An unpublished manuscript of Darden (1983) has large deviation results for the Wright-Fisher model with two alleles and heterotic selection. Morrow and Sawyer [89] derived large deviation results for a class of Markov chains arising from population genetics. They used Wright-Fisher model where average effect of forces such as selection and mutation are much stronger than effects due to finite population size. The equilibrium probability for the process to be found away from fixed point and amount of time required by the process to escape a fixed neighborhood of a fixed point is computed using large deviation principles. We are studying a similar model and would like to estimate probabilities for rare events under large population limit.

A stochastic Lotka-Volterra model was analyzed by Klebaner and Liptser [75] in 2001. They approach the problem of extinction via the theory of large deviations. The large deviation principle is established, and consequently used to obtain bounds for the asymptotics of the time to extinction of the prey population.

Demetrius, Gundlach and Ochs [30] studied complexity and demographic stability in population models. They invoked the large deviation theory to derive a fluctuation theorem in the system. This theorem says that the rate of fluctuations around a steady state is positively correlated with entropy, a point which is then used to predict correlations between ecological constraints and evolutionary trends.

A. Cercueil and O. Franq́is [18] described new models in population genetics that extend the neutral Wright-Fisher model by including strong selection and mutation. Fixation times are studied in the limit of small mutation rates within the framework of Markov chains with rare transitions. These results use the formalism of large deviations. The main result outlines the role of the discrete geometry of the fitness landscape and provides a mean for estimating the expected number of generations for an individual with better fitness value to appear.

Johansson and Sumpter [66] also calculate evolutionary stable strategies, extinction probabilities using large deviations for a wide range of site based ecological models. For these models local interactions between individuals are assumed to take place at a finite number of discrete resource sites over non-overlapping generations and individuals between generations move randomly between sites over entire system.

In 2005 N. Champagnat [19] proved a convergence result of the $m$icroscopic [35, 83] model of evolution to the adaptive dynamics trait substitution sequence model when the parameters are normalized in a non-standard way, leading to a time scales separation. Under the large population asymptotic, and small mutations asymptotic he proved the occurrence of time scale separation between the birth and death events and the mutation events. The proof uses large deviation results on branching processes and logistic Markov birth and death processes.

Metz et al. [86] have introduced an asymptotic of rare mutations to approximate the process of adaptive evolution with a monomorphic (composed of individuals holding the same trait value) jump process. The jump process describes evolutionary trajectories as trait substitution sequences developing over the timescale of mutations. Dieckmann and Law [34] have further achieved a deterministic approximation for the jump process as a solution to the so called canonical equation of adaptive dynamics. Metz et al.'s notion of trait substitution sequences and Dieckmann and Law's canonical equation form the core of the current theory of adaptive dynamics.

Dynamics for finite populations with strong selection and weak mutations were studied by Fudenberg et.al. [50]. They implement game theory to study frequency-based selection, where fitness of a phenotype depends on composition of the population. The long run behavior of the process with mutations is related to a simpler process with no mutations using large deviations. They provided a characterization of the asymptotic behavior of the absorption probabilities as the population size go to infinity.

Méléard, Jabin and Champagnat [20] described adaptation in a stochastic multi-resources chemostat model. All the traits with zero density at equilibrium are proved to actually go extinct after a time of the same order as the logarithm of the population size. Also it is proved that the exit time from a neighborhood of the equilibrium grows as an exponential of the population size using classical results from large deviation estimates.

Viet Chi Tran [119] studied a continuous-time discrete population structured by a vector of ages where individuals reproduce asexually, age and die. It is shown that in a large population limit, the microscopic process converges to the measure-valued solution of an equation that generalizes the McKendrick-Von Foerster and Gurtin-McCamy PDEs in demography and the large deviations associated with this convergence are studied.

With Ferriére and Méléard [22], Champagnat published results for unifying evolutionary dynamics from individual stochastic processes to macroscopic models. The issue of evolutionary dynamics drifting away from trajectories predicted by the canonical equation is investigated by considering the asymptotic of the probability of 'rare events' for the sample paths of the discussion. Martingale and large deviation theories are used as the probabilistic tools for deriving and unifying models of evolutionary dynamics from stochastic nonlinear processes operating at the individual level. On a timescale of very rare mutations, they establish rigorously the models of trait substitution sequences and their approximation known as the canonical equation of adaptive dynamics.

Champagnat and Méléard [21] also published results for polymorphic evolution sequence and evolutionary branching. They use adaptive dynamics based on the biologically motivated assumptions of rare mutations and large population. It is proved that such a microscopic process describing ecological dynamics can be approximated by a Markov pure jump process on the set of point measures on the trait space. They examine the asymptotic behavior of the microscopic process when the population size grows to infinity as well as the mutation rate converges to 0, in a long time scale.

We are also interested in probabilities of 'rare events' in evolutionary dynamics with assumption of rare mutations and large population. We assume the process to be a Markov chain as described in chapter 2 and use reverse shooting technique from numerical analysis to predict the trajectory and associated rate functional for desired evolutionary events.

3.3 Importance Sampling in the Context of Large Deviations

Another significant application is large deviations analysis to derive an effective strategy for importance sampling, a domain covered by an extensive literature.

Importance sampling is a variance reduction technique that has been applied successfully to the problem of estimating the probabilities of rare events. A guiding principle in the efficient estimation of rare event probabilities by Monte Carlo is that importance sampling based on the change of measure suggested by a large deviations analysis can reduce variance by many orders of magnitude.

Siegmund [113] showed that the uniquely optimal exponential change of measure for estimating a gambler's ruin probability is determined by the exponential rate of decay of the probability as one of the boundaries recedes.

As observed by Glasserman and Wang [54] in 1997, the subsequent literature can be roughly divided in two: results showing that specific estimators have provably good performance, and the development of estimators, often evaluated experimentally, suggested by, but not strictly supported by, rare-event asymptotics.

Consider a probability space Ω of the space of trajectories ω of a random process, endowed with a probability P_N, . Then consider a fixed event, $A \subset \Omega$ which satisfies either a logarithmic limit

1.
$$\lim_{N\to\infty} \frac{1}{N} \log P_N(A) = -\gamma$$

for some $\gamma > 0$, or the stronger exponential asymptotic

2.
$$P_N(A) \sim Ce^{-\gamma N},$$

for some constant $C > 0$. To estimate $\alpha(N) \triangleq P_N(A)$, straightforward simulation generates 'r' independent realizations of the process trajectories $\omega_1, ..., \omega_r$.

The standard estimator of $P_N(A)$ is the empirical frequency q_r observed for the realization of the event A among these r simulated trajectories. The variance of this estimator is then $\alpha - \alpha^2/r$. If $P_N(A) \to 0$ then this variance approaches 0. The relative error of the estimator (the ratio of its standard deviation to its mean) then satisfies

$$\text{relative error} = \frac{\sqrt{\alpha(N) - \alpha^2(N)}}{\sqrt{r}\alpha(N)} \geq \frac{1}{\sqrt{r\alpha(N)}} \to \infty.$$

If (2) is true, then

$$\text{relative error} \geq \frac{\sqrt{Ce^{\gamma N}}}{\sqrt{r}}$$

and the increase is observed to be exponential and the number of independent simulations trajectories ω_i required to achieve a fixed relative error grows exponentially in N.

Importance sampling generates samples under a different measure $\bar{P}_N$ with expectation operator $\bar{E}$ and uses the representation

(3)

$$P_N(A) = E[L_N 1_{A_N}],$$

in which L_N is the likelihood ratio of $\bar{P}_N$ to P_N.

Based on (3), we obtain an unbiased estimator of $P_N(A)$ by averaging over independent replications of the random terms $L_N(\omega_j)1_{A_N}(\omega_j)$. It can then be shown [54] that the number of independent replicate trajectories ω_j required to achieve a fixed relative error for the estimation of $P_N(A)$ grows at a grows at a sub-exponential rate and remains bounded.

However, it was noted that a successful application of an importance sampling distribution based on large deviation theory critically depends on the specific problem at hand. Glasserman and Wang [54] give variations on both the level-crossing problem and the Cramér-type problem, and show that exponential twists can be inefficient if the rare event A is irregular. Similar observations have been made by Glasserman and Kou [53] in a queuing context.

It is natural to ask whether there exist any necessary and sufficient conditions for asymptotic efficiency. In cases when the Gärtner-Ellis theorem applies, this question is studied by Sadowsky and Bucklew [110] , while Sadowsky [109] extends these findings to a general abstract large deviation setting.

Dieker and Mandjes [36] have given necessary and sufficient conditions (*Varadhan* *conditions*), which are shown to improve the conditions of Sadowsky [109].

Dupuis and Wang [40], indicate that large deviation theory suggests many possible changes of measure, which are not all are suitable for importance sampling. They consider importance sampling schemes where the exponential change of measure is adaptive, in the sense that it depends on the historical empirical mean. Their results indicate that large deviations analysis truly suggests an adaptive change of measure, rather than a static change of measure.

Using the fact that the statistical functionals usually can be represented as the functionals of empirical probability measures Ermakov [42] developed a similar approach of effective importance sampling based on the theorems about the large and moderate large deviations of empirical measures ([56], [41]). The results on efficient simulation of large deviations are obtained and expressed in terms of Kullback-Leibler information. He also shows that the effective importance sampling measures are the solutions of extremal problem involving the minimization of Kullback-Leibler information numbers on specific sets.

A more comprehensive overview of these techniques for light and heavy tailed systems can be found in paper by Blanchet and Lam [10]. They review standard (state-independent) techniques that take advantage of large deviations results for the design of efficient importance sampling estimators. State dependent techniques are also discussed in detail along with examples in which they are applicable.

Another important application is in the field of finance, where questions related to extremal events play an increasingly major role. Financial applications range from

Monte-Carlo methods and importance sampling in option pricing to estimates of large portfolio losses subject to credit risk, or long term portfolio investment. Pham [98] in his lectures has explained some essential techniques in large deviations theory, and illustrated how they are applied recently for example in stochastic volatility models to compute implied volatilities near maturities.

The use of large deviation theory to efficiently implement importance sampling is a very active area of research. In future extensions of our work on large deviations approximations for genetic population evolutions, a natural next step for us is to study an implement importance sampling to estimate probabilities of rare events involving genetic evolution trajectories.

CHAPTER 4

Large Deviations for One-step Transition Probabilities

We have specified in chapter 2 a locked box stochastic model for genotypic evolution of a population submitted to successive growth periods alternating with random selections. We now introduce large deviation approximations for the transition probabilities of this stochastic dynamics for large population size N. Thus we derive large deviation rate functionals for both multinomial sampling and Poisson probability of random mutations.

4.1 Asymptotic Contexts

We distinguish at least 3 radically different situations.

(1.) **Bounded Mean Mutations**(BMM)

N tends to ∞ but $Nm_j(N)F_j(N)$ remain bounded between fixed bounds

$$0 < a < Nm_j(N)F_j(N) < A.$$

(2.) **Unbounded Mean Mutations**(UMM)

N tends to ∞ and $m_j(N)F_j(N)$ remain between fixed bounds

$$0 < b < m_j(N)F_j(N) < B$$

so that $Nm_j(N)F_j(N)$ tends to infinity at a speed proportional to N.

(3.) **Low Mean Mutations**(LMM)

N tends to ∞ and for some $c > 0$ and $d > 0$

$$Nm_j(N)F_j(N) < \frac{c}{N^d}$$

so that $Nm_j(N)F_j(N)$ tends to zero at polynomial speed.

For the two experimental setups we will consider (TC[129], HK[59]), we are essentially in the context UMM and thus will focus only on this case. In fact we have the values for rate of mutations (m) and growth factor (F) to be constant and thus independent of N. Hence all our limits have been studied in the UMM case. Also, we isolate the cases of restricted and general non-restricted random mutations and approach them separately.

4.2 Large Deviations Approximation for Multinomial Sampling

Due to the large size N of cell populations being considered here in the biological systems, we introduce large deviations approximations for the multinomial distributions involved in the random selection step.

Let N be the fixed initial population size at the beginning of each daily growth period and N_{sat} be the saturation size reached by the population at the end of each growth period, just before random selection. the values of N and N_{sat} are given in Table 4.1.

Parameter	TC	HK
N	5×10^4	2.5×10^5
N_{sat}	10^7	8.25×10^8

Table 4.1: Parameters from TC [129] and HK[59] experiments

Call g the number of genotypes involved in the evolution model, and let U_i denote the number of cells of genotype i present in the population before random selection by multinomial sampling. Also, assume V_i to be the number of cells of genotype i chosen after the random selection. Then, we have that $\sum_{i=1}^{g} U_i = N_{\text{sat}}$ and $\sum_{i=1}^{g} V_i = N$.

Since we use the Stirling formula

$$log(n!) \simeq n \log n - n$$

for estimating factorials, we need to introduce boundary cases where Stirling approximation is not accurate since it is only valid for factorials larger than 50!.

Thus, let $p_i = U_i/N_{\text{sat}}$ and $G_i = V_i/N$, so that for boundary cases we will have

$$0 \leq G_i \leq \epsilon$$

where we systematically set $\epsilon = 50/N$. The value of ϵ in our numerical contexts of TC and HK experiments is 10^{-3} and 5×10^{-6} respectively.

We present below detailed large deviations approximations for multinomial probabilities under various boundary case assumptions.

1. Assume that all $G_i \geq \epsilon$ or in other words $V_i \geq 50$ is true for all genotypes.

   The multinomial probability for picking V_i cells of genotype i from a population with U_i cells of genotype i is expressed as [46]

   $$\mu(N) = \frac{N!}{\prod_{i=1}^{g} V_i!} \prod_{i=1}^{g} p_i^{V_i} \tag{4.1}$$

   where $p_i = U_i/N_{\text{sat}}$. Taking logarithm on both sides and dividing by population size N we get

   $$\frac{1}{N} \log \mu(N) = \frac{1}{N} \log(N!) - \sum_{i=1}^{g} \frac{1}{N} \log(V_i!) + \sum_{i=1}^{g} \frac{V_i}{N} \log p_i \tag{4.2}$$

   We now introduce Stirling formula, $log(n!) \simeq n \log n - n$ and simplify to get

   $$\frac{1}{N} \log \mu(N) \simeq \frac{1}{N}[N \log N - N] - \sum_{i=1}^{g} \frac{1}{N}[V_i \log V_i - V_i] + \sum_{i=1}^{g} \frac{V_i}{N} \log p_i \tag{4.3}$$

Thus,

$$\frac{1}{N}\log\mu(N) \simeq \frac{N}{N}\log N - \frac{N}{N} + \frac{1}{N}\sum_{i=1}^{g} V_i - \sum_{i=1}^{g}\frac{1}{N}V_i\log V_i + \sum_{i=1}^{g}\frac{V_i}{N}\log p_i \tag{4.4}$$

Since we know $\sum_{i=1}^{g} V_i = N$ and $G_i = V_i/N$ we get

$$\frac{1}{N}\log\mu(N) \simeq \sum_{i=1}^{g} G_i\log\frac{1}{G_i} + \sum_{i=1}^{g} G_i\log p_i \tag{4.5}$$

Now, take limit as $N \to \infty$ for both sides

$$\lim_{N\to\infty}\frac{1}{N}\log\mu(N) = \sum_{i=1}^{g} G_i\log\frac{p_i}{G_i} \tag{4.6}$$

or,

$$\lim_{N\to\infty}\frac{1}{N}\log(\mu(N)) = -KLD(G,p) \tag{4.7}$$

where KLD is the Kullback Leibler Divergence defined by

$$\begin{aligned} KLD(G,P) &= +\infty \quad \text{iff there is a } j \in \Gamma \text{ such that } G(j) > 0 \text{ and } P(j) = 0, \\ KLD(G,P) &= \sum_{j\,\in\,\Gamma} G(j)\log\frac{G(j)}{P(j)} \quad \text{in all other cases.} \end{aligned}$$

where Γ is the set of all genotypes and by convention, the term $0\log 0$ is defined to be 0.

2. The case where $G_i \leq \epsilon$, or equivalently $V_i \leq 50$ is true for all genotypes is not feasible since we need $\sum_{i=1}^{g} G_i = 1$. So, without loss of generality, let $G_1 \leq \epsilon$ or equivalently $V_1 \leq 50$ and $G_i \geq \epsilon$ or $V_i \geq 50$ for all other genotypes. Then we have

$$\frac{1}{N}\log\mu(N) = \frac{1}{N}\log(N!) - \sum_{i=1}^{g}\frac{1}{N}\log(V_i!) + \sum_{i=1}^{g}\frac{V_i}{N}\log p_i \tag{4.8}$$

Proceeding as above, we use Stirling approximation for all factorials except $V_1!$. We now have

$$\frac{1}{N}\log\mu(N) \simeq \frac{1}{N}[N\log N - N] - \frac{1}{N}\log(V_1!) - \sum_{i=2}^{g}\frac{1}{N}[V_i\log V_i - V_i] \qquad (4.9)$$
$$+\sum_{i=1}^{g}\frac{V_i}{N}\log p_i$$

Thus,

$$\frac{1}{N}\log\mu(N) \simeq \frac{N}{N}\log N - \frac{N}{N} - \frac{1}{N}\log(V_1!) + \frac{1}{N}\sum_{i=2}^{g}V_i - \sum_{i=2}^{g}\frac{1}{N}V_i\log V_i \qquad (4.10)$$
$$+\sum_{i=1}^{g}\frac{V_i}{N}\log p_i$$

Again using the facts that $\sum_{i=1}^{g}V_i = N$, $N = N/\langle F, H\rangle$, and $G_i = V_i/N$ we get

$$\begin{aligned}\frac{1}{N}\log\mu(N) &\simeq -\frac{1}{N}\log(V_1!) + G_1(\log N - 1) + \sum_{i=2}^{g}G_i\log\frac{1}{G_i} + \sum_{i=1}^{g}G_i\log p_i \\ &\simeq -\frac{1}{N}\log(V_1!) + \frac{V_1}{N}[\log N - 1 + \log p_1] + \sum_{i=2}^{g}G_i\log\frac{p_i}{G_i} \qquad (4.11)\end{aligned}$$

In the above expression, we analyze the sum when $V_1 \leq 50$

$$s(N) = \frac{1}{N}\log(V_1!) + \frac{V_1}{N}[1 - \log(p_1N)]$$

.

We first assume $U_1 \geq 1$. Since $p_1 = U_1/Nsat$, $p_1N = U_1\frac{N}{N_{\text{sat}}} = F_1U_1$, and so $\log(p_1N) = \log F_1 + \log U_1$.

Thus for given U_1, we have

$$\begin{aligned} s(N) &= \frac{1}{N}\log(V_1!) + \frac{V_1}{N}[1 - \log(p_1 N)], \\ &= \frac{1}{N}\log(V_1!) + \frac{V_1}{N}[1 - \log U_1 - \log F_1] \end{aligned}$$

Now, for $U_1 > 1$ clearly $\lim_{N\to\infty} s(N) = 0$.

Similarly, for $U_1 = 1$, $\log U_1 = 0$ and

$$s(N) = \frac{1}{N}\log(V_1!) + \frac{V_1}{N}[1 - \log F_1]$$

we obtain $\lim_{N\to\infty} s(N) = 0$.

Also, for the case with $N = 50000$, $g = 3$ and $p_i = G_i = V_i/N$ numerical evaluations show that maximum sum, $|\text{sum}|$ for $V1 \leq 50$ is inferior to 6×10^{-5}.

If however $U_1 = 0$, this means $p_1 = 0 \Rightarrow V_1 = 0$. Using this information in the expression for $\frac{1}{N}\log\mu(N)$ we get

$$\frac{1}{N}\log\mu(N) = \frac{1}{N}\log(N!) - \sum_{i=2}^{g}\frac{1}{N}\log(V_i!) + \sum_{i=2}^{g}\frac{V_i}{N}\log p_i \tag{4.12}$$

and using Stirling approximation gives us

$$\frac{1}{N}\log\mu(N) \simeq \sum_{i=2}^{g} G_i \log\frac{p_i}{G_i} \tag{4.13}$$

Thus the additional factorial terms are not relevant in this particular case.

Hence the rate functional for very large population, when $N \to \infty$ turns out to be

$$-\sum_{i=2}^{g} G_i \log \frac{p_i}{G_i} \tag{4.14}$$

3. Without loss of generality let $G_i \leq \epsilon$ or $V_i \leq 50$ for $i = 1, 2$ and $G_i \geq \epsilon$ or $V_i \geq 50$ for $i \geq 3$. Then similar to above case we have the final cost formula as

$$-\sum_{i=3}^{g} G_i \log \frac{p_i}{G_i} \tag{4.15}$$

Thus, for a target population with g genotypes where B is the set of genotypes that satisfy boundary condition as outlined before and $J = \Gamma - B$ is the set of genotypes away from boundary, the one-step rate functional associated to the random selection step is given by

$$\sum_{j \in J} G_j \log \frac{G_j}{p_j} \tag{4.16}$$

Note that the boundary margin $\epsilon = 50/N$ depends on the daily initial population size N and is not constant. We now derive inequalities which help us to choose and estimate the value of ϵ required for proper convergence in equation 4.11.

When $p_1 = 0$, we have already seen in equations 4.12 and 4.12 that the expression simplifies without any additional factorials. So, now we prove the following formulas for the case $p_1 \geq \frac{1}{N}$.

1.

$$\log(V!) \leq (V+1)\log(V+1) - V, \;\; \forall V \geq 1, V \in \mathbb{Z}. \tag{4.17}$$

Proof : We have that

$$\begin{aligned} (V+1)! &= (V+1)V!, \\ &\geq 2V!, \ forV \geq 1, V \in \mathbb{Z}, \\ \log((V+1)!) &\geq \log 2 + \log V!. \end{aligned}$$

Approximating the sum $\log(n!) = \sum_{j=1}^{n} \log j$ with an integral

$$\sum_{j=1}^{n} \log j \approx \int_{1}^{n} \log x \, dx = \log(n!) \approx n \log n - n + 1$$

we get

$$\begin{aligned} \log V! &\leq (V+1)\log(V+1) - (V+1) + 1 - \log 2, \\ &\leq (V+1)\log(V+1) - V + 1 - \log 2 - \log 2, \\ \log V! &\leq (V+1)\log(V+1) - V. \end{aligned}$$

2.

$$0 \leq \log(Np_i) \leq \log(N). \tag{4.18}$$

Proof : We know that $1 \leq Np_i \leq N$ and taking log on both sides gives us the required inequality, $0 \leq \log(Np_i) \leq \log(N)$.

3. Hence

$$\log(V!) + V - V\log(Np_i) \leq (V+1)\log(V+1). \tag{4.19}$$

Proof : We have shown that $\log(V!) + V - V\log(Np_i) \leq \log(V!) + V$ and also

$$\log(V!) + V - V\log(Np_i) \leq (V+1)\log(V+1).$$

From the above formulas it is clear that the sum $s(N) = \frac{1}{N}\log(V_1!) + \frac{V_1}{N}[1 - \log(p_1 N)]$ in the expression 4.11 when $V_1 \leq 50$ is always smaller than $51 \log 51/N$ for $N = 50000$ or 0.004. This upper bound is not yet small but becomes smaller as the population size increases to $N = 10^6$ and $N = 10^7$. Hence to ensure the convergence of this approximation we need to make sure that we allow for maximum value of V, $V_{max} = N\epsilon_N$ such that

$$N\epsilon_N \log(N\epsilon_N)/N \rightarrow 0.$$

Now, for the above relation to be true we need

$\epsilon_N \log(N) \rightarrow 0$ and $\epsilon_N \log(\epsilon_N) \rightarrow 0$ as $N \rightarrow \infty$

since if $\epsilon_N \log(N) \rightarrow 0$ then clearly $\epsilon_N \log(\epsilon_N) \rightarrow 0$.

Thus, in order to make sure $\epsilon_N \log(N) \rightarrow 0$ we need to pick $\epsilon_N << 1/logN$. This is true in our case as we have $N = 50000$ and so $\epsilon_N = 10^{-3} << 1/\log(50000) = 0.09$.

4.3 Large Deviations Approximation for One-step Random Mutations

4.3.1 Non-restricted Mutations

The probability expression for random Poisson mutations, $Q(R, H, N)$ actually depends on the initial population size, N, the mutation matrix, R and initial population histogram H. This expression was derived in chapter 2. So to be more precise, let

$$R_{j,k} = Nr_{j,k},$$

where $R_{j,k} = r_{j,k} = 0$ for $j = k$.

We here study the correct approximation (for large N and given $H_n = H$ and $r_{j,k}$ fixed) of the following probability

$$Q(r, H, N) = \prod_{j,k|j\neq k} \mathrm{Prob(X_{j,k} = Nr_{j,k})}$$

This is a conditional probability for $X_{j,k}$ where each $X_{j,k}$ has Poisson probability distribution. So

$$Q(r, H, N) = \prod_{j,k|j\neq k} \exp(-m_k N F_j H_j) \frac{(m_k N F_j H_j)^{Nr_{j,k}}}{[Nr_{j,k}]!}.$$

We approximate the factorials using Stirling formula.

$$\frac{1}{N} \log Q(r, H, N) = \sum_{j,k|j\neq k} \left[-m_k F_j H_j + r_{j,k}(\log N + \log(m_k F_j H_j)) - \frac{1}{N} \log((Nr_{j,k})!) \right]$$

Using Stirling formula gives

$$\begin{aligned}\frac{1}{N}\log Q(r,H,N) \simeq &\sum_{j,k|j\neq k}[-m_kF_jH_j + r_{j,k}\log(m_kF_jH_j)] + \log N\sum_{j,k|j\neq k} r_{j,k}\\ &-\frac{1}{N}\sum_{j,k|j\neq k}\left[Nr_{j,k}\log(Nr_{j,k}) - Nr_{j,k} + \frac{1}{2}\log(Nr_{j,k}) + \frac{1}{2}\log 2\pi\right],\end{aligned}$$

Taking limit as $N \to \infty$ we get,

$$\begin{aligned}L(r,H) &= \lim_{N\to\infty}\frac{-1}{N}\log Q(r,H,N)\\ &= -\sum_{j,k|j\neq k}[-m_kF_jH_j + r_{j,k}\log(m_kF_jH_j)] + \sum_{j,k|j\neq k}[r_{j,k}\log r_{j,k}] - \sum_{j,k|j\neq k} r_{j,k}\end{aligned} \tag{4.20}$$

Thus, we can write

$$Q(r,H,N) \simeq \mathrm{res}(N)\exp(-NL(r,H)).$$

where $\mathrm{res}(N)$ is equivalent to some power of $1/N$ for large N. The full derivation of $\mathrm{res}(N)$ follows in a separate section.

4.3.2 Restricted Mutations

Let

$$R_{j,k} = Nr_{j,k},$$

where $R_{j,k} = r_{j,k} = 0$ for $j \leq k$.

So we study the correct approximation (for large N and given $H_n = H$ and $r_{j,k}$ fixed) of the following probability

$$\begin{aligned} Q(r, H, N) &= \prod_{j,k|j<k} \mathrm{Prob}(\mathrm{X_{j,k} = Nr_{j,k}}) \\ &= \prod_{j,k|j<k} \exp(-m_k N F_j H(j)) \frac{(m_k N F_j H(j))^{Nr_{j,k}}}{[Nr_{j,k}]!}. \end{aligned}$$

Using Stirling formula as before and taking limit as $N \to \infty$ we get,

$$\begin{aligned} L(r, H) &= \lim_{N\to\infty} \frac{-1}{N} \log Q(r, H, N) \\ &= -\sum_{j,k|j<k} [-m_k F_j H_j + r_{j,k} \log(m_k F_j H_j)] + \sum_{j,k|j<k} [r_{j,k} \log r_{j,k}] - \sum_{j,k|j<k} r_{j,k} \end{aligned} \tag{4.21}$$

Thus, again for this case we can write

$$Q(r, H, N) \simeq \mathrm{res}(N) \exp(-NL(r, H)).$$

where $\mathrm{res}(N)$ is equivalent to some power of $1/N$ for large N. The full derivation of $\mathrm{res}(N)$ follows in a separate section.

4.4 Residual Terms in Large Deviations Approximation

4.4.1 Non-restricted Mutations

For random mutations probability we have

$$\begin{aligned}
\frac{1}{N}\log Q(r,H,N) &\simeq \sum_{j,k|j\neq k}\left[-m_kF_jH_j + r_{j,k}\log(m_kF_jH_j)\right] + \log N\sum_{j,k|j\neq k} r_{j,k}\\
&-\frac{1}{N}\sum_{j,k|j\neq k}\left[Nr_{j,k}\log(Nr_{j,k}) - Nr_{j,k} + \frac{1}{2}\log Nr_{j,k} + \frac{1}{2}\log 2\pi\right],\\
&\simeq \sum_{j,k|j\neq k}\left[-m_kF_jH_j + r_{j,k}\log(m_kF_jH_j) + r_{j,k} - r_{j,k}\log r_{j,k}\right]\\
&-\frac{1}{2N}\sum_{j,k|j\neq k}\log N - \frac{1}{2N}\sum_{j,k|j\neq k}\log r_{j,k} + \frac{1}{2N}\sum_{j,k|j\neq k}\log 2\pi,
\end{aligned}$$

Since for the case of g mutants we have $s = g^2 - g$ sums over non diagonal entries of the mutation matrix, we obtain

$$\frac{1}{N}\log Q(r,H,N) \simeq -L(r,H) - \frac{s\log N}{2N} - \frac{1}{2N}\sum_{j,k|j\neq k}\log r_{j,k} + \frac{s}{2N}\log 2\pi$$

then,

$$\frac{1}{N}\log Q(r,H,N) \simeq -L(r,H) - \eta_N$$

where

$$\eta_N = \frac{s\log N}{2N} + \frac{1}{2N}\sum_{j,k|j\neq k}\log r_{j,k} - \frac{s}{2N}\log 2\pi$$

and $\lim_{N\to\infty} \eta_N = 0$.

Thus we have

$$Q(r, H, N) \simeq \exp(-NL(r, H)) \exp(-N\eta_N)$$

and

$$\exp(-N\eta_N) = \left(\frac{2\pi}{N}\right)^{s/2} \left(\prod_{j,k|j\neq k} \exp(r_{j,k}/2)\right)^{-1}$$

Similarly, for the case one step random selections with no boundary cases we can obtain residual term in the limit as follows. We know $\sum_{i=1}^{g} V_i = N$ and $G_i = V_i/N$ and also

$$\frac{1}{N}\log\mu(N) = \frac{1}{N}\log N! - \sum_{i=1}^{g}\frac{1}{N}\log V_i! + \sum_{i=1}^{g}\frac{V_i}{N}\log p_i$$

Using Stirling formula $log(n!) \simeq n\log(n) - n + 1/2\log n + 1/2\log 2\pi$, we get

$$\begin{aligned}\frac{1}{N}\log\mu(N) \simeq{}& \frac{1}{N}[N\log N - N + \frac{1}{2}\log N + \frac{1}{2}\log 2\pi] \\ & - \sum_{i=1}^{g}\frac{1}{N}[V_i\log V_i - V_i + \frac{1}{2}\log V_i + \frac{1}{2}\log 2\pi] + \sum_{i=1}^{g}\frac{V_i}{N}\log p_i, \\ & \simeq \sum_{i=1}^{g}\frac{V_i}{N}\log\frac{N}{V_i} + \sum_{i=1}^{g}\frac{V_i}{N}\log p_i + \frac{1}{2N}(\log N - \sum_{i=1}^{g}\log V_i) - \frac{1}{N}\log 2\pi \\ & \simeq \sum_{i=1}^{g} G_i\log\frac{p_i}{G_i} + \frac{1}{2N}(\log N - \sum_{i=1}^{g}\log V_i) - \frac{1}{N}\log 2\pi\end{aligned}$$

Generalizing the above expression so as to include the cases with boundary conditions we have

$$\frac{1}{N}\log\mu(N) \simeq \sum_{i\in J} G_i\log\frac{p_i}{G_i} + \frac{1}{2N}(\log N - \sum_{i\in J}\log V_i) - \frac{a-1}{2N}\log 2\pi$$

where J is the set of genotypes away from boundary and cardinality$(J) = a$. Thus we get that

$$\frac{1}{N}\log \mu(N) \simeq -KLD - \delta_N$$

or

$$\mu(N) \simeq \exp(-N * KLD)\exp(-N\delta_N)$$

where

$$\begin{aligned}\delta_N &= -\frac{1}{2N}(\log N - \sum_{i\in J}\log V_i) + \frac{a-1}{2N}\log 2\pi \\ &= -\frac{1}{2N}(\log N - \sum_{i\in J}\log NG_i) + \frac{a-1}{2N}\log 2\pi\end{aligned}$$

and

$$KLD = \sum_{i\in J} G_i \log\frac{p_i}{G_i}.$$

Hence

$$\exp(-N\delta_N) = \frac{1}{(2\pi N)^{\frac{a-1}{2}}\prod_{i\in J}\sqrt{G_i}}$$

4.5 One-step Large Deviations Rate for Transition Probabilities

Hence, we get final probability (derived in chapter 2) as the product of Poisson and Multinomial probabilities as follows

$$\Theta(H,G) = \sum_r \exp(-N(L(r,H) + KLD(p,G)))\exp(-N(\eta_N + \delta_N)).$$

Defining $S(H,r,G) = L(r,H) + KLD(G,p)$, where $L(r,H)$ is given by the Equation 4.20 and $KLD(G,p)$ is expressed in the Equation 4.8, we obtain the final expression for the probability as follows

$$\Theta(H,G) = \sum_r \exp(-N\ S(H,r,G)) \left(\frac{(2\pi)^{\frac{s-a+1}{2}}}{N^{\frac{s+a-1}{2}}} \left(\prod_{i\in J} \sqrt{G_i} \prod_{j,k,j\neq k} \exp(r_{j,k}/2) \right)^{-1} \right). \tag{4.22}$$

where $J = \Gamma - B$ is the set of genotypes in G away from boundary, $s = g^2 - g$ and cardinality$(J) = a$.

4.5.1 Restricted Mutations

Following a similar line of argument for restricted mutations where $r_{j,k} = 0, \forall j \geq k$, we obtain

$$\frac{1}{N}\log Q(r,H,N) \simeq \sum_{j,k|j<k} [-m_k F_j H_j + r_{j,k}\log(m_k F_j H_j) + r_{j,k} - r_{j,k}\log r_{j,k}]$$
$$-\frac{1}{2N}\sum_{j,k|j<k}\log N - \frac{1}{2N}\sum_{j,k|j<k}\log r_{j,k} + \frac{1}{2N}\sum_{j,k|j<k}\log 2\pi,$$

And for the case of g mutants we have $s = (g^2 - g)/2$ sums over the possible non-zero entries of the mutation matrix and we obtain

$$\frac{1}{N}\log Q(r,H,N) \simeq -L(r,H) - \frac{s\log N}{2N} - \frac{1}{2N}\sum_{j,k|j<k}\log r_{j,k} + \frac{s}{2N}\log 2\pi$$

then

$$\frac{1}{N}\log Q(r,H,N) \simeq -L(r,H) - \eta_N$$

where

$$\eta_N = \frac{s \log N}{2N} + \frac{1}{2N} \sum_{j,k|j<k} \log r_{j,k} - \frac{s}{2N} \log 2\pi$$

and $\lim_{N\to\infty} \eta_N = 0$.

Thus we have that

$$Q(r, H, N) \simeq \exp(-NL(r, H)) \exp(-N\eta_N)$$

and

$$\exp(-N\eta_N) = \left(\frac{2\pi}{N}\right)^{s/2} \left(\prod_{j,k|j<k} \exp(r_{j,k}/2)\right)^{-1}$$

Also, for the case of one step random selections with no genotypes on boundary we can obtain the similar residual term in the limit for restricted mutations as before. Thus we get

$$\Theta(H, G) = \sum_r \exp(-N(L(r, H) + KLD)) \left(\left(\frac{2\pi}{N}\right)^{s/2} \left(\prod_{j,k|j<k} \exp(r_{j,k}/2)\right)^{-1} \frac{1}{(2\pi N)^{\frac{a-1}{2}} \prod_{i\in J} \sqrt{G_i}}\right),$$

Defining $Sr(H, r, G) = L(r, H) + KLD(G, p)$, where $L(r, H)$ is given by the Equation 4.21 and $KLD(G, p)$ is expressed in the Equation 4.8, we obtain the final expression for the probability as follows

$$\Theta(H, G) = \sum_r \exp(-N\ Sr(H, r, G)) \left(\frac{(2\pi)^{\frac{4-a}{2}}}{N^{\frac{2+a}{2}} \prod_{i\in J} \sqrt{G_i}} \left(\prod_{j,k|j<k} \exp(r_{j,k}/2)\right)^{-1}\right). \tag{4.23}$$

4.6 Analysis of One-step Rate Functional

From the discussion in previous sections, if we have a target population with g genotypes, B as the set of genotypes that satisfy boundary condition and $J = \Gamma - B$ as the set of genotypes away from boundary. Then rate functional for one-step is

$$RF = L(r, H) + KLD(G, p)$$

where $L(r, H)$ is given by the Equation 4.21 and $KLD(G, p)$ is expressed in the Equation 4.8.

Let μ be a probability distribution of Frechet type on the Borel σ algebra of a separable locally convex vector space E. Let $\lambda : E \to [0, +\infty]$ be the general Cramer transform of μ defined in chapter 3. The Cramer Set Functional $A \to \Lambda(A)$ associated to μ is then defined for all subsets A of E by

$$\Lambda(A) = \inf_{x \in A} \lambda(x)$$

Then Λ takes values in $[0, +\infty]$. The function λ uniquely determines the functional Λ and conversely. For more detailed discussion and proofs refer to [4], [38], [39] and [6].

Recall here that the sum is being taken over the non-zero entries of mutation matrix r and hence is a finite sum. So applying this generic principle of large deviations to a a finite sum of exponentials $exp(-NA_q)$ with $1 \leq q \leq q_{\max}$, we get its equivalence to $cte \times exp(-NA)$ where $A = \min(A_q)$ and cte is some constant.

Here we have a sum over all matrices r which gives a polynomial number of terms

$< N^{g^2/2}$. Then the optimal large deviations rate functional, RF will be

$$\mathrm{RF}_{\mathrm{opt}} = \min_{\mathrm{r}} \left[\mathrm{L}(\mathrm{r}, \mathrm{H}) + \sum_{\mathrm{j} \in \mathrm{J}} \mathrm{G}_{\mathrm{j}} \log \frac{\mathrm{G}_{\mathrm{j}}}{\mathrm{p}_{\mathrm{j}}} \right] \quad \text{as} \;\; \mathrm{N} \to \infty,$$

As before, we analyze the restricted and non-restricted mutation cases separately.

4.6.1 Non-restricted Mutations

For our model the mutation rate, m is assumed to be same for all genotypes. We deliberately study this slightly simplified context but all the arguments we develop here also apply to the generic case of distinct mutations. The one-step large deviations rate, RF is given by

$$\begin{aligned}\mathrm{RF}_{\mathrm{opt}} = \min_r [&\sum_{j,k|j\neq k} (m_k F_j H_j) - \sum_{j,k|j\neq k} r_{j,k} \log(m_k F_j H_j) + \sum_{j,k|j\neq k} r_{j,k} \log r_{j,k} \\ &- \sum_{j,k|j\neq k} r_{j,k} + \sum_{j\in J} G_j \log \frac{G_j}{p_j}]. \end{aligned} \tag{4.24}$$

Thus, for the optimal value of r

$$\mathrm{RF} = (g-1)m\langle F,H\rangle - \sum_{j,k|j\neq k} r_{j,k}(1+\log(mF_jH_j) - \log r_{j,k}) + \sum_{j\in J} G_j \log \frac{G_j}{p_j}, \tag{4.25}$$

Differentiating the function RF given in Equation 4.25 w.r.t r in order to find the minimum, we get

$$\begin{aligned}\frac{\partial \mathrm{RF}}{\partial r_{l,n}} &= -(1+\log(mF_lH_l)) + (1+\log r_{l,n}) - \sum_{j\in J} \frac{G_j}{p_j}\frac{\partial p_j}{\partial r_{l,n}}, \\ &= -\log(mF_lH_l) + \log r_{l,n} - \sum_{j\in J} \frac{G_j}{p_j}\frac{\partial p_j}{\partial r_{l,n}}. \end{aligned}$$

Now we know that

$$p_j = \frac{F_jH_j}{\langle F,H\rangle} - \frac{\sum_{k|k\neq j} r_{j,k}}{\langle F,H\rangle} + \frac{\sum_{k|k\neq j} r_{k,j}}{\langle F,H\rangle},\ \forall j = 1,...,g$$

and so,

$$\frac{\partial p_j(r,H)}{\partial r_{l,n}} = \begin{cases} 0 & j \neq l,\ j \neq n \\ -\langle F,H\rangle^{-1} & j = l \\ \langle F,H\rangle^{-1} & j = n \end{cases}$$

Let $q_j = p_j\langle F,H\rangle$, then

$$q_j = F_j H_j - \sum_{k|k\neq j} r_{j,k} + \sum_{k|k\neq j} r_{k,j}$$

and

$$\frac{\partial q_j}{\partial r_{l,n}} = \begin{cases} 0 & j \neq l,\ j \neq n \\ -1 & j = l \\ 1 & j = n \end{cases}$$

Thus, we get the following expression

$$\frac{\partial \text{RF}}{\partial r_{l,n}} = \begin{cases} -\log(mF_lH_l) + \log r_{l,n} + \left(\frac{G_l}{q_l} - \frac{G_n}{q_n}\right) & l,n \in J \\ -\log(mF_lH_l) + \log r_{l,n} + \frac{G_l}{q_l} & l \in J,\ n \notin J \\ -\log(mF_lH_l) + \log r_{l,n} - \frac{G_n}{q_n} & l \notin J,\ n \in J \\ -\log(mF_lH_l) + \log r_{l,n} & l \notin J,\ n \notin J \end{cases}$$

or

$$\frac{\partial \text{RF}}{\partial r_{l,n}} = \begin{cases} \log \frac{r_{l,n}}{mF_lH_l} + \left(\frac{G_l}{q_l} - \frac{G_n}{q_n}\right) & l,n \in J \\ \log \frac{r_{l,n}}{mF_lH_l} + \frac{G_l}{q_l} & l \in J,\ n \notin J \\ \log \frac{r_{l,n}}{mF_lH_l} - \frac{G_n}{q_n} & l \notin J,\ n \in J \\ \log \frac{r_{l,n}}{mF_lH_l} & l \notin J,\ n \notin J \end{cases}$$

4.6.2 Restricted Mutations

For the case of restricted mutations, we have our modified cost function as follows

$$\begin{aligned}\mathrm{RF}_{\mathrm{opt}} = \min_r [\sum_{j,k|j<k} m_k F_j H_j - \sum_{j,k|j<k} r_{j,k} \log(m_k F_j H_j) + \sum_{j,k|j<k} r_{j,k} \log r_{j,k} \\ - \sum_{j,k|j<k} r_{j,k} + \sum_{j \in J} G_j \log \frac{G_j}{p_j}].\end{aligned}$$

Thus,

$$\mathrm{RF} = \sum_{j,k|j<k} m_k F_j H_j - \sum_{j,k|j<k} r_{j,k}(1 + \log mF_j H_j) - \log r_{j,k}) + \sum_{j \in J} G_j \log \frac{G_j}{p_j}. \quad (4.26)$$

Again,

$$p_j = \frac{F_j H_j}{\langle F, H \rangle} - \frac{\sum_{k|k>j} r_{j,k}}{\langle F, H \rangle} + \frac{\sum_{k|k<j} r_{k,j}}{\langle F, H \rangle}$$

Thus, introducing q_j as before where

$$q_j = F_j H_j - \sum_{k|k>j} r_{j,k} + \sum_{k|k<j} r_{k,j}$$

we have that

$$\frac{\partial q_j}{\partial r_{l,n}} = \begin{cases} 0 & j \neq l,\ j \neq n \\ 0 & l \geq n \\ -1 & j = l, l < n \\ 1 & j = n, l < n \end{cases}$$

Differentiating the rate functional, RF given by Equation 4.26 w.r.t r in order to find the minimum, we get

$$\frac{\partial \mathrm{RF}}{\partial r_{l,n}} = \begin{cases} 0 & l \geq n \\ -\log mF_l H_l + \log r_{l,n} - \sum_{j \in J} \frac{G_j}{p_j} \frac{\partial p_j}{\partial r_{l,n}}. & l < n \end{cases}$$

Thus, we get the following expression

$$\frac{\partial \mathrm{RF}}{\partial r_{l,n}} = \begin{cases} 0 & l \geq n \\ -\log mF_lH_l + \log r_{l,n} + \left(\frac{G_l}{q_l} - \frac{G_n}{q_n}\right) & l < n, l, n \in J \\ -\log mF_lH_l + \log r_{l,n} + \frac{G_l}{q_l} & l < n, l \in J,\ n \notin J \\ -\log mF_lH_l + \log r_{l,n} - \frac{G_n}{q_n} & l < n, l \notin J,\ n \in J \\ -\log mF_lH_l + \log r_{l,n} & l < n, l \notin J,\ n \notin J \end{cases}$$

or

$$\frac{\partial \mathrm{RF}}{\partial r_{l,n}} = \begin{cases} 0 & l \geq n \\ \log \frac{r_{l,n}}{mF_lH_l} + \left(\frac{G_l}{q_l} - \frac{G_n}{q_n}\right) & l < n, l, n \in J \\ \log \frac{r_{l,n}}{mF_lH_l} + \frac{G_l}{q_l} & l < n, l \in J,\ n \notin J \\ \log \frac{r_{l,n}}{mF_lH_l} - \frac{G_n}{q_n} & l < n, l \notin J,\ n \in J \\ \log \frac{r_{l,n}}{mF_lH_l} & l < n, l \notin J,\ n \notin J \end{cases}$$

For the special case of $g = 3$ genotypes and restricted mutations we have that there are only 3 non-zero values in the mutation matrix.

Let $r_{1,2} = x,\ r_{1,3} = y,\ r_{2,3} = z$. Then our equations simplify as follows

$$\begin{aligned} \mathrm{RF} = &\sum_{j,k|j<k} m_kF_jH_j - x(1 + \log mF_1H_1) - y(1 + \log mF_1H_1) - z(1 + \log mF_2H_2) \\ &+ x\log x + y\log y + z\log z + \sum_{j\in J} G_j \log \frac{G_j}{p_j}. \end{aligned} \tag{4.27}$$

Also, the intermediate probability vector is given as

$$\begin{aligned}
p(1) &= \frac{F_1H_1}{\langle F,H\rangle} - \frac{x+y}{\langle F,H\rangle},\\
p(2) &= \frac{F_2H_2}{\langle F,H\rangle} - \frac{z}{\langle F,H\rangle} + \frac{x}{\langle F,H\rangle},\\
p(3) &= \frac{F_3H_3}{\langle F,H\rangle} + \frac{y+z}{\langle F,H\rangle}
\end{aligned}$$

So,

$$\begin{aligned}
\frac{\partial p_1}{\partial x} &= \frac{\partial p_1}{\partial y} = \frac{\partial p_2}{\partial z} = -\frac{1}{\langle F,H\rangle},\\
\frac{\partial p_1}{\partial z} &= \frac{\partial p_2}{\partial y} = \frac{\partial p_3}{\partial x} = 0,\\
\frac{\partial p_2}{\partial x} &= \frac{\partial p_3}{\partial y} = \frac{\partial p_3}{\partial z} = \frac{1}{\langle F,H\rangle}.
\end{aligned}$$

In terms of q, we get following expressions

$$\begin{aligned}
q_1 &= F_1H_1 - x - y,\\
q_2 &= F_2H_2 - z + x,\\
q_3 &= F_3H_3 + y + z
\end{aligned}$$

So,

$$\begin{aligned}
\frac{\partial q_1}{\partial x} &= \frac{\partial q_1}{\partial y} = \frac{\partial q_2}{\partial z} = -1,\\
\frac{\partial q_1}{\partial z} &= \frac{\partial q_2}{\partial y} = \frac{\partial q_3}{\partial x} = 0,\\
\frac{\partial q_2}{\partial x} &= \frac{\partial q_3}{\partial y} = \frac{\partial q_3}{\partial z} = 1,
\end{aligned}$$

Again, differentiating RF given by Equation 4.27 w.r.t r in order to find minimum, we get

$$\frac{\partial \mathrm{RF}}{\partial r_{l,n}} = \begin{cases} 0 & l \geq n \\ -\log mF_l H_l + \log r_{l,n} - \sum_{j \in J} \frac{G_j}{p_j} \frac{\partial p_j}{\partial r_{l,n}}. & l < n \end{cases}$$

Expressing them with new variables we get the following equations

$$\frac{\partial \mathrm{RF}}{\partial x} = \begin{cases} -\log mF_1 H_1 + \log x + \left(\frac{G_1}{q_1} - \frac{G_2}{q_2}\right) & 1,2 \in J \\ -\log mF_1 H_1 + \log x + \frac{G_1}{q_1} & 1 \in J,\ 2 \notin J \\ -\log mF_1 H_1 + \log x - \frac{G_2}{q_2} & 1 \notin J,\ 2 \in J \\ -\log mF_1 H_1 + \log x & 1,2 \notin J \end{cases}$$

$$\frac{\partial \mathrm{RF}}{\partial y} = \begin{cases} -\log mF_1 H_1 + \log y + \left(\frac{G_1}{q_1} - \frac{G_3}{q_3}\right) & 1,3 \in J \\ -\log mF_1 H_1 + \log y + \frac{G_1}{q_1} & 1 \in J,\ 3 \notin J \\ -\log mF_1 H_1 + \log y - \frac{G_3}{q_3} & 1 \notin J,\ 3 \in J \\ -\log mF_1 H_1 + \log y & 1,3 \notin J \end{cases}$$

and

$$\frac{\partial \mathrm{RF}}{\partial z} = \begin{cases} -\log mF_2 H_2 + \log z + \left(\frac{G_2}{q_2} - \frac{G_3}{q_3}\right) & 2,3 \in J \\ -\log mF_2 H_2 + \log z + \frac{G_2}{q_2} & 2 \in J,\ 3 \notin J \\ -\log mF_2 H_2 + \log z - \frac{G_3}{q_3} & 2 \notin J,\ 3 \in J \\ -\log mF_2 H_2 + \log z & 2,3 \notin J \end{cases}$$

Clearly the above expressions give explicit formulas for x, y, z depending on the non boundary set J. Various cases are outlined below.

(i) If $1, 2, 3 \in J$, then

$$\begin{aligned} x &= mF_1H_1 \exp\left(\frac{G_2}{q_2} - \frac{G_1}{q_1}\right), \\ y &= mF_1H_1 \exp\left(\frac{G_3}{q_3} - \frac{G_1}{q_1}\right), \\ z &= mF_2H_2 \exp\left(\frac{G_3}{q_3} - \frac{G_2}{q_2}\right). \end{aligned}$$

(ii) If $1 \notin J$, and $2, 3 \in J$, then

$$\begin{aligned} x &= mF_1H_1 \exp\left(\frac{G_2}{q_2}\right), \\ y &= mF_1H_1 \exp\left(\frac{G_3}{q_3}\right), \\ z &= mF_2H_2 \exp\left(\frac{G_3}{q_3} - \frac{G_2}{q_2}\right). \end{aligned}$$

(iii) If $2 \notin J$, and $1, 3 \in J$, then

$$\begin{aligned} x &= mF_1H_1 \exp\left(\frac{-G_1}{q_1}\right), \\ y &= mF_1H_1 \exp\left(\frac{G_3}{q_3} - \frac{G_1}{q_1}\right), \\ z &= mF_2H_2 \exp\left(\frac{G_3}{q_3}\right). \end{aligned}$$

(iv) If $3 \notin J$, and $1, 2 \in J$, then

$$\begin{aligned} x &= mF_1H_1 \exp\left(\frac{G_2}{q_2} - \frac{G_1}{q_1}\right), \\ y &= mF_1H_1 \exp\left(\frac{-G_1}{q_1}\right), \\ z &= mF_2H_2 \exp\left(\frac{-G_2}{p_2}\right). \end{aligned}$$

(v) If $2, 3 \notin J$, and $1 \in J$, then

$$\begin{aligned} x &= mF_1H_1 \exp\left(\frac{-G_1}{q_1}\right), \\ y &= mF_1H_1 \exp\left(\frac{-G_1}{q_1}\right), \\ z &= mF_2H_2. \end{aligned}$$

(vi) If $1, 3 \notin J$, and $2 \in J$, then

$$\begin{aligned} x &= mF_1H_1 \exp\left(\frac{G_2}{q_2}\right), \\ y &= mF_1H_1, \\ z &= mF_2H_2 \exp\left(\frac{-G_2}{q_2}\right). \end{aligned}$$

(vii) If $1, 2 \notin J$,and $3 \in J$, then

$$\begin{aligned} x &= mF_1H_1, \\ y &= mF_1H_1 \exp\left(\frac{G_3}{q_3}\right), \\ z &= mF_2H_2 \exp\left(\frac{G_3}{q_3}\right). \end{aligned}$$

(viii) If $1, 2, 3 \notin J$, then

$$\begin{aligned} x &= mF_1H_1, \\ y &= mF_1H_1, \\ z &= mF_2H_2. \end{aligned}$$

Thus, we can say that $x = mA$, $y = mB$, and $z = mC$ where the values of

A, B, C are determined as explained above. This gives the following expression

$$\begin{aligned}\mathrm{RF} = {} & m(2F_1H_1 + F_2H_2) - mA(1 + \log mF_1H_1) - mB(1 + \log mF_1H_1) \\ & - mC(1 + \log mF_2H_2) + mA \log mA + mB \log mB + mC \log mC \\ & + \sum_{j \in J} G_j \log \frac{G_j}{p_j}.\end{aligned}$$

Simplifying the above expression we obtain

$$\begin{aligned}\mathrm{RF} = {} & m(2F_1H_1 + F_2H_2) - mA - mA \log m - mA \log F_1H_1 \\ & - mB - mB \log m - mB \log F_1H_1 \\ & - mC - mC \log m - mC \log F_2H_2 \\ & + mA \log m + mA \log A + mB \log m + mB \log B \\ & + mC \log m + mC \log C + \sum_{j \in J} G_j \log \frac{G_j}{p_j}.\end{aligned}$$

or

$$\begin{aligned}\mathrm{RF} = {} & m(2F_1H_1 + F_2H_2) - mA - mA \log F_1H_1 - mB - mB \log F_1H_1 \\ & - mC - mC \log F_2H_2 + mA \log A + mB \log B + mC \log C + \sum_{j \in J} G_j \log \frac{G_j}{p_j}.\end{aligned}$$

and so

$$\begin{aligned}\mathrm{RF} = {} & m(2F_1H_1 + F_2H_2 - A - A \log F_1H_1 - B - B \log F_1H_1 \\ & - C - C \log F_2H_2 + A \log A + B \log B + C \log C) + \sum_{j \in J} G_j \log \frac{G_j}{p_j}.\end{aligned}$$

In other words we observe that the final cost given by rate functional, RF is composed of two parts where the Poisson part is a multiple of mutation rate. Another interesting observation to make here is that the cost from Poisson part is positive so presence of mutations actually leads to increase in overall cost. This was verified numerically and it was also observed that cost from multinomial process is in general much higher than Poisson cost.

Recall that N is large (for instance 10^5) and the m_j are small (i.e. 10^{-6}), and the F_j can be of the order of 200. Other realistic combinations are also possible.

We present the comparison of these two costs for a particular case of 2 genotypes and 1% discretization in sample space.

The non-zero values of Poisson cost vary between 2×10^{-17} and 8×10^{-6}. On the other hand, non-zero values of multinomial cost have a minimum of 1.2×10^{-8} and a maximum of Infinity.

The histograms of Poisson cost and multinomial cost for the case of 2 genotypes and 1% discretization are shown in Figure 4.1.

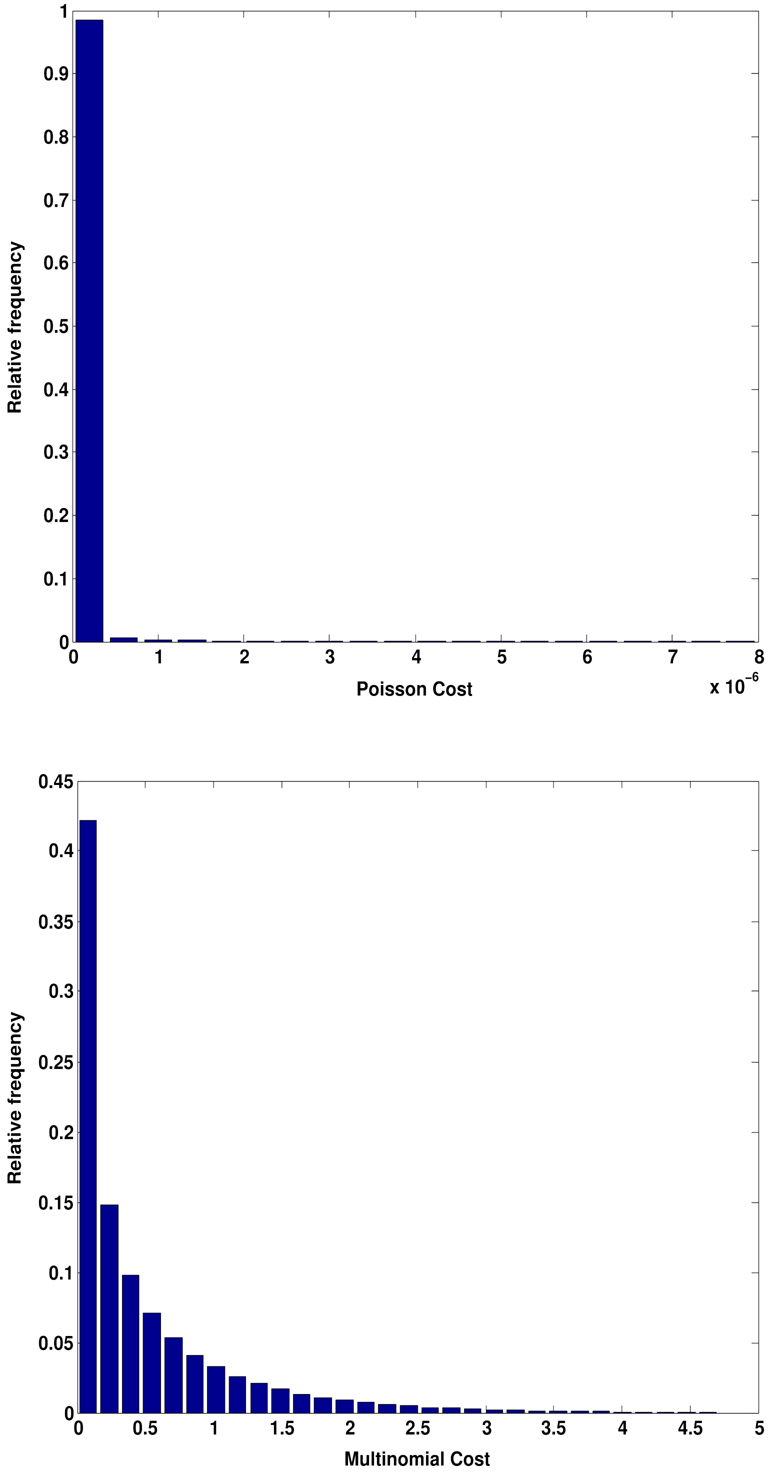

Figure 4.1: The Histograms of Poisson Cost and Multinomial Cost for the Case of 2 Genotypes and 1% Discretization.

Note that in the case of general mutations one cannot derive closed form expressions for the matrix of random mutations, r which minimizes the one step rate function for transition between H and G easily. The higher number of variables and equations increases the complexity of optimization.

As could be expected from fairly general large deviation principles, we have numerical evidence for multiple cases that our one-step rate function $S(H, r, G)$ is a convex function of mutations, r and thus we assume the success of above approach in estimating the global optimal value of cost for optimal value of r. The Hessian of the cost matrix turns out to be symmetric and positive definite. For the case of 3 genotypes and restricted mutations, the Hessian was computed theoretically and its eigen values were analyzed numerically. The symmetric Hessian matrix is presented below

$$\begin{bmatrix} \frac{1}{x} + \frac{G_1}{q_1^2} + \frac{G_2}{q_2^2} & \frac{G_1}{q_1^2} & -\frac{G_2}{q_2^2} \\ \frac{G_1}{q_1^2} & \frac{1}{y} + \frac{G_1}{q_1^2} + \frac{G_3}{q_3^2} & \frac{G_3}{q_3^2} \\ -\frac{G_2}{q_2^2} & \frac{G_3}{q_3^2} & \frac{1}{z} + \frac{G_2}{q_2^2} + \frac{G_3}{q_3^2} \end{bmatrix}$$

All the eigen values were observed to be strictly positive. However, we did observe that for histograms with genotypes near boundary the eigen values were very small. Some of them were of the order 10^{-10}. This is acceptable since the expression for first and second derivative of cost are well defined for histograms away from boundary.

For this particular case with $g = 3$ genotypes and restricted mutations, it is easy to verify that the hessian matrix is indeed positive definite. The determinants for all the three principal square submatrices are clearly positive and are presented below.

1.
$$\frac{1}{x}+\frac{G_1}{q_1^2}+\frac{G_2}{q_2^2},$$

2.
$$\left(\frac{1}{x}+\frac{G_2}{q_2^2}\right)\left(\frac{1}{y}+\frac{G_1}{q_1^2}+\frac{G_3}{q_3^2}\right)+\frac{G_1}{q_1^2}\left(\frac{1}{y}+\frac{G_1}{q_1^2}\right),\ \text{and}$$

3.
$$\begin{aligned}&\frac{1}{xyz}+\frac{G_2}{xyq_2^2}+\frac{G_3}{xyq_3^2}+\frac{G_1}{xzq_1^2}+\frac{G_1G_2}{xq_1^2q_2^2}+\frac{G_1G_3}{xq_1^2q_3^2}+\frac{G_3}{xzq_3^2}+\frac{G_2G_3}{xq_2^2q_3^2}\\&+\frac{G_1}{yzq_1^2}+\frac{G_1G_2}{yq_1^2q_2^2}+\frac{G_1G_3}{yq_1^2q_3^2}+\frac{G_1^2G_3}{q_1^4q_3^2}+\frac{G_1G_3}{zq_1^2q_3^2}\\&+\frac{G_2}{yzq_2^2}+\frac{G_2G_3}{yq_2^2q_3^2}+\frac{G_1G_2}{zq_1^2q_2^2}+\frac{G_2G_3}{zq_2^2q_3^2}.\end{aligned}$$

Next, we present a few histograms showing distribution of minimum eigen values for each cost matrix observed.

The Figure 4.2 shows a histogram of minimum eigen values of histograms away from boundary.

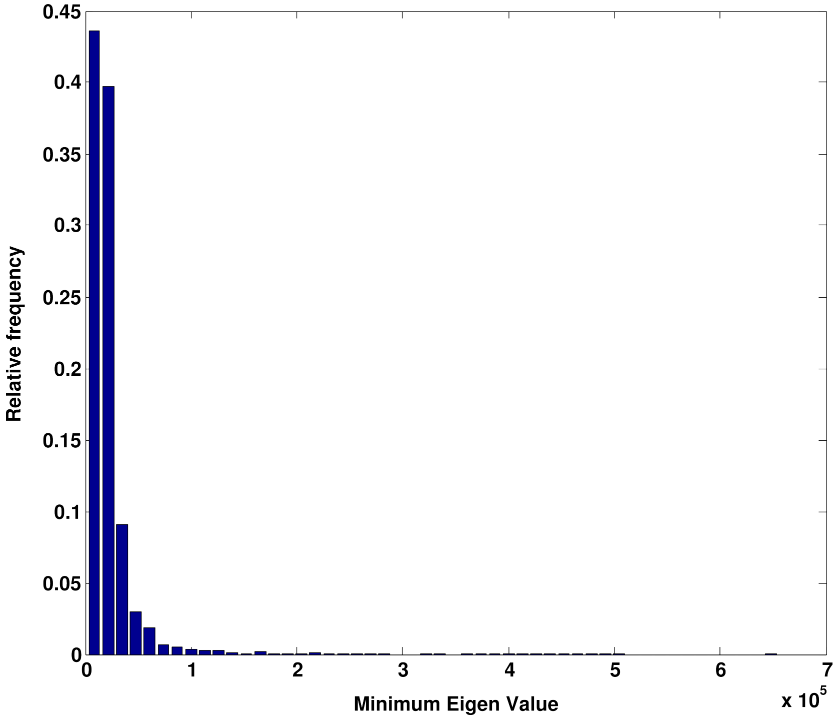

Figure 4.2: Histogram of Minimum Eigen Values for Interior Points

Some of the eigen values have very large magnitude resulting in the skewed histogram. It is important to note here that x-axis has a scale of the order 10^5. The minimum of all eigen values plotted here is 2000.

and a zoomed in version of the previous histogram near the origin is shown here in figure 4.3.

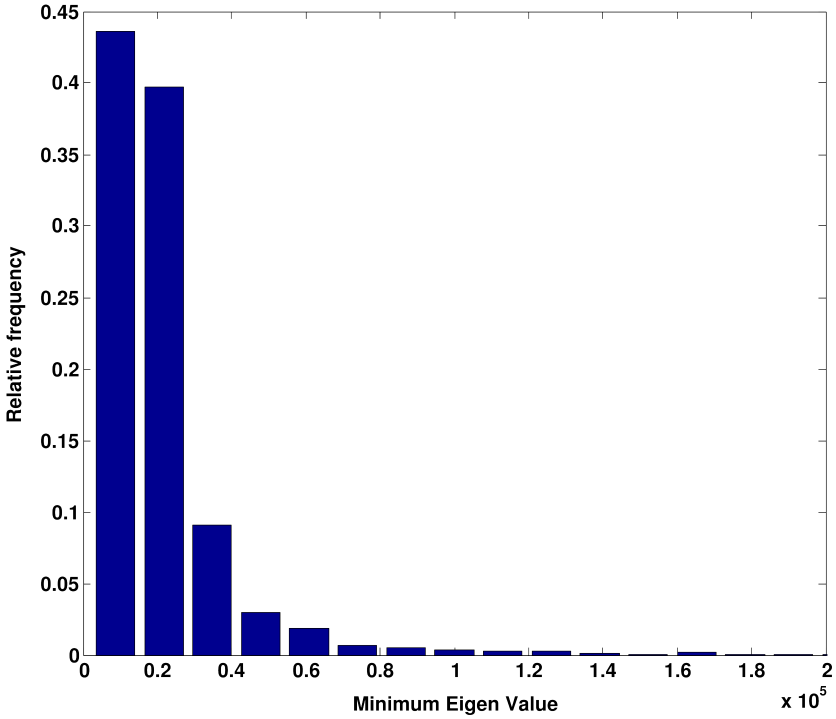

Figure 4.3: Histogram (near Origin) of Minimum Eigen Values for Interior Points

The following figure 4.4 shows a histogram of minimum eigen values of histograms near the boundary.

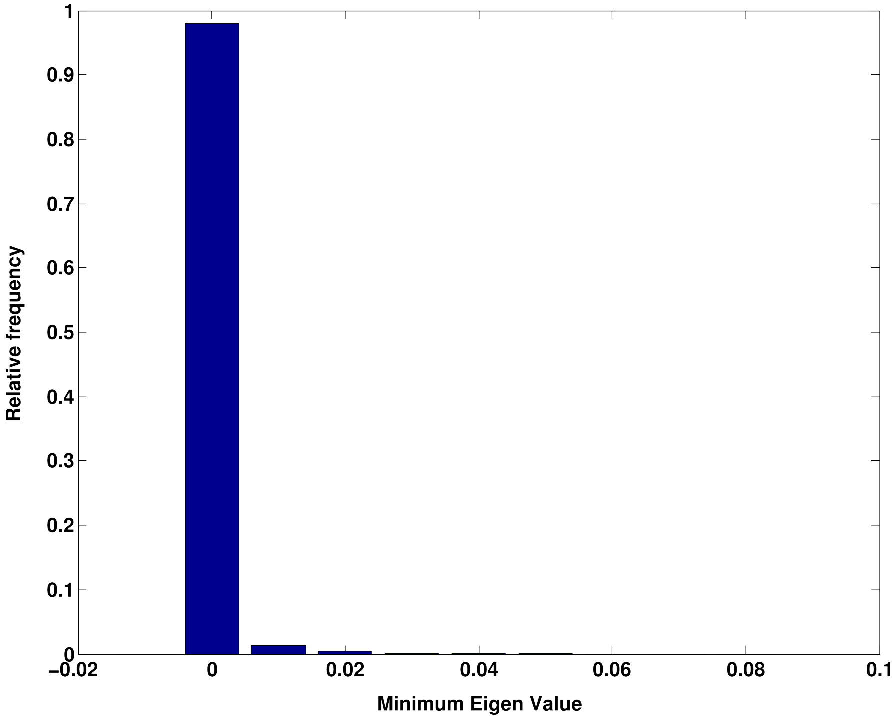

Figure 4.4: Histogram of Minimum Eigen Values for Boundary Points

As, explained before we observe the highest frequency near origin.

For the case with $g = 3$ genotypes, we can view the state space as a $2-$ dimensional simplex which is a unit triangle with one of the vertices on origin. So the distribution of points which have minimum eigen values for the cost matrix less than 10^{-7} is along the edges of this triangle as shown in figure 4.5.

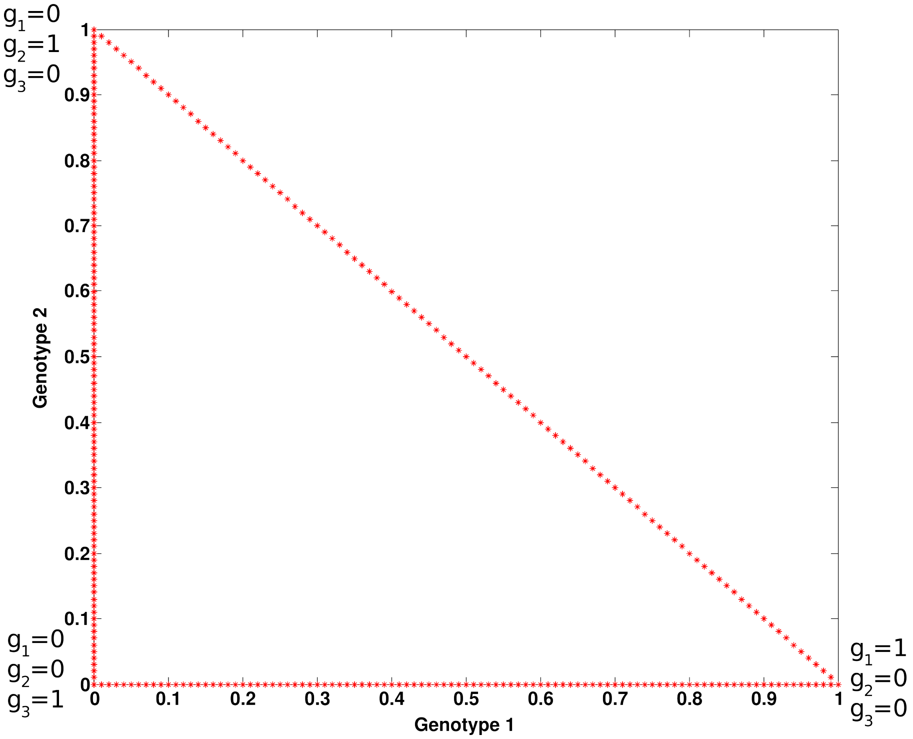

Figure 4.5: Boundary Points

Note that in equation 4.25, if we assume mean mutations, i.e., $r_{j,k} = mF_jH_j$ and mean density for multinomial random sampling, i.e., $G_i = p_i$, the expression for one-step rate function transforms to

$$\mathrm{RF} = (g-1)m\langle F, H\rangle - \sum_{j,k|j\neq k} mF_jH_j(1+\log(mF_jH_j) - \log mF_jH_j) + \sum_{j\in J} p_j \log\frac{p_j}{p_j},$$

or

$$
\begin{aligned}
\mathrm{RF} &= (g-1)m\langle F, H\rangle - \sum_{j,k|j\neq k} mF_j H_j, \\
&= (g-1)m\langle F, H\rangle - (g-1)m\langle F, H\rangle \qquad\qquad = 0
\end{aligned}
$$

Hence we achieve minimal cost or optimal value of rate functional for a trajectory which follows path given by mean mutations and mean sampling distribution. This optimal trajectory can be computed explicitly and it is the most likely trajectory which always directs the population towards fittest genotype. Hence we can not use this approach for finding optimal trajectories for other events which are rare in nature.

CHAPTER 5

Numerical Computation of One-step Rate Functional

To compute optimal value of rate functional given initial and target histograms, we need to find corresponding cost optimizing mutation matrix, r in the evolutionary step. In chapter 4 we derived implicit equations for random mutations in the system using the derivative of rate functional to be zero. We also demonstrated that except for few particular cases with restricted mutations, it was not feasible to solve the implicit system and derive closed form solutions for value of mutations.

One of the techniques for solving directly the implicit equations is to use optimization toolbox in MATLAB. When optimizing optimal mutations using an inbuilt subroutine 'fmincon' with all the constraints for $g = 3$ genotypes, we find that the complete computation of one-step rate functional which includes computation of optimal mutations takes around $2-5$ seconds per pair of initial and target histogram. So for a discretization of 2% with $g=3$ genotypes, where we have 1326 states in the state space, we would need atleast 60 days to finish computations for one-step rate functional.

Hence we need to devise an alternative approach which allows us to estimate cost optimizing values of the mutation matrix.

Next, we describe an efficient way to find optimal value of mutations. Consider, the equation

$$0 = \frac{\partial \mathrm{RF}}{\partial r_{l,n}}$$

Rearranging the terms we get following equations for r, where J is the set of genotypes in target state G which are away from boundary, i.e. $G_j \geq \epsilon,\ j \in J$.

$$\log r_{l,n} = \begin{cases} \log(mF_l H_l) - \left(\frac{G_l}{q_l} - \frac{G_n}{q_n}\right) & l, n \in J \\ \log(mF_l H_l) - \frac{G_l}{q_l} & l \in J,\ n \notin J \\ \log(mF_l H_l) + \frac{G_n}{q_n} & l \notin J,\ n \in J \\ \log(mF_l H_l) & l \notin J,\ n \notin J \end{cases}$$

We want to solve these equations for the value of r.

5.1 Optimal Intermediary Mutation Step

We will study two situations, the generic case with non-restricted mutations and special case with restricted mutations separately.

5.1.1 Non-restricted Mutations

For an initial histogram H and target histogram G with mutation matrix r, where set of boundary genotypes B is empty, we have that the rate functional is given by

$$\mathrm{RF} = \sum_{j,k|j\neq k} m_k F_j H_j - \sum_{j,k|j\neq k} r_{j,k}(1+\log(m_k F_j H_j)) + \sum_{j,k|j\neq k} r_{j,k}\log r_{j,k} + \sum_j G_j \log\frac{G_j}{p_j}.$$

Now we know that mean of r is $\bar{r}_{j,k} = m_k F_j H_j$ and we also assume $m = m_i, \forall i$ to simplify results.

then we have

$$\mathrm{RF} = (g-1)m\langle F,H\rangle + \sum_{j,k|j\neq k} r_{j,k}(\log\frac{r_{j,k}}{\bar{r}_{j,k}} - 1) + \sum_j G_j\log\frac{G_j}{p_j}. \tag{5.1}$$

and the derivatives w.r.t. r are given by

$$\frac{\partial \mathrm{RF}}{\partial r_{j,k}} = -\log(mF_jH_j) + \log r_{j,k} - \frac{1}{\langle F,H\rangle}\left(\frac{G_j}{p_j} - \frac{G_k}{p_k}\right).$$

So rearranging the terms, we get the stationarity conditions

$$\begin{aligned} 0 &= \frac{\partial \mathrm{RF}}{\partial r_{j,k}} \\ &= \log\left(\frac{r_{j,k}}{mF_jH_j}\right) - \frac{1}{\langle F,H\rangle}\left(\frac{G_j}{p_j} - \frac{G_k}{p_k}\right). \end{aligned} \tag{5.2}$$

So we obtain for a given H and G, r as the solution of the following system of g^2 equations in g^2 unknowns given by $r_{j,k}$ and p_j

$$r_{j,k} = mF_jH_j\exp\left(\frac{1}{\langle F,H\rangle}\left(\frac{G_k}{p_k} - \frac{G_j}{p_j}\right)\right),$$

$$p_j = p(j) = \frac{F_jH_j}{\langle F,H\rangle} - \frac{\sum_k r_{j,k}}{\langle F,H\rangle} + \frac{\sum_k r_{k,j}}{\langle F,H\rangle}.$$

or it can be written as a solution to the following system,

$$r_{j,k} = mF_jH_j \exp\left(\frac{G_k}{q_k} - \frac{G_j}{q_j}\right),$$
$$q_j = q(j) = F_jH_j - \sum_k r_{j,k} + \sum_k r_{k,j}.$$

where $q_j = p_j\langle F, H\rangle$ and $q_j \geq 0$. Also let

$$\rho_{j,k} = \frac{r_{j,k}}{\bar{r}\ \ _{j,k}} = \frac{r_{j,k}}{m_kF_jH_j} = \frac{r_{j,k}}{mF_jH_j}. \tag{5.3}$$

Then, from the expression for q,

$$\begin{aligned} q_j &= F_jH_j - \sum_k r_{j,k} + \sum_k r_{k,j} \\ &= F_jH_j - \sum_{k|k\neq j} \rho_{j,k}\bar{r}_{j,k} + \sum_{k|k\neq j} \rho_{k,j}\bar{r}_{k,j}, \\ &= F_jH_j - F_jH_j \sum_{k|k\neq j} \rho_{j,k}m + m\sum_{k|k\neq j} \rho_{k,j}F_kH_k, \\ &= F_jH_j\left(1 - m\sum_{k|k\neq j}\rho_{j,k} + \frac{m}{F_jH_j}\sum_{k|k\neq j}\rho_{k,j}F_kH_k\right). \end{aligned}$$

Thus, q is given by Equation 5.4

$$q_j = F_jH_j\left(1 - m\sum_{k|k\neq j}\rho_{j,k} + \frac{m}{F_jH_j}\sum_{k|k\neq j}\rho_{k,j}F_kH_k\right). \tag{5.4}$$

If $m = 0$, (i.e.) there are no mutations in the system then we have that $q_j^0 = F_jH_j$ where clearly $q^0(j) \geq 0$.

Thus from equations 5.3 and 5.2

$$\rho^0_{j,k} = \exp\left(\frac{G_k}{F_kH_k} - \frac{G_j}{F_jH_j}\right) = \frac{\gamma_k}{\gamma_j} \tag{5.5}$$

where $\gamma_i = \exp\left(\frac{G_i}{F_iH_i}\right)$. Thus we have, for $m = 0$,

$$r^0_{j,k} = mF_jH_j\frac{\gamma_k}{\gamma_j}. \tag{5.6}$$

In case of realistic cell populations, the mutation rate m is very small, generally smaller than 10^{-6}. So for these small values of m, we may use first-order approximations, which we write as follows

$$q_j \simeq q^0_j(1 + m\beta_j),$$

and

$$\rho_{j,k} \simeq \rho^0_{j,k}(1 + m\alpha_{j,k}).$$

where $\alpha j, k$ and β_j are known coefficients.

$$\frac{G_j}{q_j} \simeq \frac{G_j}{q^0_j}(1 + m\beta_j)^{-1} \simeq \frac{G_j}{q^0_j}(1 - m\beta_j) \tag{5.7}$$

Using this approximation in expression for $\rho_{j,k}$, we obtain

$$\begin{aligned}
\rho_{j,k} &= \exp\left(\frac{G_k}{q_k} - \frac{G_j}{q_j}\right), \\
&\simeq \exp\left(\frac{G_k(1 - m\beta_k)}{q^0_k} - \frac{G_j(1 - m\beta_j)}{q^0_j}\right), \\
&\simeq \exp\left(\frac{G_k}{q^0_k} - \frac{mG_k\beta_k}{q^0_k} - \frac{G_j}{q^0_j} + \frac{mG_j\beta_j}{q^0_j}\right), \\
&\simeq \exp\left(\frac{G_k}{F_kH_k} - \frac{G_j}{F_jH_j}\right)\exp\left(m\left(-\frac{G_k\beta_k}{q^0_k} + \frac{G_j\beta_j}{q^0_j}\right)\right), \\
&\simeq \rho^0_{j,k}\exp\left(m\left(-\frac{G_k\beta_k}{q^0_k} + \frac{G_j\beta_j}{q^0_j}\right)\right), \\
&\simeq \rho^0_{j,k}\left(1 + m\left(\frac{G_j\beta_j}{F_jH_j} - \frac{G_k\beta_k}{F_kH_k}\right)\right).
\end{aligned}$$

This yields the following expressions for the $\alpha j, k$ in terms of β_j and β_k.

$$\alpha_{j,k} = \left(\frac{G_j \beta_j}{F_j H_j} - \frac{G_k \beta_k}{F_k H_k} \right) \tag{5.8}$$

Now we use Equation 5.4 for q_j and rewrite it to obtain,

$$q_j = F_j H_j \left(1 + m \left(\sum_{k|k \neq j} -\rho_{j,k} + \frac{1}{F_j H_j} \sum_{k|k \neq j} \rho_{k,j} F_k H_k \right) \right) \tag{5.9}$$

This implies the relations

$$\begin{aligned}
\beta_j &= \sum_{l|l \neq j} -\rho^0_{j,l} + \frac{1}{F_j H_j} \sum_{l|l \neq j} \rho^0_{l,j} F_l H_l, \\
&= \sum_{l,l \neq j} \frac{-\gamma_l}{\gamma_j} + + \frac{1}{F_j H_j} \sum_{l|l \neq j} \frac{\gamma_j}{\gamma_l} F_l H_l, \\
&= -\gamma_j \sum_{l|l \neq j} \frac{1}{\gamma_l} + \frac{1}{F_j H_j \gamma_j} \sum_{l|l \neq j} \gamma_l F_l H_l, \\
&= -\gamma_j \sum_{l|l \neq j} \frac{1}{\gamma_l} + \frac{1}{F_j H_j \gamma_j} \sum_{l|l \neq j} \gamma_l F_l H_l
\end{aligned}$$

Updating values of β_j and β_k in equation 5.8, we derive

$$\begin{aligned}
\alpha_{j,k} = &\frac{G_j}{F_j H_j} \left(\gamma_j \sum_{l|l \neq j} -\frac{1}{\gamma_l} + \frac{1}{F_j H_j \gamma_j} \sum_{l|l \neq j} \gamma_l F_l H_l \right) \\
- &\frac{G_k}{F_k H_k} \left(\gamma_k \sum_{l|l \neq k} -\frac{1}{\gamma_l} + \frac{1}{\gamma_k F_k H_k} \sum_{l|l \neq k} \gamma_l F_l H_l \right)
\end{aligned}$$

where $\gamma_i = \exp\left(\frac{G_i}{F_i H_i}\right)$ and thus we have the first-order approximations for the optimal intermediary mutation step between H and G.

$$r_{j,k} = m F_j H_j \exp\left(\frac{G_k}{F_k H_k} - \frac{G_j}{F_j H_j} \right) (1 + m\alpha_{j,k})$$

or

$$r_{j,k} = mF_jH_j \exp\left(\frac{G_k}{F_kH_k} - \frac{G_j}{F_jH_j}\right)[1 + m\left(\frac{G_k\gamma_k}{F_kH_k} - \frac{G_j\gamma_j}{F_jH_j}\right)\sum_{l|l\neq j}\frac{1}{\gamma_l}$$
$$+ \left(\frac{G_j}{F_j^2H_j^2\gamma_j} - \frac{G_k}{\gamma_kF_k^2H_k^2}\right)\sum_{l|l\neq k}\gamma_lF_lH_l].$$

Thus, using the values of γ we get an approximation for the values of r and q and a very precise approximation of the minimal rate functional for a one-step jump from H to G. The corresponding optimal rate functional is derived in section 5.2 later.

Extending the above discussion to the case which allows one or several genotypes to be near boundary frequencies ($< \epsilon$), we derive similar results and approximations for r and q. As before let J be the set of genotypes in target population G which are away from boundary values ($G_j \geq \epsilon$) and B be the set of remaining boundary genotypes. Also, call I the set of genotypes i such that $H_i \neq 0$.

We claim that $j \notin I \Rightarrow j \notin J$. The mean for random Poisson mutations, $\bar{r}$ is given by $\bar{r}_{j,k} = m_kF_jH_j$.

For the case when $j \notin I$, then $H_j = 0$ and thus

$$\bar{r}_{j,k} = m_kF_jH_j = 0, \forall k.$$

In order for $G_j > \epsilon$ or $j \in J$, we need that $r_{j,k} \geq \epsilon$. Since, our tolerance for boundary cases for population of size $N = 50000$ is $\epsilon = 10^{-3}$, we would need atleast $r_{j,k} \geq 10^{-3}$. This is highly unlikely since $\bar{r}_{j,k} = 0$.

This fact can also be verified numerically. We present the histogram for the values of $\rho_{j,k}$ which is ratio of values of $r_{j,k}$ to its mean value $\bar{r_{j,k}}$ in Figure 5.1. This figure has been generated for every pair of initial and target histogram in the state space of $g = 3$ genotypes generated using 2% discretization. The values of growth factor and mutation rates are used from TC experiments. It is clear from the figure that random mutations have values very close to mean mutations.

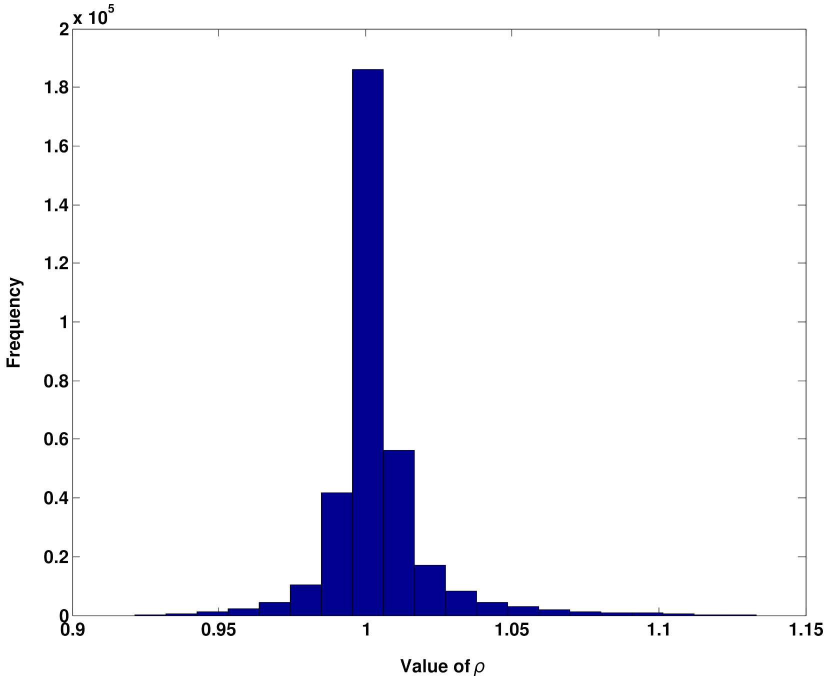

Figure 5.1: Histogram for Ratio of Random Mutations to Mean Mutations.

Hence we can safely assume that if $H_j = 0$, then $p_j << 10^{-3}$ and thus $G_j < 10^{-3}$ or $j \in B$ for optimal trajectory.

Thus, $j \notin I \Rightarrow j \notin J$ or in other words $j \in J \Rightarrow j \in I$. The set of genotypes for which p_i is very small or zero can be similarly included in the set of boundary genotypes in G.

These relations are required so that we do not divide by 0 or extremely small values in the expression for $r^0_{j,k}$.

We have computed the expression for approximate random mutation matrix, r in equation 5.6 when the target histogram G is away from boundary. Next we introduce target histograms which may satisfy boundary conditions and recompute an approximation for the mutation matrix for these special cases.

The generalized one-step rate functional or the cost function for any pair of initial (H) and target (G) histogram using Equation 4.24 is given as

$$\mathrm{RF}{=}\sum_{\mathrm{j}\in\mathrm{I},\mathrm{k}|\mathrm{j}\neq\mathrm{k}}\mathrm{m_kF_jH_j}-\sum_{\mathrm{j}\in\mathrm{I},\mathrm{k}|\mathrm{j}\neq\mathrm{k}}\mathrm{r_{j,k}}(1+\log(\mathrm{m_kF_jH_j}))+\sum_{\mathrm{j}\in\mathrm{I},\mathrm{k}|\mathrm{j}\neq\mathrm{k}}\mathrm{r_{j,k}}\log(\mathrm{r_{j,k}})+\sum_{\mathrm{j}\in\mathrm{J}}\mathrm{G_j}\log\frac{\mathrm{G_j}}{\mathrm{p_j}}. \tag{5.10}$$

The derivative of the rate functional in equation 5.10 for these special cases is given in Equation 5.11.
Recall here that $H_k \neq 0, \forall k \in I$.

$$\frac{\partial \mathrm{RF}}{r_{j,k}} = \begin{cases} 0 & j \notin I; j,k \notin J \\ \frac{-1}{\langle F,H\rangle}\frac{G_k}{p_k} & j \notin I; j \notin J, k \in J \\ \log(\rho_{j,k}) & j \in I; j,k \notin J \\ \log(\rho_{j,k}) + \frac{1}{\langle F,H\rangle}\left(\frac{G_j}{p_j} - \frac{G_j}{p_j}\right) & j \in I; j,k \in J \\ \log(\rho_{j,k}) + \frac{1}{\langle F,H\rangle}\frac{G_j}{p_j} & j \in I; j \in J, k \notin J \\ \log(\rho_{j,k}) + \frac{-1}{\langle F,H\rangle}\frac{G_k}{p_k} & j \in I; j \notin J, k \in J \end{cases} \tag{5.11}$$

Using the fact that $q_j = p_j\langle F, H\rangle$ and rewriting, we get

$$\rho_{j,k} = \begin{cases} 0 & j \notin I \\ 1 & j \in I; j,k \notin J \\ \exp\left(\frac{G_k}{q_k} - \frac{G_j}{q_j}\right) & j \in I; j,k \in J \\ \exp\left(-\frac{G_j}{q_j}\right) & j \in I; j \in J, k \notin J \\ \exp\left(\frac{G_k}{q_k}\right) & j \in I; j \notin J, k \in J \end{cases} \tag{5.12}$$

and thus

$$r_{j,k} = \begin{cases} 0 & j \notin I \\ mF_jH_j & j \in I; j,k \notin J \\ mF_jH_j \exp\left(\frac{G_k}{q_k} - \frac{G_j}{q_j}\right) & j \in I; j,k \in J \\ mF_jH_j \exp\left(\frac{-G_j}{q_j}\right) & j \in I; j \in J, k \notin J \\ mF_jH_j \exp\left(\frac{G_k}{q_k}\right) & j \in I; j \notin J, k \in J \end{cases} \tag{5.13}$$

Proceeding as before, for the case when $m = 0$ we will get

$$q_j = \begin{cases} 0 & j \notin I \\ F_jH_j & j \in I \end{cases}$$

which gives

$$\rho^0_{j,k} = \begin{cases} 0 & j \notin I \\ 1 & j \in I; j,k \notin J \\ \exp\left(\frac{G_k}{F_kH_k} - \frac{G_j}{F_jH_j}\right) & j \in I; j,k \in J \\ \exp\left(-\frac{G_j}{F_jH_j}\right) & j \in I; j \in J, k \notin J \\ \exp\left(\frac{G_k}{F_kH_k}\right) & j \in I; j \notin J, k \in J \end{cases} \tag{5.14}$$

The above expression is valid since if $k \in J$ then $k \in I$ and hence $q_k \neq 0$. So,

$$r^0_{j,k} = \begin{cases} 0 & j \notin I \\ mF_jH_j & j \in I; j, k \notin J \\ mF_jH_j \exp\left(\frac{G_k}{F_kH_k} - \frac{G_j}{F_jH_j}\right) & j \in I; j, k \in J \\ mF_jH_j \exp\left(-\frac{G_j}{F_jH_j}\right) & j \in I; j \in J, k \notin J \\ mF_jH_j \exp\left(\frac{G_k}{F_kH_k}\right) & j \in I; j \notin J, k \in J \end{cases} \tag{5.15}$$

The above calculation completes the zero order approximation for optimal non restricted mutation matrix, r for any pair of initial (H) and target (G) histograms.

5.1.2 Restricted Mutations

Now we consider the special case of restricted mutations in the above system and determine the optimal intermediary values of r and p. Random mutations in a population are said to be restricted when they are only allowed to increase the fitness of population. This means that a genotype with higher growth factor is not allowed to mutate into a genotype with lower growth factor.

$$\mathrm{RF}{=}\sum_{\mathrm{j\in I,k|j<k}} \mathrm{m_k F_j H_j} - \sum_{\mathrm{j\in I,k|j<k}} \mathrm{r_{j,k}}(1+\log(\mathrm{m_k F_j H_j})) + \sum_{\mathrm{j\in I,k|j<k}} \mathrm{r_{j,k}}\log(\mathrm{r_{j,k}}) + \sum_{\mathrm{j\in J}} \mathrm{G_j}\log\frac{\mathrm{G_j}}{\mathrm{p_j}}. \tag{5.16}$$

The derivative of rate functional in equation 5.16 for various possibilities is given in Equation 5.17.

Recall again that $H_k \neq 0, \forall k \in I$.

$$\frac{\partial \mathrm{RF}}{r_{j,k}} = \begin{cases} 0 & j \notin I; j,k \notin J \\ 0 & j \in I, j \geq k; j,k \notin J \\ \frac{-1}{\langle F,H\rangle}\frac{G_k}{p_k} & j \notin I; j \notin J, k \in J \\ \log(\rho_{j,k}) & j \in I, j < k; j,k \notin J \\ \log(\rho_{j,k}) + \frac{1}{\langle F,H\rangle}\left(\frac{G_j}{p_j} - \frac{G_j}{p_j}\right) & j \in I, j < k; j,k \in J \\ \log(\rho_{j,k}) + \frac{1}{\langle F,H\rangle}\frac{G_j}{p_j} & j \in I, j < k; j \in J, k \notin J \\ \log(\rho_{j,k}) + \frac{-1}{\langle F,H\rangle}\frac{G_k}{p_k} & j \in I, j < k; j \notin J, k \in J \end{cases} \tag{5.17}$$

Using the fact that $q_j = p_j\langle F, H\rangle$ and rewriting, we get

$$\rho_{j,k} = \begin{cases} 0 & j \notin I \\ 0 & j \in I, j \geq k \\ 1 & j \in I, j < k; j,k \notin J \\ \exp\left(\frac{G_k}{q_k} - \frac{G_j}{q_j}\right) & j \in I, j < k; j,k \in J \\ \exp\left(\frac{-G_j}{q_j}\right) & j \in I, j < k; j \in J, k \notin J \\ \exp\left(\frac{G_k}{q_k}\right) & j \in I, j < k; j \notin J, k \in J \end{cases} \tag{5.18}$$

and thus

$$r_{j,k} = \begin{cases} 0 & j \notin I \\ 0 & j \in I, j \geq k \\ mF_jH_j & j \in I, j < k; j,k \notin J \\ mF_jH_j \exp\left(\frac{G_k}{q_k} - \frac{G_j}{q_j}\right) & j \in I, j < k; j,k \in J \\ mF_jH_j \exp\left(\frac{-G_j}{q_j}\right) & j \in I, j < k; j \in J, k \notin J \\ mF_jH_j \exp\left(\frac{G_k}{p_k}\right) & j \in I, j < k; j \notin J, k \in J \end{cases} \tag{5.19}$$

Proceeding as before, for the case when $m = 0$ we will get

$$\rho^0_{j,k} = \begin{cases} 0 & j \notin I \\ 0 & j \in I, j \geq k \\ 1 & j \in I, j < k; j,k \notin J \\ \exp\left(\frac{G_k}{F_kH_k} - \frac{G_j}{F_jH_j}\right) & j \in I, j < k; j,k \in J \\ \exp\left(-\frac{G_j}{F_jH_j}\right) & j \in I, j < k; j \in J, k \notin J \\ \exp\left(\frac{G_k}{F_kH_k}\right) & j \in I, j < k; j \notin J, k \in J \end{cases} \tag{5.20}$$

and

$$r^0_{j,k} = \begin{cases} 0 & j \notin I \\ 0 & j \in I, j \geq k \\ mF_jH_j & j \in I, j < k; j,k \notin J \\ mF_jH_j \exp\left(\frac{G_k}{F_kH_k} - \frac{G_j}{F_jH_j}\right). & j \in I, j < k; j,k \in J \\ mF_jH_j \exp\left(-\frac{G_j}{F_jH_j}\right) & j \in I, j < k; j \in J, k \notin J \\ mF_jH_j \exp\left(\frac{G_k}{F_kH_k}\right) & j \in I, j < k; j \notin J, k \in J \end{cases} \tag{5.21}$$

5.1.3 Mathematical Justification

The above approximations for the value of optimal mutation matrix r are not only numerically correct but can also be shown to be a good approximation to the optimal mutation matrix using the implicit function theorem. We focus on a special case where we have $g = 3$, $m_i = m, \forall i$ and $r_{i,i} = 0, \forall i$.

We know that the optimal r is the solution to following system of implicit equations

$$r_{j,k} = mF_jH_j \exp\left(\frac{G_k}{q_k} - \frac{G_j}{q_j}\right), \tag{5.22}$$

$$q_j = F_jH_j - \sum_{k|k\neq j} r_{j,k} + \sum_{k|k\neq j} r_{k,j}. \tag{5.23}$$

This is a system of 8 implicit equations in 8 unknowns $r_{12}, r_{13}, r_{21}, r_{23}, r_{31}, r_{32}, q_1$ and q_2 since $\sum_j q_j = \langle F, H\rangle$, and thus $q_3 = \langle F, H\rangle - q_1 - q_2$. We now appeal to the implicit function theorem [91].

Theorem 5.1.1. *Let A be an open set in $\mathbb{R}^{n+k}$ and let $f : A \to \mathbb{R}^n$ be a C^r function. Write f in the form $f(x, y)$, where x and y are elements of $\mathbb{R}^k$ and $\mathbb{R}^n$. Suppose that (a, b) is a point in A such that $f(a, b) = 0$ and the determinant of the $n \times n$ Jacobian matrix whose elements are the derivatives of the n component functions of f with respect to the n variables, written as y, evaluated at (a, b), is not equal to zero.*

Then there exists a neighborhood B of a in $\mathbb{R}^k$ and a unique C^r function $g : B \to \mathbb{R}^n$ such that $g(a) = b$ and $f(x, g(x)) = 0$ for all $x \in B$.

Hence it says that there is a unique solution $r(m)$ of the system for m close to zero and that, $r(m)$ is a smooth function of m, and hence has a Taylor expansion for m close to zero.

We will show that the Jacobian of system given by 5.22 is non singular and hence using the Implicit function theorem we can approximate r in the close neighborhood of $m = 0$.

The Jacobian matrix is as follows

$$\begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & -e_{1,2}G_1/q_1^2 & e_{1,2}G_2/q_2^2 \\ 0 & 1 & 0 & 0 & 0 & 0 & -e_{1,3}(G_3/q_3^2 + G_1/q_1^2) & -e_{1,3}G_3/q_3^2 \\ 0 & 0 & 1 & 0 & 0 & 0 & e_{2,1}G_1/q_1^2 & -e_{2,1}G_2/q_2^2 \\ 0 & 0 & 0 & 1 & 0 & 0 & -e_{2,3}G_3/q_3^2 & -e_{2,3}(G_2/q_2^2 + G_3/q_3^2) \\ 0 & 0 & 0 & 0 & 1 & 0 & e_{3,1}(G_1/q_1^2 + G_3/q_3^2) & e_{3,1}G_3/q_3^2 \\ 0 & 0 & 0 & 0 & 0 & 1 & e_{3,2}G_3/q_3^2 & e_{3,2}(G_2/q_2^2 + G_3/q_3^2) \\ 1 & 1 & -1 & 0 & -1 & 0 & 1 & 0 \\ -1 & 0 & 1 & 1 & 0 & -1 & 0 & 1 \end{bmatrix}$$

where $e_{i,j} = mF_iH_i \exp\left(\frac{G_i}{q_i} - \frac{G_j}{q_j}\right)$.

The determinant of the above matrix is given by

$$\begin{aligned}
\text{Determinant} = 1 + (&G_1e_{1,2}q_2^2q_3^2 + G_2e_{1,2}q_1^2q_3^2 + G_1e_{1,3}q_2^2q_3^2 + G_3e_{1,3}q_1^2q_2^2 + G_1e_{2,1}q_2^2q_3^2 \\
&+ G_2e_{2,1}q_1^2q_3^2 + G_2e_{2,3}q_1^2q_3^2 + G_3e_{2,3}q_1^2q_2^2 + G_1e_{3,1}q_2^2q_3^2 + G_3e_{3,1}q_1^2q_2^2 + G_2e_{3,2}q_1^2q_3^2 \\
&+ G_3e_{3,2}q_1^2q_2^2 + G_1G_2e_{1,2}e_{1,3}q_3^2 + G_1G_3e_{1,2}e_{1,3}q_2^2 \\
&+ G_2G_3e_{1,2}e_{1,3}q_1^2 + G_1G_2e_{1,3}e_{2,1}q_3^2 + G_1G_3e_{1,3}e_{2,1}q_2^2 \\
&+ G_2G_3e_{1,3}e_{2,1}q_1^2 + G_1G_2e_{1,2}e_{2,3}q_3^2 + G_1G_3e_{1,2}e_{2,3}q_2^2 \\
&+ G_2G_3e_{1,2}e_{2,3}q_1^2 + G_1G_2e_{1,3}e_{2,3}q_3^2 + G_1G_3e_{1,3}e_{2,3}q_2^2 \\
&+ G_2G_3e_{1,3}e_{2,3}q_1^2 + G_1G_2e_{1,2}e_{3,1}q_3^2 + G_1G_3e_{1,2}e_{3,1}q_2^2 \\
&+ G_2G_3e_{1,2}e_{3,1}q_1^2 + G_1G_2e_{1,2}e_{3,2}q_3^2 + G_1G_2e_{2,1}e_{2,3}q_3^2 \\
&+ G_1G_3e_{1,2}e_{3,2}q_2^2 + G_1G_3e_{2,1}e_{2,3}q_2^2 + G_2G_3e_{1,2}e_{3,2}q_1^2 \\
&+ G_2G_3e_{2,1}e_{2,3}q_1^2 + G_1G_2e_{1,3}e_{3,2}q_3^2 + G_1G_3e_{1,3}e_{3,2}q_2^2 \\
&+ G_2G_3e_{1,3}e_{3,2}q_1^2 + G_1G_2e_{2,1}e_{3,1}q_3^2 + G_1G_3e_{2,1}e_{3,1}q_2^2 \\
&+ G_2G_3e_{2,1}e_{3,1}q_1^2 + G_1G_2e_{2,1}e_{3,2}q_3^2 + G_1G_3e_{2,1}e_{3,2}q_2^2 \\
&+ G_2G_3e_{2,1}e_{3,2}q_1^2 + G_1G_2e_{2,3}e_{3,1}q_3^2 + G_1G_3e_{2,3}e_{3,1}q_2^2 \\
&+ G_2G_3e_{2,3}e_{3,1}q_1^2 + G_1G_2e_{3,1}e_{3,2}q_3^2 + G_1G_3e_{3,1}e_{3,2}q_2^2 \\
&+ G_2G_3e_{3,1}e_{3,2}q_1^2)/(q_1^2q_2^2q_3^2);
\end{aligned}$$

Clearly in the above expression all terms are positive and the very first term $1 > 0$. Thus the Jacobian matrix is non singular, hence proving our result. We can follow similar argument in case with arbitrary g number of genotypes and derive a non singular Jacobian matrix.

5.2 Efficient Approximation of One-step Rate Functional

Now using the intermediary optimal values of mutations derived in previous section, we rewrite the corresponding approximation for the associated rate functional or the cost expression. As done before, first we consider the rate functional for the case where all genotypes in the target histogram G are away from boundary, $G_i \geq \epsilon$.

$$\mathrm{RF} = \sum_{j,k|j\neq k} m_k F_j H_j + \sum_{j,k|j\neq k} r_{j,k}(\log(\rho_{j,k}) - 1) + \sum_{j=1}^{ng} G_j \log \frac{G_j}{p_j}. \tag{5.24}$$

Now, using optimal intermediary values for ρ, r, p

$$\begin{aligned} \rho_{j,k} &= \exp\left(\frac{G_k}{F_k H_k} - \frac{G_j}{F_j H_j}\right), \\ r_{j,k} &= m_k F_j H_j \exp\left(\frac{G_k}{F_k H_k} - \frac{G_j}{F_j H_j}\right), \\ p_j &= \frac{F_j H_j}{\langle F, H\rangle}. \end{aligned}$$

$$\begin{aligned} \mathrm{RF} = &\sum_{j,k|j\neq k} m_k F_j H_j + \sum_{j,k|j\neq k} m_k F_j H_j \left(\frac{G_k}{F_k H_k} - \frac{G_j}{F_j H_j} - 1\right) \exp\left(\frac{G_k}{F_k H_k} - \frac{G_j}{F_j H_j}\right) \\ &+ \log\langle F, H\rangle + \sum_{j=1}^{ng} G_j \log\left(\frac{G_j}{F_j H_j}\right). \end{aligned} \tag{5.25}$$

Extending the cost function so that it reflects the boundary cases, $H_i = 0$ genotypes and restricted mutations we can write it as

$$\begin{aligned}\mathrm{RF} = &\sum_{j \in I, k \in I | j<k} m_k F_j H_j + \sum_{j \in I, k \in I | j<k} m_k F_j H_j \left(\frac{G_k}{F_k H_k} - \frac{G_j}{F_j H_j} - 1 \right) \exp\left(\frac{G_k}{F_k H_k} - \frac{G_j}{F_j H_j} \right) \\ &+ \log\langle F, H \rangle \sum_{j \in J} G_j + \sum_{j \in J} G_j \log\left(\frac{G_j}{F_j H_j} \right). \qquad (5.26)\end{aligned}$$

CHAPTER 6

Optimality Conditions for Rate Minimizing Evolution Trajectories

Consider a trajectory of histograms from initial point a to final stage b in a population with g genotypes with deterministic growth factor F. Let (x, y, z) be any sequence of 3 consecutive points along a trajectory from a to b minimizing the rate functional among all paths going from a to b. Our motivation here is to find an optimal intermediary step y given x and z which minimizes the rate functional. Now, assuming that all the points preceding and following y are fixed along the trajectory we can write

$$\mathrm{RF}(x, y) + \mathrm{RF}(y, z) = f(y)$$

(i.e.) the one-step rate functional (RF) can be considered a function of y assuming x and z are fixed.

So, for a minimal path we must have

$$\min_{\tilde{y}} \mathrm{RF}(x, \tilde{y}) + \mathrm{RF}(\tilde{y}, z) = \mathrm{RF}(x, y) + \mathrm{RF}(y, z).$$

Moreover, since it is a trajectory of histograms, we have $\sum_{i=1}^{g} \tilde{y}_i = 1$. Now, minimizing the cost function as a function of histogram y using Lagrange multipliers we obtain Equation 6.1

$$\frac{\partial \mathrm{RF}(x, y)}{\partial y_j} + \frac{\partial \mathrm{RF}(y, z)}{\partial y_j} = \lambda, \ \forall j = 1, 2, ..., g \tag{6.1}$$

where λ is corresponding Lagrange multiplier and the constraint is

$$\sum_{j=1}^{g} y_j - 1 = 0.$$

6.1 Generic Optimality Conditions

For the trajectory $x \to y \to z$, let $\bar{r}, \bar{p}$ and r, p be the corresponding random matrix of mutations and the intermediary population histogram before dilution for x and y respectively. Also, let $\vec{m} = (m_i)$ be the vector of mutation rates where $m_i = m$, $\forall i$, i.e., we consider that the rates of mutation are equal regardless of the species.

We start with the case of no genotypes on boundary in both the target histograms y and z and will discuss relevant boundary cases in a later section.

Using the expression of cost as derived in Equation 5.25 with approximation for r, ρ, p, we have

$$\begin{aligned}\mathrm{RF}(x,y) = (g-1)m\langle F,x\rangle + m\sum_{j,k|j\neq k} F_j x_j\left(\frac{y_k}{F_k x_k} - \frac{y_j}{F_j x_j} - 1\right)\exp\left(\frac{y_k}{F_k x_k} - \frac{y_j}{F_j x_j}\right)\\ + \sum_{j=1}^{g} y_j \log\left(\frac{y_j\langle F,x\rangle}{F_j x_j}\right), \qquad (6.2)\end{aligned}$$

$$\begin{aligned}\mathrm{RF}(y,z) = (g-1)m\langle F,y\rangle + m\sum_{j,k|j\neq k} F_j y_j\left(\frac{z_k}{F_k y_k} - \frac{z_j}{F_j y_j} - 1\right)\exp\left(\frac{z_k}{F_k y_k} - \frac{z_j}{F_j y_j}\right)\\ + \sum_{j=1}^{g} z_j \log\left(\frac{z_j\langle F,y\rangle}{F_j y_j}\right). \qquad (6.3)\end{aligned}$$

We first we explore the case with zero mutations, $m = 0$. The rate functional or cost becomes

$$\begin{aligned}\mathrm{RF}(x,y) &= \sum_{j=1}^{g} y_j \log\left(\frac{y_j}{F_j x_j/\langle F,x\rangle}\right)\\ &= \sum_{j=1}^{g} y_j \log(y_j) + y_j \log\langle F,x\rangle - y_j \log(F_j x_j)\\ &= \sum_{j=1}^{g} y_j \log(y_j) + \log\langle F,x\rangle - \sum_{j=1}^{g} y_j \log(F_j x_j)\end{aligned}$$

where $\langle F,x\rangle = \sum_{j=1}^{g} F_j x_j$ and $F = (F_j)$ is the vector of deterministic growth factors where $F_1 < F_2 < ... < F_g$.

Similarly,

$$\mathrm{RF}(x,y) = \sum_{j=1}^{g} z_j \log(z_j) + \log\langle F,y\rangle - \sum_{j=1}^{g} z_j \log(F_j y_j)$$

Taking derivative w.r.t y gives

$$\begin{aligned}\frac{\partial \mathrm{RF}(x,y)}{\partial y_j} &= 1+\log(y_j)-\log(F_j x_j),\ \forall j && (6.4)\\ \frac{\partial \mathrm{RF}(y,z)}{\partial y_j} &= -\frac{z_j}{y_j}+\frac{F_j}{\langle F,y\rangle},\ \forall j. && (6.5)\end{aligned}$$

Thus, using Lagrange optimality condition Equation 6.1.

$$\lambda = 1+\log(y_j)-\log(F_j)-\log(x_j)-\frac{z_j}{y_j}+\frac{F_j}{\langle F,y\rangle}, \tag{6.6}$$

which simplifies to

$$\log(x_j) = 1+\log(y_j)-\log(F_j)-\frac{z_j}{y_j}+\frac{F_j}{\langle F,y\rangle}-\lambda.$$

Hence $\log(x_j)$ can be expressed as a function of λ, y, z. Taking exponential of both sides we get

$$x_j = \exp\left(1+\log(y_j)-\log(F_j)-\frac{z_j}{y_j}+\frac{F_j}{\langle F,y\rangle}\right)\exp(-\lambda) = \exp(-\lambda)\Psi_j(y,z) \tag{6.7}$$

where

$$\Psi_j(y,z) = \exp\left(1+\log(y_j)-\log(F_j)-\frac{z_j}{y_j}+\frac{F_j}{\langle F,y\rangle}\right). \tag{6.8}$$

We know that x is a histogram, so

$$\sum_{j=1}^{g} x_j = 1 = \exp(-\lambda)\sum_{j=1}^{g}\Psi_j(y,z),$$

or

$$\exp(\lambda) = \sum_{j=1}^{g}\Psi_j(y,z).$$

Thus, given the histograms z and y, the unknown histogram x must be given by Equation 6.9

$$x_j = \frac{\Psi_j(y,z)}{\sum_{j=1}^{g} \Psi_j(y,z)}, \forall j. \tag{6.9}$$

where Ψ is given by Equation 6.8.

So if we fix a last target state z for any cost minimizing trajectory, we can generate by successive reverse steps the unique cost minimizing trajectory ending with the last two successive points y and z. This will be valid for every choice of the penultimate position histogram y. So, for a fixed z we only need to explore the possible values of the penultimate step y in the state space.

Recall that these formulas are valid for the case where $m = 0$. Now, we come back to the generic case of small but non-zero mutations rate m and calculate the new derivatives of the two-step rate functional using expressions in Equations 6.2 and 6.3.

$$\frac{\partial \mathrm{RF}(x,y)}{\partial y_i} = 1 + \log\left(\frac{y_i}{F_i x_i}\right)$$

$$+ m \sum_{k|k\neq i} F_i x_i \left(\frac{-1}{F_i x_i} \exp\left(\frac{y_k}{F_k x_k} - \frac{y_i}{F_i x_i}\right) + \frac{-1}{F_i x_i}\left(\frac{y_k}{F_k x_k} - \frac{y_i}{F_i x_i} - 1\right) \exp\left(\frac{y_k}{F_k x_k} - \frac{y_i}{F_i x_i}\right)\right)$$

$$+ m \sum_{j|j\neq i} F_j x_j \left(\frac{1}{F_i x_i} \exp\left(\frac{y_i}{F_i x_i} - \frac{y_j}{F_j x_j}\right) + \frac{1}{F_i x_i}\left(\frac{y_i}{F_i x_i} - \frac{y_j}{F_j x_j} - 1\right) \exp\left(\frac{y_i}{F_i x_i} - \frac{y_j}{F_j x_j}\right)\right)$$

and

$$\frac{\partial \mathrm{RF}(y,z)}{\partial y_i} = (g-1)mF_i+\frac{F_i}{\langle F,y\rangle}-\frac{z_i}{y_i}-mF_i$$

$$+\, m\sum_{k\neq i}\left(\left(\frac{F_i z_k}{F_k y_k}-F_i\right)\exp(-b_{i,k})+\left(\frac{F_i y_i z_k}{F_k y_k}-z_i-F_i y_i\right)\left(\frac{z_i}{F_i y_i^2}\right)\exp(-b_{i,k})\right)$$
$$+\, m\sum_{j\neq i}\left(\left(-\frac{F_j y_j z_i}{F_i y_i^2}\right)\exp(b_{i,j})+\left(\frac{F_j y_j z_i}{F_i y_i}-z_j+F_j y_j\right)\left(\frac{-z_i}{F_i y_i^2}\right)\exp(b_{i,j})\right)$$

Simplifying, we obtain

$$\frac{\partial \mathrm{RF}(x,y)}{\partial y_i} = 1+\log\left(\frac{y_i}{F_i x_i}\right)+m\sum_{j|j\neq i} a_{i,j}\exp(-a_{i,j})+\frac{F_j x_j}{F_i x_i}a_{i,j}\exp(a_{i,j})$$

$$\frac{\partial \mathrm{RF}(y,z)}{\partial y_i} = (g-1)mF_i+\frac{F_i}{\langle F,y\rangle}-\frac{z_i}{y_i}-mF_i$$
$$+\, m\sum_{k|k\neq i}\left(\frac{F_i z_k}{F_k y_k}-F_i+\frac{z_k z_i}{F_k y_k y_i}-\frac{z_i^2}{F_i y_i^2}-\frac{z_i}{y_i}\right)\exp(-b_{i,k})$$
$$+\, m\sum_{k|k\neq i}\left(-\frac{F_k y_k z_i^2}{F_i^2 y_i^3}+\frac{z_i z_k}{F_i y_i^2}\right)\exp(b_{i,k})$$

where

$$a_{i,j}=\frac{y_i}{F_i x_i}-\frac{y_j}{F_j x_j}.$$

and

$$b_{i,j}=\frac{z_i}{F_i y_i}-\frac{z_j}{F_j y_j}.$$

Now, using Lagrange optimality conditions from Equation 6.1 we have

$$\lambda=\frac{\partial \mathrm{RF}(x,y)}{\partial y_i}+\frac{\partial \mathrm{RF}(y,z)}{\partial y_i},\ \forall i$$

$$
\begin{aligned}
\lambda = 1 + \log\left(\frac{y_i}{F_i x_i}\right) + (g-1)mF_i + \frac{F_i}{\langle F, y\rangle} - \frac{z_i}{y_i} - mF_i \\
+ m \sum_{j|j\neq i} a_{i,j}\exp(-a_{i,j}) + \frac{F_j x_j}{F_i x_i} a_{i,j}\exp(a_{i,j}) \\
+ m \sum_{j|j\neq i} \left(\frac{F_i z_j}{F_j y_j} - F_i + \frac{z_j z_i}{F_j y_j y_i} - \frac{z_i^2}{F_i y_i^2} - \frac{z_i}{y_i}\right)\exp(-b_{i,j}) \\
+ m \sum_{j|j\neq i} \left(-\frac{F_j y_j z_i^2}{F_i^2 y_i^3} + \frac{z_i z_j}{F_i y_i^2}\right)\exp(b_{i,j})
\end{aligned}
$$

The above equations give an implicit system of equations for x along with the condition that $\sum_i x_i = 1$.

So we have the following system of implicit equations in x given y, z

$$\begin{aligned}\lambda &= 1 + \log\left(\frac{y_i}{F_i x_i}\right) + (g-2)mF_i + \frac{F_i}{\langle F, y\rangle} - \frac{z_i}{y_i}\\ &+ m\sum_{j|j\neq i} a_{i,j}\exp(-a_{i,j}) + \frac{F_j x_j}{F_i x_i} a_{i,j}\exp(a_{i,j})\\ &+ m\sum_{j|j\neq i}\left(\frac{F_i z_j}{F_j y_j} - F_i + \frac{z_j z_i}{F_j y_j y_i} - \frac{z_i^2}{F_i y_i^2} - \frac{z_i}{y_i}\right)\exp(-b_{i,j})\\ &+ m\sum_{j|j\neq i}\left(-\frac{F_j y_j z_i^2}{F_i^2 y_i^3} + \frac{z_i z_j}{F_i y_i^2}\right)\exp(b_{i,j})\end{aligned}$$

If we compare the derivatives of the two-step rate functional for the case with $m = 0$ with the derivatives of two-step rate functional with random mutations, we are led to introducing two terms of order 0 in m denoted by $U(x, y), \tilde{U}(y, z)$ where

$$U_i(x, y) = 1 + \log\frac{y_i}{F_i x_i}, \tag{6.10}$$

$$\tilde{U}_i(y, z) = \frac{-z_i}{y_i} + \frac{F_i}{\langle F, y\rangle} \tag{6.11}$$

Call $V(x, y), \tilde{V}(x, y)$ to be the coefficients of the first-order terms in m where

$$V_i(x, y) = \sum_{j|j\neq i} a_{i,j}\exp(-a_{i,j}) + \frac{F_j x_j}{F_i x_i} a_{i,j}\exp(a_{i,j}), \tag{6.12}$$

$$\begin{aligned}\tilde{V}_i(y, z) &= (g-2)F_i + \sum_{k|k\neq i}\left(\frac{F_i z_k}{F_k y_k} - F_i + \frac{z_k z_i}{F_k y_k y_i} - \frac{z_i^2}{F_i y_i^2} - \frac{z_i}{y_i}\right)\exp(-b_{i,k})\\ &+ \sum_{k|k\neq i}\left(-\frac{F_k y_k z_i^2}{F_i^2 y_i^3} + \frac{z_i z_k}{F_i y_i^2}\right)\exp(b_{i,k})\end{aligned} \tag{6.13}$$

and thus we have

$$\frac{\partial \mathrm{RF}(x, y)}{\partial y_i} = U_i(x, y) + mV_i(x, y), \tag{6.14}$$

$$\frac{\partial \mathrm{RF}(y, z)}{\partial y_i} = \tilde{U}_i(y, z) + m\tilde{V}_i(y, z). \tag{6.15}$$

Then,

$$\lambda = U_i(x, y) + \tilde{U}_i(y, z) + m(V_i(x, y) + \tilde{V}_i(y, z)). \tag{6.16}$$

where

$$\lambda(0) = U_i(x, y) + \tilde{U}_i(y, z). \tag{6.17}$$

is the system of equations for the case $m = 0$. We have an explicit solution as derived in Equation 6.9 for such a case. Let the solution for $m = 0$ be $x(0), \lambda(0)$.

Let $x(1)$ and $\lambda(1)$ be the first-order approximations of the solution of the system given by 6.16.

$$x_i(1) = x_i(0) + mX_i, \ \ \forall i$$

and

$$\lambda(1) = \lambda(0) + m\Lambda$$

Substituting into Equation 6.16 we get,

$$\lambda(0) + m\Lambda = U_i(x(1), y) + \tilde{U}_i(y, z) + m(V_i(x(0), y) + \tilde{V}_i(y, z)). \tag{6.18}$$

Now, using Taylor expansion for $U_i(x(1), y)$, we have

$$U_i(x(1), y) = U_i(x(0) + mX, y) = U_i(x(0), y) + m \sum_j X_j \frac{\partial U_i}{\partial x_j}$$

Replacing it back we get

$$\lambda(0) + m\Lambda = U_i(x(0), y) + m \sum_j X_j \frac{\partial U_i}{\partial x_j} + \tilde{U}_i(y, z) + m(V_i(x(0), y) + \tilde{V}_i(y, z)).$$

Now $x(0), \lambda(0)$ is a solution of the Equation 6.17, so

$$\lambda(0) = U_i(x(0), y) + \tilde{U}_i(y, z)$$

Hence, we need to solve the following system for value of X

$$\Lambda = \sum_j X_j \frac{\partial U_i}{\partial x_j} + V_i(x(0), y) + \tilde{V}_i(y, z). \tag{6.19}$$

Also, since $\sum_i x_i(1) = \sum_i x_i(0) = 1$, we have

$$\sum_i X_i = 0$$

Moreover, using the fact that $U_i(x, y) = 1 + \log(y_i) - \log(F_i x_i)$, we get

$$\frac{\partial U_i}{\partial x_j} = \begin{cases} 0 & i \neq j \\ \frac{-1}{x_i(0)} & i = j \end{cases}$$

so the system of equations for X transform to,

$$\Lambda = \frac{-X_i}{x_i(0)} + V_i(x(0), y) + \tilde{V}_i(y, z).$$

or,

$$X_i \;=\; x_i(0)(-\Lambda + V_i(x(0), y) + \tilde{V}_i(y, z)) \tag{6.20}$$

and using the fact that $\sum_i X_i = 0$, we have

$$\begin{aligned} 0 = \sum_i X_i &= \sum_i x_i(0)(-\Lambda + V_i(x(0), y) + \tilde{V}_i(y, z)), \\ 0 &= -\sum_i x_i(0)\Lambda + \sum_i x_i(0)(V_i(x(0), y) + \tilde{V}_i(y, z)), \\ 0 &= -\Lambda + \sum_i x_i(0)(V_i(x(0), y) + \tilde{V}_i(y, z)), \\ \Lambda &= \sum_i x_i(0)(V_i(x(0), y) + \tilde{V}_i(y, z)) \end{aligned}$$

Replacing the value of Λ in Equation 6.20, we obtain

$$X_i = -x_i(0) \sum_{j \neq i} x_j(0)(V_j(x(0), y) + \tilde{V}_j(y, z)) + x_i(0)(1 - x_i(0))(V_i(x(0), y) + \tilde{V}_i(y, z)). \tag{6.21}$$

and so finally we have

$$x_i(1) = x_i(0) + mX_i \tag{6.22}$$

as the solution where $x_i(0)$ is given by Equation 6.9 and X is given by Equation 6.21.

6.2 Iterative Procedure

In the evolution trajectory, $x \longrightarrow y \longrightarrow z$ assuming

$$h_n = z,\ h_{n-1} = y,\ h_{n-2} = x$$

or as the population histograms at $n, n-1$ and $n-2$ steps respectively, we can rewrite expression in the Equation 6.9 as

$$h_{n-2}(j) = \frac{\Psi_j(h_{n-1}, h_n)}{\sum_{j=1}^{g} \Psi_j(h_{n-1}, h_n)}, \forall j. \tag{6.23}$$

where Ψ is given by

$$\Psi_j(h_{n-1}, h_n) = \exp\left(1 + \log h_{n-1}(j) - \log F(j) - \frac{h_n(j)}{h_{n-1}(j)} + \frac{F(j)}{\langle F, h_{n-1}\rangle}\right). \tag{6.24}$$

The above gives us a reverse iterative scheme to estimate the optimal histogram h_{n-2} given histograms h_{n-1} and h_n when $m = 0$.

The value of n is not fixed in our model and we keep building the reverse trajectory with the iterative formula in 6.23 which uses the next 2 steps in the chain as input.

At the first step we fix $z = h_n$ as the target histogram and $y = h_{n-1}$ as the penultimate step. Then we iterate and estimate optimal $x = h_{n-2}$. Then h_n and h_{n-1} are updated to be y and x respectively. Thus we continue the procedure and estimate new h_{n-2} in the optimal rate minimizing trajectory.

This procedure is repeated until the histograms get close to the boundary or reach a corner since our formulas are for interior points only.

This iterative scheme also works for the population trajectories with non zero random mutation matrix. The corresponding formulas are derived at the end of previous section 6.1.

6.3 Optimality Conditions for Restricted Mutations

For the trajectory $x \to y \to z$, let $\bar{r}, \bar{p}$ and r, p be the corresponding random matrix of restricted mutations and the intermediary population histogram before dilution for x and y respectively. Also, let $\vec{m} = (m_i)$ be the vector of mutation rates where $m_i = m$, $\forall i$, i.e., we consider that the rates of mutation are equal regardless of the species.

Again, using the expression for rate functional as derived in Equation 5.26 with approximation for r, ρ, p, we have

$$\begin{aligned}\mathrm{RF}(x,y) = m\sum_{j=1}^{g}(g-j)F_jx_j + m\sum_{j,k|j<k} F_jx_j\left(\frac{y_k}{F_kx_k} - \frac{y_j}{F_jx_j} - 1\right)\exp\left(\frac{y_k}{F_kx_k} - \frac{y_j}{F_jx_j}\right)\\ + \sum_{j=1}^{g} y_j \log\left(\frac{y_j\langle F, x\rangle}{F_jx_j}\right). \qquad (6.25)\end{aligned}$$

$$\begin{aligned}\mathrm{RF}(y,z) = m\sum_{j=1}^{g}(g-j)F_jy_j + m\sum_{j,k|j<k} F_jy_j\left(\frac{z_k}{F_ky_k} - \frac{z_j}{F_jy_j} - 1\right)\exp\left(\frac{z_k}{F_ky_k} - \frac{z_j}{F_jy_j}\right)\\ + \sum_{j=1}^{g} z_j \log\left(\frac{z_j\langle F, y\rangle}{F_jy_j}\right). \qquad (6.26)\end{aligned}$$

Recomputing the expressions for the optimal value of x given y and z using the Lagrange parameter λ as explained in section 6.1 we derive the following formulas.

$$\begin{aligned}
\lambda = 1 + \log\left(\frac{y_i}{F_i x_i}\right) + (g-i)mF_i + \frac{F_i}{\langle F, y\rangle} - \frac{z_i}{y_i} \\
+ m \sum_{j|j>i} a_{i,j} \exp(-a_{i,j}) + m \sum_{j|j<i} \frac{F_j x_j}{F_i x_i} a_{i,j} \exp(a_{i,j}) \\
+ m \sum_{j|j>i} \left(\frac{F_i z_j}{F_j y_j} - F_i + m\frac{z_j z_i}{F_j y_j y_i} - \frac{z_i^2}{F_i y_i^2} - \frac{z_i}{y_i}\right) \exp(-b_{i,j}) \\
+ m \sum_{j|j<i} \left(-\frac{F_j y_j z_i^2}{F_i^2 y_i^3} + \frac{z_i z_j}{F_i y_i^2}\right) \exp(b_{i,j})
\end{aligned}$$

And similar to the case of general random mutations in section 6.1, we obtain

$$x_i(1) = x_i(0) + mX_i \tag{6.27}$$

as the solution where $x_i(0)$ is given by Equation 6.9 and X is given by Equation 6.21 where

$$\begin{aligned}
U_i(x, y) &= 1 + \log \frac{y_i}{F_i x_i}, &(6.28)\\
\tilde{U}_i(y, z) &= \frac{-z_i}{y_i} + \frac{F_i}{\langle F, y\rangle} &(6.29)
\end{aligned}$$

$$V_i(x, y) = \sum_{j|j>i} a_{i,j} \exp(-a_{i,j}) + \sum_{j|j<i} \frac{F_j x_j}{F_i x_i} a_{i,j} \exp(a_{i,j}), \tag{6.30}$$

and

$$\begin{aligned}
\tilde{V}_i(y, z) = (g-i)F_i + \sum_{k|k>i} \left(\frac{F_i z_k}{F_k y_k} - F_i + \frac{z_k z_i}{F_k y_k y_i} - \frac{z_i^2}{F_i y_i^2} - \frac{z_i}{y_i}\right) \exp(-b_{i,k}) \\
+ \sum_{k|k<i} \left(-\frac{F_k y_k z_i^2}{F_i^2 y_i^3} + \frac{z_i z_k}{F_i y_i^2}\right) \exp(b_{i,k}). \qquad (6.31)
\end{aligned}$$

6.4 Optimality Conditions for Boundary Targets

In this section we consider the special case of restricted random mutations in the trajectory and target histograms to be on boundary. Here the boundary cases are only explored for $g = 3$ genotypes in the system. case.

We use the same strategy as developed before for the case of non boundary targets. We determine the corresponding functions $U, \tilde{U}, V, \tilde{V}$ and then solve to get the value for $x_i(1)$ using $x_i(0)$.

We use the rate functional or cost expression given by Equation 5.26. The rate functional is as follows

$$\begin{aligned}\mathrm{RF(H,G)} = &\sum_{j\in I,k\in I|j<k} m_k F_j H_j + \sum_{j\in I,k\in I|j<k} m F_j H_j \left(\frac{G_k}{F_k H_k} - \frac{G_j}{F_j H_j} - 1\right) \exp\left(\frac{G_k}{F_k H_k} - \frac{G_j}{F_j H_j}\right)\\ &+ \log\langle F, H\rangle \sum_{j\in J} G_j + \sum_{j\in J} G_j \log\left(\frac{G_j}{F_j H_j}\right).\end{aligned}$$

where J is the set of genotypes which are away from boundary in G and I is the set of genotypes for which H is non zero (i.e.), $H_i \neq 0, \forall i \in I$. Since $I \subset J$, for notational convenience we rewrite the rate functional as

$$\begin{aligned}\mathrm{RF(H,G)} = &\sum_{j,k|j<k} m_k F_j H_j \left(\frac{G_k}{F_k H_k} - \frac{G_j}{F_j H_j} - 1\right) \exp\left(\frac{G_k}{F_k H_k} - \frac{G_j}{F_j H_j}\right)\\ &+ m\sum_{j=1}^{g}(g-j)F_j x_j + \log\langle F, H\rangle \sum_{j\in J} G_j + \sum_{j\in J} G_j \log\left(\frac{G_j}{F_j H_j}\right).\end{aligned}$$

Similar to our approach for interior points, we will derive expressions for the case $m = 0$ and then develop formulas for the case with non-zero mutations. Again, let B be the set of genotypes that satisfy boundary condition as outlined before and rest

J be the set of genotypes away from boundary such that $B \cup J = [1:g]$.

The feasible cases to be considered while discussing boundary cases are outlined next.

1. In the chain $x \to y \to z$, only y is on boundary and without loss of generality let $J_y = \{1,2\}$ and $B_y = \{3\}$. Then we get for $m = 0$,

$$\begin{aligned}
\mathrm{RF}(x,y) &= \sum_{j \in J_y} y_j \log\left(\frac{y_j}{F_j x_j / \langle F, x\rangle}\right) \\
&= \sum_{j \in J_y} y_j \log y_j + y_j \log\langle F, x\rangle - y_j \log F_j x_j \\
&= \sum_{j \in J_y} y_j \log y_j + (1 - \sum_{j \in B_y} y_j) \log\langle F, x\rangle - \sum_{j \in J_y} y_j \log F_j x_j .
\end{aligned}$$

$$\begin{aligned}
\mathrm{RF}(y,z) &= \sum_{j=1}^{3} z_j \log\left(\frac{z_j}{F_j y_j / \langle F, y\rangle}\right) \\
&= \sum_{j=1}^{3} z_j \log(z_j) + \log\langle F, y\rangle - \sum_{j=1}^{3} z_j \log(F_j y_j).
\end{aligned}$$

This gives

$$\frac{\partial \mathrm{RF}(x,y)}{\partial y_j} = \begin{cases} -\log\langle F, x\rangle & j \notin J_y \\ 1 + \log y_j - \log F_j x_j & j \in J_y \end{cases}$$

and

$$\frac{\partial \mathrm{RF}(y,z)}{\partial y_j} = -\frac{z_j}{y_j} + \frac{F_j}{\langle F, y\rangle}$$

Thus using Lagrange optimality conditions in 6.1

$$\lambda = \begin{cases} -\log(\langle F, x\rangle) - \frac{z_j}{y_j} + \frac{F_j}{\langle F, y\rangle} & j \notin J_y \\ 1 + \log(y_j) - \log(F_j x_j) - \frac{z_j}{y_j} + \frac{F_j}{\langle F, y\rangle} & j \in J_y \end{cases}$$

Since $J_y = \{1,2\}$ and $B_y = \{3\}$, then

$$\lambda = -\log\langle F, x\rangle - \frac{z_3}{y_3} + \frac{F_3}{\langle F, y\rangle}$$

which gives

$$\langle F, x\rangle = \exp(-\lambda)\exp\left(-\frac{z_3}{y_3} + \frac{F_3}{\langle F, y\rangle}\right) \tag{6.32}$$

Also, we have

$$\log(x_j) = 1 + \log(y_j) - \log(F_j) - \frac{z_j}{y_j} + \frac{F_j}{\langle F, y\rangle} - \lambda,\ j = 1,2$$

Taking exponential on both sides we get for $j = 1, 2$

$$\begin{aligned} x_j &= \exp\left(1 + \log(y_j) - \log(F_j) - \frac{z_j}{y_j} + \frac{F_j}{\langle F, y\rangle}\right)\exp(-\lambda), \\ &= \exp(-\lambda)\Psi_j(y, z) \end{aligned} \tag{6.33}$$

where

$$\Psi_j(y, z) = \exp\left(1 + \log(y_j) - \log(F_j) - \frac{z_j}{y_j} + \frac{F_j}{\langle F, y\rangle}\right),\ j = 1, 2. \tag{6.34}$$

Thus, from Equation 6.32

$$F_1x_1 + F_2x_2 + F_3x_3 = \exp(-\lambda)\exp\left(-\frac{z_3}{y_3} + \frac{F_3}{\langle F, y\rangle}\right),$$

$$F_1\exp(-\lambda)\Psi_1(y, z) + F_2\exp(-\lambda)\Psi_2(y, z) + F_3x_3 = \exp(-\lambda)\exp\left(-\frac{z_3}{y_3} + \frac{F_3}{\langle F, y\rangle}\right)$$

or,

$$x_3 = \exp(-\lambda)\left(\frac{1}{F_3}\exp\left(-\frac{z_3}{y_3} + \frac{F_3}{\langle F, y\rangle}\right) - \frac{F_1}{F_3}\Psi_1(y, z) - \frac{F_2}{F_3}\Psi_2(y, z)\right)$$

Using $\sum_{j=1}^{3} x_j = 1$ we get

$$\exp(\lambda) = \left(\frac{1}{F_3}\exp\left(-\frac{z_3}{y_3} + \frac{F_3}{\langle F, y\rangle}\right) - \sum_{i=1}^{2}\frac{F_i}{F_3}\Psi_i(y,z) + \sum_{i=1}^{2}\Psi_i(y,z)\right) \tag{6.35}$$

thus giving us explicit formulas for x_i which depend only on y, z as follows

$$x_1 = \exp(-\lambda)\Psi_1(y,z), \tag{6.36}$$

$$x_2 = \exp(-\lambda)\Psi_2(y,z), \tag{6.37}$$

$$x_3 = \exp(-\lambda)\left(\frac{1}{F_3}\exp\left(-\frac{z_3}{y_3} + \frac{F_3}{\langle F, y\rangle}\right) - \frac{F_1}{F_3}\Psi_1(y,z) - \frac{F_2}{F_3}\Psi_2(y,z)\right) \tag{6.38}$$

with Ψ given by 6.34 and λ given by 6.35.

Now, bringing back random mutations into the equations, we compute once again the functions $U(x,y), \tilde{U}(y,z), V(x,y)$ and $\tilde{V}(y,z)$.

$$U_i(x,y) = \begin{cases} -\log\langle F, x\rangle & i = 3 \\ 1 + \log\frac{y_i}{F_i x_i} & i \neq 3 \end{cases}$$

$$\tilde{U}_i(y,z) = \frac{-z_i}{y_i} + \frac{F_i}{\langle F, y\rangle}$$

$$V_i(x,y) = \sum_{j|j>i} a_{i,j}\exp(-a_{i,j}) + \sum_{j|j<i}\frac{F_j x_j}{F_i x_i} a_{i,j}\exp(a_{i,j}),$$

and

$$\tilde{V}_i(y,z) = (g-i)F_i + \sum_{k|k>i}\left(\frac{F_i z_k}{F_k y_k} - F_i + \frac{z_k z_i}{F_k y_k y_i} - \frac{z_i^2}{F_i y_i^2} - \frac{z_i}{y_i}\right)\exp(-b_{i,k})$$
$$+ \sum_{k|k<i}\left(-\frac{F_k y_k z_i^2}{F_i^2 y_i^3} + \frac{z_i z_k}{F_i y_i^2}\right)\exp(b_{i,k})$$

where $a_{i,j} = \frac{y_i}{F_i x_i} - \frac{y_j}{F_j x_j}$ and $b_{i,j} = \frac{z_i}{F_i y_i} - \frac{z_j}{F_j y_j}$.

This gives

$$\frac{\partial U_i}{\partial x_j} = \begin{cases} \frac{-F_i}{\langle F, x(0)\rangle} & i = 3 \\ 0 & i \neq 3, i \neq j \\ \frac{-1}{x_i(0)} & i \neq 3, i = j \end{cases}$$

with $x(0)$ given by Equations 6.36, 6.37 and 6.38. So,

$$\Lambda = \begin{cases} \frac{-F_i X_i}{\langle F, x(0)\rangle} + V_i(x(0), y) + \tilde{V}_i(y, z) & i = 3 \\ \\ \frac{-X_i}{x_i(0)} + V_i(x(0), y) + \tilde{V}_i(y, z) & i \neq 3 \end{cases}$$

and hence

$$X_i = \begin{cases} \frac{\langle F, x(0)\rangle}{F_i}(-\Lambda + V_i(x(0), y) + \tilde{V}_i(y, z)) & i = 3 \\ \\ x_i(0)(-\Lambda + V_i(x(0), y) + \tilde{V}_i(y, z)) & i \neq 3 \end{cases}$$

Now, using the fact that $\sum_i X_i = 0$, we have

$$\sum_{i=1}^{2} x_i(0)(-\Lambda + V_i(x(0), y) + \tilde{V}_i(y, z)) + \frac{\langle F, x(0)\rangle}{F_3}(-\Lambda + V_3(x(0), y) + \tilde{V}_3(y, z)) = 0,$$

Hence we get the value of Λ as follows

$$\Lambda = \frac{\sum_{i=1}^{2} x_i(0)(V_i(x(0), y) + \tilde{V}_i(y, z)) + \frac{\langle F, x(0)\rangle}{F_3}(V_3(x(0), y) + \tilde{V}_3(y, z))}{\sum_{i=1}^{2} x_i(0) + \frac{\langle F, x(0)\rangle}{F_3}},$$

and thus

$$X_i = \begin{cases} \frac{\langle F, x(0)\rangle}{F_i}(-\Lambda + V_i(x(0), y) + \tilde{V}_i(y, z)) & i = 3 \\ \\ x_i(0)(-\Lambda + V_i(x(0), y) + \tilde{V}_i(y, z)) & i \neq 3 \end{cases}$$

where value of Λ is as derived above. So we have

$$x_i(1) = x_i(0) + mX_i$$

as the new solution.

2. In the chain $x \to y \to z$, both y, z are on boundary and $J_y = J_z$ with $\#J_y = 2$. Without loss of generality, let $J_y = J_z = \{1, 2\}$ and $B_y = B_z = \{3\}$. Then we get that

   Then proceeding as before we get

$$\begin{aligned}\mathrm{RF}(x, y) &= \sum_{j \in J_y} y_j \log\left(\frac{y_j}{F_j x_j / \langle F, x\rangle}\right)\\ &= \sum_{j \in J_y} y_j \log y_j + (1 - \sum_{j \notin J_y} y_j) \log\langle F, x\rangle - \sum_{j \in J_y} y_j \log F_j x_j.\end{aligned}$$

$$\begin{aligned}\mathrm{RF}(y, z) &= \sum_{j \in J_z} z_j \log\left(\frac{z_j}{F_j y_j / \langle F, y\rangle}\right)\\ &= \sum_{j \in J_z} z_j \log z_j + z_j \log\langle F, y\rangle - z_j \log F_j y_j.\end{aligned}$$

   This gives

$$\frac{\partial \mathrm{RF}(x, y)}{\partial y_j} = \begin{cases} -\log\langle F, x\rangle & j \notin J_y \\ 1 + \log y_j - \log F_j x_j & j \in J_y \end{cases}$$

   and

$$\frac{\partial \mathrm{RF}(y, z)}{\partial y_j} = \begin{cases} \frac{F_j}{\langle F,y\rangle} \sum_{i \in J_z} z_i & j \notin J_z \\ -\frac{z_j}{y_j} + \frac{z_j F_j}{\langle F,y\rangle} & j \in J_z \end{cases}$$

   Thus,

$$\lambda = \begin{cases} -\log\langle F, x\rangle + \frac{F_j}{\langle F,y\rangle} \sum_{i \in J_z} z_i & j \notin J_y, j \notin J_z \\ -\log\langle F, x\rangle - \frac{z_j}{y_j} + \frac{z_j F_j}{\langle F,y\rangle} & j \notin J_y, j \in J_z \\ 1 + \log(y_j) - \log(F_j x_j) + \frac{F_j}{\langle F,y\rangle} \sum_{i \in J_z} z_i & j \in J_y, j \notin J_z \\ 1 + \log(y_j) - \log(F_j x_j) - \frac{z_j}{y_j} + \frac{z_j F_j}{\langle F,y\rangle} & j \in J_y, j \in J_z \end{cases}$$

using Lagrange optimality conditions in 6.1. Using $J_y = J_z = \{1, 2\}$ and $B_y = B_z = \{3\}$ in the above system gives

$$\lambda = -\log\langle F, x\rangle + \frac{F_3}{\langle F, y\rangle}\sum_{i \in J_z} z_i,$$

or

$$\langle F, x\rangle = \exp(-\lambda)\exp\left(\frac{F_3}{\langle F, y\rangle}\sum_{i \in J_z} z_i\right). \tag{6.39}$$

For, $j = 1, 2$, we have as before

$$\begin{aligned} x_j &= \exp(-\lambda)\exp\left(1 + \log y_j - \frac{z_j}{y_j} + \frac{z_j F_j}{\langle F, y\rangle} - \log F_j\right), \\ &= \exp(-\lambda)\Psi_j(y, z) \end{aligned}$$

where

$$\Psi_j(y, z) = \exp\left(1 + \log y_j - \frac{z_j}{y_j} + \frac{z_j F_j}{\langle F, y\rangle} - \log F_j\right),\ j = 1, 2. \tag{6.40}$$

Now

$$\begin{aligned} \langle F, x\rangle &= F_1 x_1 + F_2 x_2 + F_3 x_3 = \exp(-\lambda)\exp\left(\frac{F_3}{\langle F, y\rangle}\sum_{i \in J_z} z_i\right), \\ &= F_1\exp(-\lambda)\Psi_1 + F_2\exp(-\lambda)\Psi_2 + F_3 x_3, \\ x_3 &= \exp(-\lambda)\left(\exp\left(\frac{F_3}{\langle F, y\rangle}\sum_{i \in J_z} z_i\right) - F_1\Psi_1 - F_2\Psi_2\right)/F_3, \end{aligned}$$

Finally, using the fact that $\sum_j x_j = 1$, we get that

$$\exp(\lambda) = \exp\left(\frac{F_3}{\langle F, y\rangle}\sum_{i \in J_z} z_i\right) + \left(1 - \frac{F_1}{F_3}\right)\Psi_1 + \left(1 - \frac{F_2}{F_3}\right)\Psi_2. \tag{6.41}$$

So that we have deterministic values of x_i as follows

$$x_1 = \exp(-\lambda)\Psi_1, \tag{6.42}$$

$$x_2 = \exp(-\lambda)\Psi_2, \tag{6.43}$$

$$x_3 = \exp(-\lambda)\left(\exp\left(\frac{F_3}{\langle F, y\rangle}\sum_{i\in J_z} z_i\right) - F_1\Psi_1 - F_2\Psi_2\right)/F_3. \tag{6.44}$$

where Ψ and λ are given by Equations 6.40 and 6.41.

Bringing back random mutations into the equations, we compute once again the functions $U(x,y), \tilde{U}(y,z), V(x,y)$ and $\tilde{V}(y,z)$.

$$U_i(x,y) = \begin{cases} -\log\langle F, x\rangle & i = 3 \\ 1 + \log\frac{y_i}{F_i x_i} & i \neq 3 \end{cases}$$

$$\tilde{U}_i(x,y) = \begin{cases} \frac{F_i(z_1+z_2)}{\langle F,y\rangle} & i = 3 \\ \frac{-z_i}{y_i} + \frac{z_i F_i}{\langle F,y\rangle} & i \neq 3 \end{cases}$$

$$V_i(x,y) = \sum_{j,j>i} a_{i,j}\exp(-a_{i,j}) + \sum_{j,j<i}\frac{F_j x_j}{F_i x_i} a_{i,j}\exp(a_{i,j}),$$

$$\tilde{V}_i(y,z) = (g-i)F_i + \sum_{k|k>i}\left(\frac{F_i z_k}{F_k y_k} - F_i + \frac{z_k z_i}{F_k y_k y_i} - \frac{z_i^2}{F_i y_i^2} - \frac{z_i}{y_i}\right)\exp(-b_{i,k})$$
$$+ \sum_{k|k<i}\left(-\frac{F_k y_k z_i^2}{F_i^2 y_i^3} + \frac{z_i z_k}{F_i y_i^2}\right)\exp(b_{i,k})$$

where $a_{i,j} = \frac{y_i}{F_i x_i} - \frac{y_j}{F_j x_j}$ and $b_{i,j} = \frac{z_i}{F_i y_i} - \frac{z_j}{F_j y_j}$. This gives

$$\frac{\partial U_i}{\partial x_j} = \begin{cases} \frac{-F_i}{\langle F, x(0)\rangle} & i = 3 \\ 0 & i \neq 3, i \neq j \\ \frac{-1}{x_i(0)} & i \neq 3, i = j \end{cases}$$

with $x(0)$ given by Equations 6.42, 6.43 and 6.44. So,

$$\Lambda = \begin{cases} \frac{-F_i X_i}{\langle F, x(0)\rangle} + V_i(x(0), y) + \tilde{V}_i(y, z) & i = 3 \\ \\ \frac{-X_i}{x_i(0)} + V_i(x(0), y) + \tilde{V}_i(y, z) & i \neq 3 \end{cases}$$

and hence

$$X_i = \begin{cases} \frac{\langle F, x(0)\rangle}{F_i}(-\Lambda + V_i(x(0), y) + \tilde{V}_i(y, z)) & i = 3 \\ \\ x_i(0)(-\Lambda + V_i(x(0), y) + \tilde{V}_i(y, z)) & i \neq 3 \end{cases}$$

and using the fact that $\sum_i X_i = 0$, we have

$$\sum_{i=1}^{2} x_i(0)(-\Lambda + V_i(x(0), y) + \tilde{V}_i(y, z)) + \frac{\langle F, x(0)\rangle}{F_3}(-\Lambda + V_3(x(0), y) + \tilde{V}_3(y, z)) = 0,$$

Hence we get the value of Λ as follows

$$\Lambda = \frac{\sum_{i=1}^{2} x_i(0)(V_i(x(0), y) + \tilde{V}_i(y, z)) + \frac{\langle F, x(0)\rangle}{F_3}(V_3(x(0), y) + \tilde{V}_3(y, z))}{\sum_{i=1}^{2} x_i(0) + \frac{\langle F, x(0)\rangle}{F_3}},$$

and thus

$$X_i = \begin{cases} \frac{\langle F, x(0)\rangle}{F_i}(-\Lambda + V_i(x(0), y) + \tilde{V}_i(y, z)) & i = 3 \\ \\ x_i(0)(-\Lambda + V_i(x(0), y) + \tilde{V}_i(y, z)) & i \neq 3 \end{cases}$$

where value of Λ is as derived above. So we have

$$x_i(1) = x_i(0) + mX_i$$

as the new solution.

3. In the chain $x \to y \to z$, both y, z are on boundary and $J_z \in J_y$ with $\#J_y = 2$ and $\#J_z = 1$. Without loss of generality let $J_y = \{1, 2\}$ and $J_z = \{1\}$, and so $B_y = \{3\}$ and $B_z = \{2, 3\}$. Then we get that

$$\lambda = \begin{cases} -\log\langle F, x\rangle + \frac{F_3}{\langle F,y\rangle}\sum_{i\in J_z} z_i & j = 3 \\ \\ 1 + \log y_2 - \log F_2 x_2 + \frac{F_2}{\langle F,y\rangle}\sum_{i\in J_z} z_i & j = 2 \\ \\ 1 + \log y_2 - \log F_2 x_2 - \frac{z_2}{y_2} + \frac{z_2 F_2}{\langle F,y\rangle} & j = 1 \end{cases}$$

using Lagrange optimality conditions 6.1.

Thus we have,

$$\langle F, x\rangle = \exp(-\lambda)\exp\left(\frac{F_3}{\langle F, y\rangle}\sum_{i\in J_z} z_i\right),$$

$$x_1 = \exp(-\lambda)\exp\left(1 + \log y_1 - \log F_1 - \frac{z_1}{y_1} + \frac{z_1 F_1}{\langle F, y\rangle}\right),$$

$$x_1 = \exp(-\lambda)\Psi_1(y, z) \tag{6.45}$$

and

$$x_2 = \exp(-\lambda)\exp\left(1 + \log y_2 - \log F_2 + \frac{F_2}{\langle F, y\rangle}\sum_{i\in J_z} z_i\right),$$

$$x_2 = \exp(-\lambda)\Psi_2(y, z) \tag{6.46}$$

where

$$\Psi_1(y, z) = \exp\left(1 + \log y_1 - \log F_1 - \frac{z_1}{y_1} + \frac{z_1 F_1}{\langle F, y\rangle}\right),$$

$$\Psi_2(y, z) = \exp\left(1 + \log y_2 - \log F_2 + \frac{F_2}{\langle F, y\rangle}\sum_{i\in J_z} z_i\right)$$

Then

$$\begin{aligned}
\langle F, x\rangle &= \exp(-\lambda)\exp\left(\frac{F_3}{\langle F, y\rangle}\sum_{i\in J_z} z_i\right)\\
&= F_1\exp(-\lambda)\Psi_1 + F_2\exp(-\lambda)\Psi_2 + F_3 x_3,\\
x_3 &= \exp(-\lambda)\left(\exp\left(\frac{F_3}{\langle F, y\rangle}\sum_{i\in J_z} z_i\right) - F_1\Psi_1 - F_2\Psi_2\right)/F_3. \qquad (6.47)
\end{aligned}$$

Again using the fact that x is a histogram we get

$$\exp(\lambda) = \left(\exp\left(\frac{F_3}{\langle F, y\rangle}\sum_{i\in J_z} z_i\right) - F_1\psi_1 - F_2\psi_2\right)/F_3.$$

which gives deterministic values for x given y, z as before.

Bringing back random mutations into the equations, we compute once again the functions $U(x,y), \tilde{U}(y,z), V(x,y)$ and $\tilde{V}(y,z)$.

$$U_i(x,y) = \begin{cases} -\log\langle F, x\rangle & i = 3\\ 1 + \log\frac{y_i}{F_i x_i} & i \neq 3\end{cases}$$

$$\tilde{U}_i(x,y) = \begin{cases} \frac{F_i z_1}{\langle F,y\rangle} & i = 2,3\\ \frac{-z_i}{y_i} + \frac{z_i F_i}{\langle F,y\rangle} & i = 1\end{cases}$$

$$V_i(x,y) = \sum_{j|j>i} a_{i,j}\exp(-a_{i,j}) + \sum_{j|j<i}\frac{F_j x_j}{F_i x_i} a_{i,j}\exp(a_{i,j}),$$

$$\begin{aligned}
\tilde{V}_i(y,z) = (g-i)F_i &+ \sum_{k|k>i}\left(\frac{F_i z_k}{F_k y_k} - F_i + \frac{z_k z_i}{F_k y_k y_i} - \frac{z_i^2}{F_i y_i^2} - \frac{z_i}{y_i}\right)\exp(-b_{i,k})\\
&+ \sum_{k|k<i}\left(-\frac{F_k y_k z_i^2}{F_i^2 y_i^3} + \frac{z_i z_k}{F_i y_i^2}\right)\exp(b_{i,k})
\end{aligned}$$

where $a_{i,j} = \frac{y_i}{F_i x_i} - \frac{y_j}{F_j x_j}$ and $b_{i,j} = \frac{z_i}{F_i y_i} - \frac{z_j}{F_j y_j}$. This gives

$$\frac{\partial U_i}{\partial x_j} = \begin{cases} \frac{-F_i}{\langle F, x(0) \rangle} & i = 3 \\ 0 & i \neq 3, i \neq j \\ \frac{-1}{x_i(0)} & i \neq 3, i = j \end{cases}$$

where $x(0)$ is given by Equations 6.45, 6.46 and 6.47. So

$$\Lambda = \begin{cases} \frac{-F_i X_i}{\langle F, x(0) \rangle} + V_i(x(0), y) + \tilde{V}_i(y, z) & i = 3 \\ \\ \frac{-X_i}{x_i(0)} + V_i(x(0), y) + \tilde{V}_i(y, z) & i \neq 3 \end{cases}$$

and hence

$$X_i = \begin{cases} \frac{\langle F, x(0) \rangle}{F_i}(-\Lambda + V_i(x(0), y) + \tilde{V}_i(y, z)) & i = 3 \\ \\ x_i(0)(-\Lambda + V_i(x(0), y) + \tilde{V}_i(y, z)) & i \neq 3 \end{cases}$$

and using the fact that $\sum_i X_i = 0$, we have

$$\sum_{i=1}^{2} x_i(0)(-\Lambda + V_i(x(0), y) + \tilde{V}_i(y, z)) + \frac{\langle F, x(0) \rangle}{F_3}(-\Lambda + V_3(x(0), y) + \tilde{V}_3(y, z)) = 0,$$

Hence we get the value of Λ as follows

$$\Lambda = \frac{\sum_{i=1}^{2} x_i(0)(V_i(x(0), y) + \tilde{V}_i(y, z)) + \frac{\langle F, x(0) \rangle}{F_3}(V_3(x(0), y) + \tilde{V}_3(y, z))}{\sum_{i=1}^{2} x_i(0) + \frac{\langle F, x(0) \rangle}{F_3}},$$

and thus

$$X_i = \begin{cases} \frac{\langle F, x(0) \rangle}{F_i}(-\Lambda + V_i(x(0), y) + \tilde{V}_i(y, z)) & i = 3 \\ \\ x_i(0)(-\Lambda + V_i(x(0), y) + \tilde{V}_i(y, z)) & i \neq 3 \end{cases}$$

where value of Λ is as derived above. So we have

$$x_i(1) = x_i(0) + mX_i$$

as the new solution.

6.5 Optimality Conditions for Mean-mutation Approximation

Next consider another stochastic model which is a process where mutations are considered to be deterministic and equal to the mean number of the Poisson random mutations at each mutation step.

For the trajectory $x \to y \to z$, let $\bar{p}$ and p be the intermediary population histogram before dilution for populations x and y respectively. Also, let vector $\vec{m} = (m_i)$ be the vector of mutation rates and $\bar{m} = \sum_{j=1}^{g} m_j$.

The complete model with its parameters is described in the chapter 4. We know for a model with mean mutations, the intermediary population histogram is given by

$$\bar{p}_j = (1 - \bar{m})\frac{F_j x_j}{\langle F, x \rangle} + m_j. \tag{6.48}$$

and

$$p_j = (1 - \bar{m})\frac{F_j y_j}{\langle F, y \rangle} + m_j, \tag{6.49}$$

The corresponding expression for one-step rate functional or the cost is given by

$$\mathrm{RF}(x, y) = \sum_{j=1}^{g} y_j \log \frac{y_j}{\bar{p}_j} = \sum_{j=1}^{g} y_j \log y_j - y_j \log \bar{p}_j$$

and

$$\mathrm{RF}(y, z) = \sum_{j=1}^{g} z_j \log z_j - z_j \log p_j$$

Thus we have, for the derivative of rate functionals w.r.t y

$$\frac{\partial \mathrm{RF}(x,y)}{\partial y_j} = 1 + \log y_j - \log \bar{p}_j$$

and

$$\frac{\partial \mathrm{RF}(y,z)}{\partial y_j} = -\sum_k \frac{z_j}{p_k}\frac{\partial p_k}{\partial y_j} = (1-\bar{m})\frac{F_j}{\langle F, y\rangle}$$

So, using Lagrange optimality conditions from 6.1 we get

$$\lambda = 1 + \log y_j - \log \bar{p}_j + (1-\bar{m})\frac{F_j}{\langle F, y\rangle}$$

which simplifies to give

$$\bar{p}_j = \exp(-\lambda)\exp\left(1 + \log y_j + (1-\bar{m})\frac{F_j}{\langle F, y\rangle}\right)$$

Let

$$\Psi_j(y,z) = \exp\left(1 + \log(y_j) + (1-\bar{m})\frac{F_j}{\langle F, y\rangle}\right). \tag{6.50}$$

Then using the fact that $\bar{p}$ is a histogram we get

$$\exp(\lambda) = \sum_{j=1}^{g} \Psi_j(y,z).$$

Thus we have the value of $\bar{p}$ given y and z as follows

$$\bar{p}_j = \frac{\Psi_j(y,z)}{\sum_{j=1}^{g}\Psi_j(y,z)}$$

We know from Equation 6.48 that

$$\bar{p}_j = (1-\bar{m})\frac{F_j x_j}{\langle F, x\rangle} + m_j,$$

Rearranging we obtain

$$\frac{x_j}{\langle F, x\rangle} = \frac{\bar{p}_j - m_j}{F_j(1-\bar{m})}, \tag{6.51}$$

Now, taking sum over all genotypes, we get

$$\sum_{k=1}^{g} \frac{x_j}{\langle F, x\rangle} = \sum_{k=1}^{g} \left[\frac{\bar{p}_k - m_k}{F_k(1-\bar{m})}\right].$$

This gives

$$\langle F, x\rangle = \frac{(1-\bar{m})}{\sum_{k=1}^{g} \frac{\bar{p}_k - m_k}{F_k}} \tag{6.52}$$

Using Equations 6.51 and 6.52 we express the value for x for the case of mean random mutations as follows

$$x_j = \frac{\bar{p}_j - m_j}{F_j} \frac{1}{\sum_{k=1}^{g} \frac{\bar{p}_k - m_k}{F_k}} \tag{6.53}$$

CHAPTER 7

Shooting Algorithms to Compute Rate Minimizing Evolution Trajectories

The motivation for building reverse trajectories using only penultimate steps comes from shooting algorithms [116] used to numerically solve two-point boundary value problems. Our problem of optimizing rate functional for a fixed initial and target histograms can be characterized broadly as a two-point boundary value problem.

Also, the problem can be expressed from a geodesic point of view where we are trying to find the best geodesic between two histograms based on a metric defined by some rate functional.

However, the task of solving a two-point boundary value problem is in general very difficult. The available solution techniques usually face serious convergence difficulties because of the lack of a good initial guess. A lot of global algorithms for finding a geodesic joining two given points in general spaces have been developed over the time.

For our applications, theoretically we can generate rate minimizing trajectory in both forward and reverse directions, i.e., projecting from either initial state histogram or target state histogram. However as can be seen by formulas in chapter 6 for the chain $x \to y \to z$, isolating explicit formula for z given x, y is not as simple as isolating expression for x given y, z. So we build our trajectories for a fixed target state and use feasible shooting directions to generate required rate minimizing trajectory which starts at our required initial state.

The explicit formulas derived in previous chapter for generating reverse rate minimizing optimal evolutionary trajectories are used to develop MATLAB subroutines to implement the above approach numerically. The detailed explanation of the algorithm to obtain these trajectories is explained in the following section.

7.1 Building the Most Likely Trajectory

7.1.1 First-stage Rate Minimizing Trajectory

We generate a state space $HIST$ of all possible histograms with a given discretization $d\%$. Numerically, having a $d\%$ discretization in a state space $HIST$ with g genotypes implies that for any 2 distinct histograms $H, G \in HIST$, the difference at any two indices is always a multiple of d. So, for all $j = 1, ..., g$

$$|H(j) - G(j)| = l(j)d, \ \forall H, G \in HIST \tag{7.1}$$

where $l(j) \geq 0$ is a scalar. For instance, a state space with $g = 4$ genotypes with $d = 2\%$ discretization has about 24000 histograms.

When concentrating on evolutionary trajectories which lead to fixation of a certain genotype starting from a population with some other fixed genotype, the region of interest ($RI \subset HIST$) is generally classified by the sets of histogram which have the required genotype frequency greater than a given threshold say 99%.

The core subroutine developed fixes a target histogram $H_{\text{tar}} \in RI$. Let Ω be a fixed neighborhood of H_{tar} in the discretized set of histograms $HIST$. Possible penultimate steps H_{pen} are chosen from set $\Omega \subset HIST$. Denote by H_{in} the required initial histogram for the rate minimizing trajectory.

The iterative scheme developed in chapter 6 builds the trajectory step by step using H_{tar} and H_{pen} as first inputs and giving $H_{\text{pen}-1}$ as optimal step on trajectory. Recall here that in the formula 6.23, after first step of iteration we update h_n and h_{n-1} to be H_{pen} and $H_{\text{pen}-1}$ respectively and compute corresponding h_{n-2} as the next optimal step in the chain as follows.

$$h_{n-2}(j) = \frac{\Psi_j(h_{n-1}, h_n)}{\sum_{j=1}^{g} \Psi_j(h_{n-1}, h_n)}, \forall j. \tag{7.2}$$

where Ψ is given by

$$\Psi_j(h_{n-1}, h_n) = \exp\left(1 + \log h_{n-1}(j) - \log F(j) - \frac{h_n(j)}{h_{n-1}(j)} + \frac{F(j)}{\langle F, h_{n-1}\rangle}\right). \tag{7.3}$$

This recursive procedure can be repeated until the current histogram h_n becomes too close to some boundary, i.e., $h_n(j) < \epsilon$ for some $j = 1, ..., g$ where ϵ is the tolerance threshold for boundary as discussed in detail in chapter 4, since formula 6.23 holds for interior points only.

For the initial histogram H_{in}, we construct the unique mean trajectory starting at H_{in} which has mean mutations and mean sampling frequency in multinomial sampling. Call this mean trajectory mtr and its length to be L. The cost or the rate functional associated to mtr is 0.

For each point mtr_k on the trajectory mtr, define a neighborhood V_k of mtr_k. Similarly, let U be a neighborhood of H_{in}. Then we define a starting zone W_{init} as $W_{\text{init}} = \cup_k V_k \cup U$.

Then, we compute the set TR_Ω containing all the feasible rate minimizing trajectories using $H_{\text{pen}} \in \Omega$ with target H_{tar}. For every trajectory $tr \in \text{TR}_\Omega$ a cost, RF(tr) given by the rate functional in Equation 5.26 is computed and attached to it. Hence TR_Ω contains one trajectory for each choice of H_{pen} in Ω.

The minimal one-step cost from any histogram A to another histogram B is computed using equation 5.26 and defined as $\text{cost}(A, B)$ and in general $\text{cost}(A, B) \neq \text{cost}(B, A)$.

Since these trajectories are allowed to "reverse" shoot in every possible direction, we have reverse trajectories in TR_{HIST} starting at H_{tar} and reverse shooting towards H_{pen}, where H_{pen} is arbitrary in $HIST$. Now the aim is to isolate trajectories starting at a given initial histogram H_{in} and ending at H_{tar}. However we only fixed the target state, so it is possible that none of the trajectories in $\mathrm{TR}_{\mathbf{HIST}}$ have H_{in} as their initial state.

We now filter the trajectories on the basis of their initial state and call $\mathrm{TR}_{\mathrm{com}}$ to be the set of trajectories in TR_{Ω} with initial histogram in W_{init} and remaining trajectories in TR_{Ω} are classified as incomplete and collected in the set $\mathrm{TR}_{\mathrm{inc}}$. Thus we consider the trajectory to be complete if it starts in the neighborhood W_{init}.

Now, we want to generate trajectories which first follow mean trajectory mtr for a while before going towards H_{tar}. The procedure is described below.

1. For every trajectory, $tr \in \mathrm{TR}_{\mathrm{inc}}$, we follow steps $2-7$ and get a corresponding complete trajectory, ctr.

2. For every point $tr_k,\ k = 1 : l$ on tr with length l, compute the one step cost, $\mathrm{cost}(mtr_j, tr_k)$ from all the points $mtr_j,\ j = 1 : L$ on mean trajectory mtr.

3. Among all the points mtr_j of the mean trajectory mtr, call mtr_K the point which achieves the least one-step cost. Hence we compute $\mathrm{cost}(mtr, tr_j) = \min_{mtr_i \in mtr} \mathrm{cost}(mtr_i, tr_j)$.

4. Now we have a minimal one-step cost $\mathrm{RF}_k = \mathrm{cost}(mtr, tr_k)$ from mtr to every point tr_k on trajectory tr.

5. Next we optimize over all these one step costs, $\mathrm{cost}(mtr, tr) = \min_{k=1:l}\mathrm{RF}_k = \min_{tr_k \in tr}\mathrm{cost}(mtr, tr_k)$. This gives an optimal point on tr which minimizes jump cost from mtr to tr. Let such a histogram on tr be tr_j, its corresponding histogram on mtr be mtr_J and minimal cost be RF_j.

6. A complete new trajectory, ctr is constructed by concatenating the partial trajectory $tr_j : tr_l$ at the end of partial mean trajectory, $mtr_1 : mtr_J$.

7. The corresponding cost is the sum of cost from mtr_J to tr_j, RF_j =cost(mtr, tr) and the cost of partial trajectory $tr_j : tr_l$.

These new complete trajectories thus obtained are integrated with the set $\mathrm{TR}_{\mathrm{in}}$ to form a new set $\mathrm{CTR}_{\mathrm{in}}$ of trajectories since they all start in the required neighborhood of H_{in}. Finally for obtaining first-stage rate minimizing trajectory, we choose the optimal minimal cost trajectory from the set $\mathrm{TR}_{\mathrm{in}}$ and call it tr_{opt}. The corresponding value of cost or rate functional is $\mathrm{RF}_{\mathrm{opt}}$.

The set of all the above trajectories is referred to as first-stage rate minimizing evolution trajectories.

We still save the original set of incomplete trajectories $\mathrm{TR}_{\mathrm{inc}}$ and use it to obtain multi-stage rate minimizing trajectory as described next.

7.1.2 Multi-stage Rate Minimizing Trajectory

All the incomplete first-stage trajectories in $\mathrm{TR}_{\mathrm{inc}}$ are now used to generate multi-stage rate minimizing trajectories.

Using $\mathrm{RF}_{\mathrm{opt}}$, the minimal cost obtained from complete first-stage trajectories we filter from the set of incomplete trajectories $\mathrm{TR}_{\mathrm{inc}}$, the ones which have lower cost than $\mathrm{RF}_{\mathrm{opt}}$. Call the new set of incomplete trajectories thus obtained LTR_{inc}, which will become the end segments of our potential multi-stage rate minimizing trajectories.

The initial point of each of these incomplete trajectories is considered a potential new target point $newH_{\text{tar}}$. Denote $newHt$ to be the set of all points $newH_{\text{tar}}$ obtained at this stage. For each $newH_{\text{tar}} \in newHt$, we use the same starting zone W_{init} as before for the neighborhood of required initial histogram H_{in}. Penultimate step histogram at this stage can be chosen from the set $n\Omega_H$, where $n\Omega_H$ is the set of neighborhood histograms around $newH_{\text{tar}}$.

Now first-stage reverse optimal trajectories ending at $newH_{\text{tar}}$ are generated for every $newH_{\text{tar}}$. Thus for a fixed $newH_{\text{tar}}$, we obtain new sets of trajectories as complete and incomplete trajectories as HTR_{com} and HTR_{inc} respectively.

The procedure is repeated for every $newH_{\text{tar}} \in newHt$ to obtain 2 new sets of trajectories as follows

$$\text{nTR}_{\text{com}} = \cup_{\text{newH}_{\text{tar}} \in \text{newHt}} \text{HTR}_{\text{com}}$$

and

$$\text{nTR}_{\text{inc}} = \cup_{\text{newH}_{\text{tar}} \in \text{newHt}} \text{HTR}_{\text{inc}}$$

Now we have a new optimal cost, RF_{new} from the set of all complete trajectories, nTR_{com}. We find the new optimal two-stage rate minimizing trajectory and update $\text{RF}_{\text{opt}} = \min\{\text{RF}_{\text{new}}, \text{RF}_{\text{opt}}\}$ as the new minimal cost.

If we still have any incomplete trajectories at this stage which have a cost lower than RF_{opt}, then the above described procedure for finding a next three-stage rate minimizing trajectory using their respective $newH_{\text{tar}}$ is followed. This multi-stage algorithm is applied till all of the incomplete trajectories have been either completed or have a cost higher than $\mathrm{RF}_{\mathrm{opt}}$.

The final minimal evolution trajectory thus computed is called the optimal multi-stage rate minimizing trajectory.

The algorithmic steps for the complete iterative procedure is presented below.

1. Input Target state H_{tar}, discretization level d, required initial state H_{in} and mean trajectory, mtr starting at H_{in}.

2. Compute neighborhood Ω of H_{tar}.

3. Define starting zone W_{init} to be a neighborhood of H_{in} and points on mtr.

4. For every penultimate step H_{pen} in Ω implement reverse trajectory subroutine to get a set of first-stage reverse optimal trajectories TR_{Ω}.

5. Compute corresponding costs RF(tr) for every trajectory, $tr \in \mathrm{TR}_{\Omega}$.

6. Call $\mathrm{TR}_{\mathrm{in}}$ to be the set of trajectories in TR_{Ω} with initial histogram in W_{init} and remaining trajectories in TR_{Ω} are classified as incomplete and collected in the set $\mathrm{TR}_{\mathrm{inc}}$.

7. Complete the incomplete trajectories by concatenating mean trajectory, mtr using the optimal jump cost from mtr to $tr \in \mathrm{TR}_{\mathrm{inc}}$ and form a new set of complete trajectories $\mathrm{CTR}_{\mathrm{in}}$.

8. The minimal cost trajectory is extracted from the set $\mathrm{CTR}_{\mathrm{in}}$ and its cost is called $\mathrm{RF}_{\mathrm{opt}}$. This concludes the computation for first-stage rate minimizing

trajectory.

9. Set $newH_{\text{tar}}$ = initial state of the incomplete trajectories with cost less than RF_{opt}.

10. Compute $n\Omega_H$ for each H_{ntar} and repeat steps $4 - 9$.

11. Following the above steps, if we obtain a trajectory from H_{in} to H_{tar} with lower cost than RF_{opt}, then we call it as the new rate-minimizing trajectory and update the value of RF_{opt} as the cost of this new trajectory.

12. Repeat till all trajectories are complete or until all incomplete trajectories have higher cost than RF_{opt}. This concludes the computation for multi-stage rate minimizing trajectory from H_{in} to H_{tar} with optimal cost given by RF_{opt}.

7.1.3 Multi-scale Rate Minimizing Trajectory

The search for cost minimizing trajectories starting at H_{in} and ending at H_{tar} can be improved further by introducing finer scaling in addition to the multi-stage approach described before.

Using a finer discretization scale $\delta\%$ to compute new discretized state space $HIST_\delta$ for possible penultimate steps H_{pen} increases the computation time exponentially and also requires a lot of computing memory to be available. For instance, a state space with $g = 4$ genotypes with $d = 2\%$ discretization has about 24000 histograms.

One approach is to generate whole state space $HIST_\delta$ using very fine discretization $\delta\%$ but use a very restricted $\Omega_\delta \subset HIST_\delta$ as a set of histograms available for the choice of H_{pen}. We can select a neighborhood of H_{tar} in $HIST_\delta$ as our new Ω_δ. We can assume that optimal H_{pen} will lie in some small neighborhood of H_{tar}.

Another approach to finer scaling is to first generate an optimal multi-stage trajectory with coarse discretization and then selectively introduce finer discretization near some selected histograms in the optimal trajectory. Once we know the optimal penultimate step H_{pen} (or the reverse shooting direction) for the optimal trajectory, we can generate closer neighbors around H_{pen} with finer discretization. Call Ω_δ as these new set of points and generate a new optimal multi-stage trajectory using Ω_δ in place of Ω instead. This is based on the continuity assumptions that the shooting direction required for an optimal trajectory in finer scale would not deviate too far from the shooting direction used to compute optimal trajectory in a coarser discretization.

In our numerical computations we have implemented the first described approach for estimating multi-scale rate minimizing trajectories. Thus using Ω_δ as the new state space and for the same target state H_{tar}, multi-stage rate minimizing trajectory is computed and the optimal rate minimizing trajectory TR_{opt} is updated along with its cost RF_{opt}.

As expected the multiscale approach leads to a significant reduction in RF_{opt} as we introduce finer and finer discretization. However after a certain discretization level the value of RF_{opt} stops changing and stabilizes.

A big reduction in required computation time and memory resources needed is observed using this technique. Some estimates for cpu times observed are presented in next chapter. The particular cases for different number of genotypes along with figures are presented in next chapter.

CHAPTER 8

Most Likely Optimal Paths Realizing Rare Events

In this chapter we work with parameters derived from realistic estimates [129] concerning the TC experiment and present step by step generation of optimal multi-stage trajectory. For the purpose of our simulations, genotypes are arranged in order of their increasing growth factor, i.e, $F_1 < F_2 < ... < F_g$ and the rate of mutations is considered to be same for all genotypes in the population.

8.1 Numerical Results for the Population with 2 Genotypes

We show in following figures the comparison of the mean trajectory with the multi-stage optimal trajectory obtained via our approach of estimating rate minimizing trajectory ($\mathrm{TR_{opt}}$) using the technique outlined in chapter 7.

The realistic growth factors and mutation rates adopted for this example are derived from [129] and are shown in Table 8.1.

Genotype	Growth Factor	Mutation Rate
1	200	0.5×10^{-6}
2	$200^{1.1}$	0.5×10^{-6}

Table 8.1: Parameters for TC experiment

Without loss of generality let H be the initial and G be the terminal state in the trajectory. Then H_i, G_i denote the frequency of genotype i in initial and terminal target state respectively.

Figure 8.1 shows the mean trajectory. The initial state has genotype 1 dominant in the population, $H_1 \geq 0.99$ and the target state has genotype 2 dominant, $G_2 \geq 0.99$.

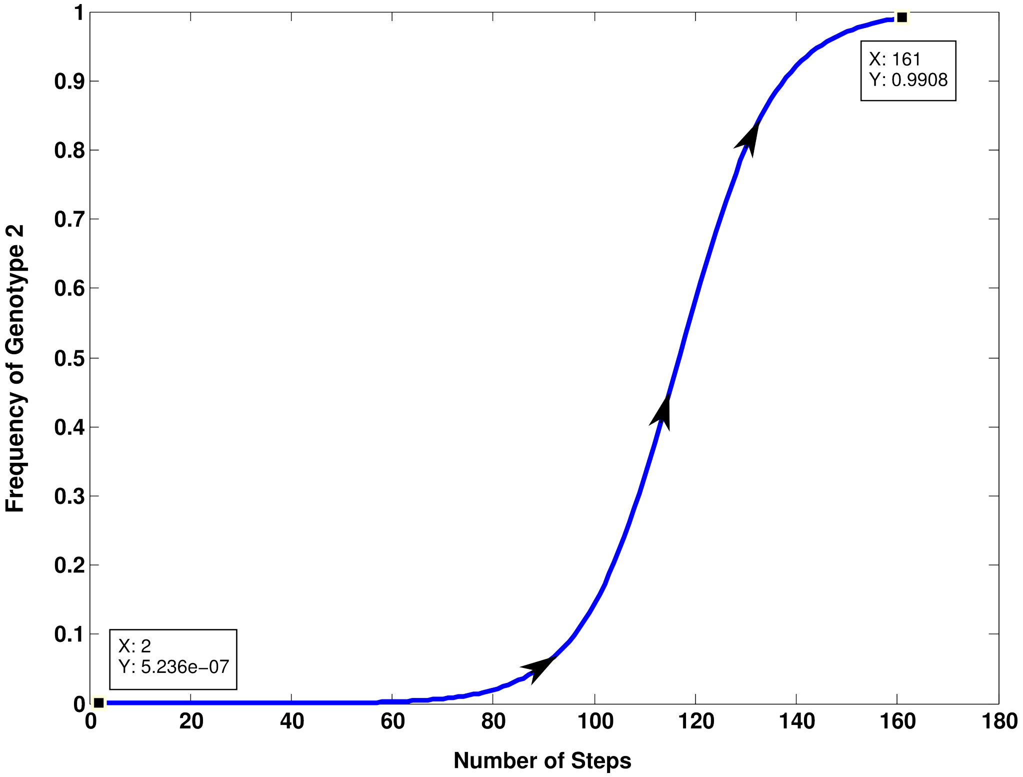

Figure 8.1: Mean Trajectory in Population with $g = 2$ Genotypes.

The step by step computation for the $\mathrm{TR}_{\mathrm{opt}}$ is presented below. Table 8.2 shows the possible target histograms used which satisfy $G_2 \geq 0.99$. In this case we obtain 4 possible targets using discretization of 0.25%.

Frequency of Genotype 1	Frequency of Genotype 2
0.0025	0.9975
0.0050	0.9950
0.0075	0.9925
0.0100	0.9900

Table 8.2: Possible Target Histograms in a Population with $g = 2$ Genotypes

Table 8.3 on the next page lists histograms in the neighborhood of a target histogram. They form the set Ω which is the set of all possible penultimate steps. In this case we obtain 19 neighbors of terminal state $[0.0025, 0.9975]$ using discretization of 0.25%.

Frequency of Genotype 1	Frequency of Genotype 2
0.0025	0.9975
0.0050	0.9950
0.0075	0.9925
0.0100	0.9900
0.0125	0.9875
0.0150	0.9850
0.0175	0.9825
0.0200	0.9800
0.0225	0.9775
0.0250	0.9750
0.0275	0.9725
0.0300	0.9700
0.0325	0.9675
0.0350	0.9650
0.0375	0.9625
0.0400	0.9600
0.0425	0.9575
0.0450	0.9550
0.0475	0.9525

Table 8.3: Possible Penultimate Histograms in a Population with $g = 2$ Genotypes

The one-stage rate minimizing trajectory can be seen in Figure 8.2.

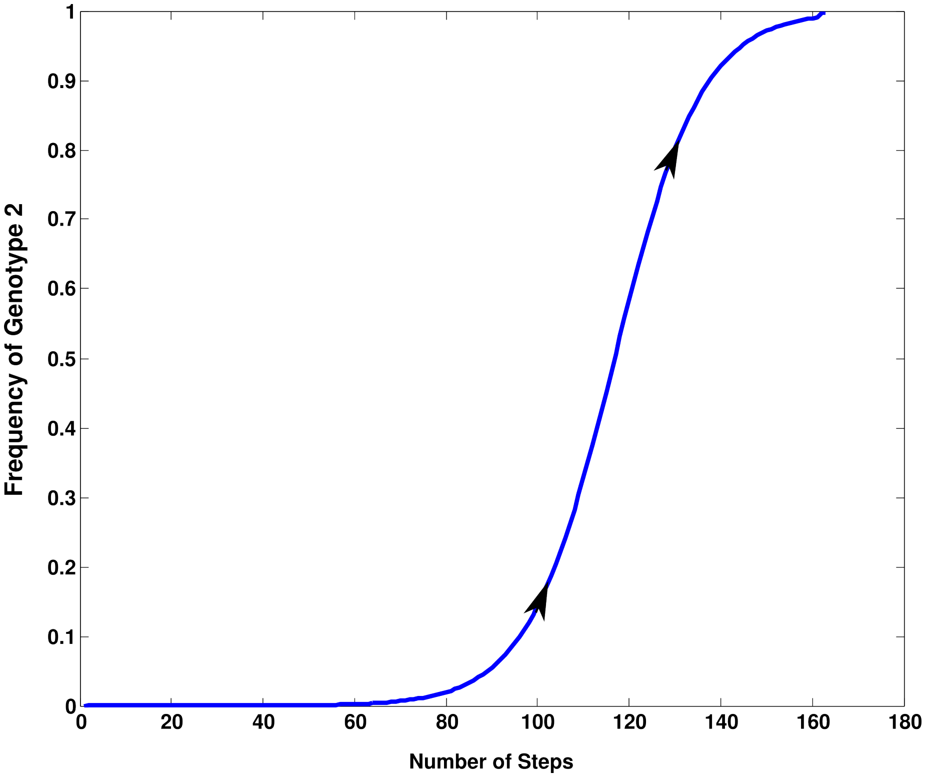

Figure 8.2: Rate Minimizing Optimal Trajectory at Step 1

We do not observe any incomplete trajectories at first stage in this case. So, figure 8.3 shows the mean trajectory with blue stars and multi-stage optimal rate minimizing trajectory ($\mathrm{TR_{opt}}$) in red.

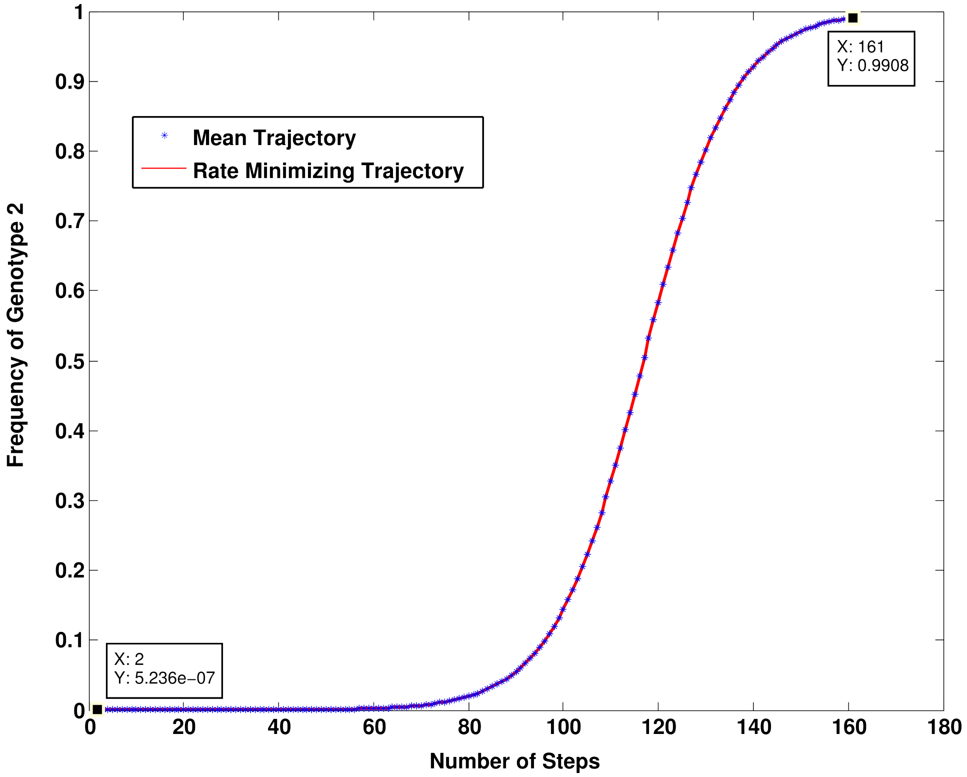

Figure 8.3: Mean Trajectory and Rate Minimizing Optimal Trajectory in Population with $g = 2$ Genotypes.

Mean trajectory has 0 cost. On the other hand, $\mathrm{TR_{opt}}$ (takes 6.3 minutes to generate) has a cost of 4.7×10^{-5}. The mean trajectory has 161 steps and $\mathrm{TR_{opt}}$ has 163 steps. In this case, we observe that both the trajectories overlap and thus are essentially identical. This also validates our numerical cost optimization technique and gives us confidence to proceed and use it for larger number of genotypes.

8.2 Numerical Results for the Population with 3 Genotypes

We now explore the case of 3 genotype and first study two events of obvious practical interest. The first event has probability close to 1, and is realized if we have population fixation at fittest genotype 3. The second event is a rare event, namely the population fixation at genotype 2 when starting at population where genotype 1 dominates overwhelmingly. We also study a comparatively less rare event where population has frequency of genotype 2 to be $\geq 60\%$ and total frequency of genotype 1 and 3 to be $\leq 40\%$.

The realistic growth factors and mutation rates adopted for this example are derived from [129] and are shown in Table 8.4.

Genotype	Growth Factor	Mutation Rate
1	200	0.5×10^{-6}
2	$200^{1.08}$	0.5×10^{-6}
3	$200^{1.15}$	0.5×10^{-6}

Table 8.4: Parameters Derived from Realistic Estimates (see [129]) Concerning the TC Experiment

Again, without loss of generality let H be the initial and G be the terminal state in the trajectory. Then H_i, G_i denote the frequency of genotype i in initial and terminal target state respectively.

Figure 8.4 shows the mean trajectory. The initial state is one with genotype 1 dominant in the population, $H_1 \geq 0.99$ and the terminal target state has genotype 3 dominant, $G_3 \geq 0.99$.

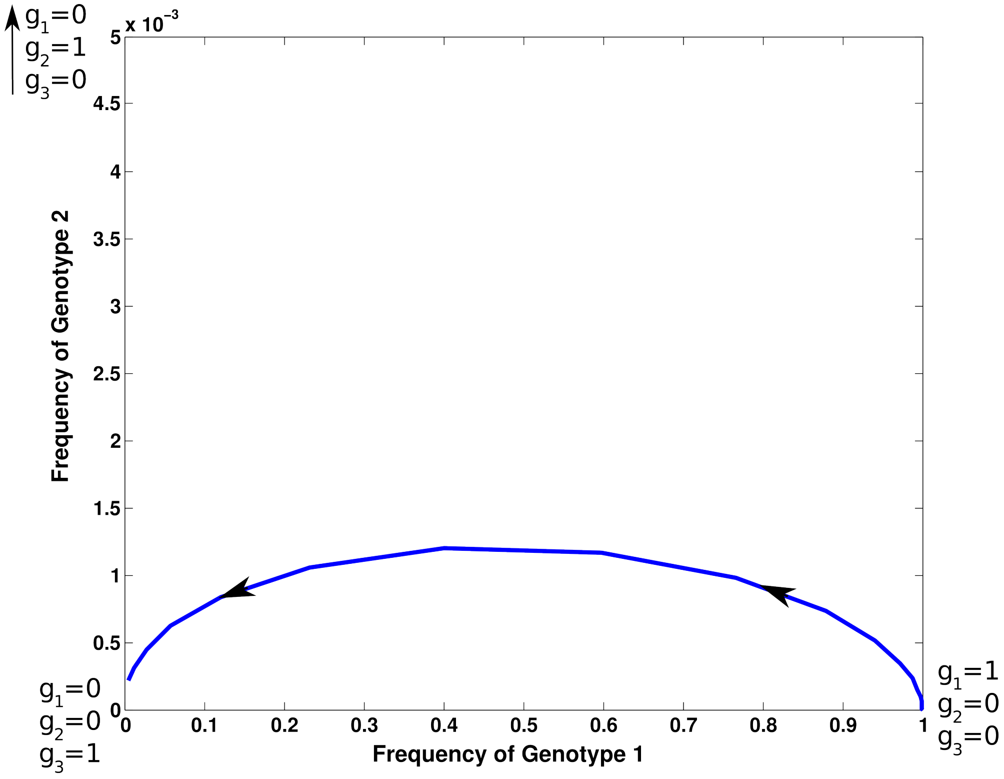

Figure 8.4: Mean Trajectory in Population with $g = 3$ Genotypes.

First we present the graphs for the most likely trajectory from initial state with $H_1 \geq 0.99$ and target state with $G_3 \geq 0.99$.

Figure 8.5 shows the mean trajectory in blue and multi-stage optimal rate minimizing trajectory ($\mathrm{TR}_{\mathrm{opt}}$) in red.

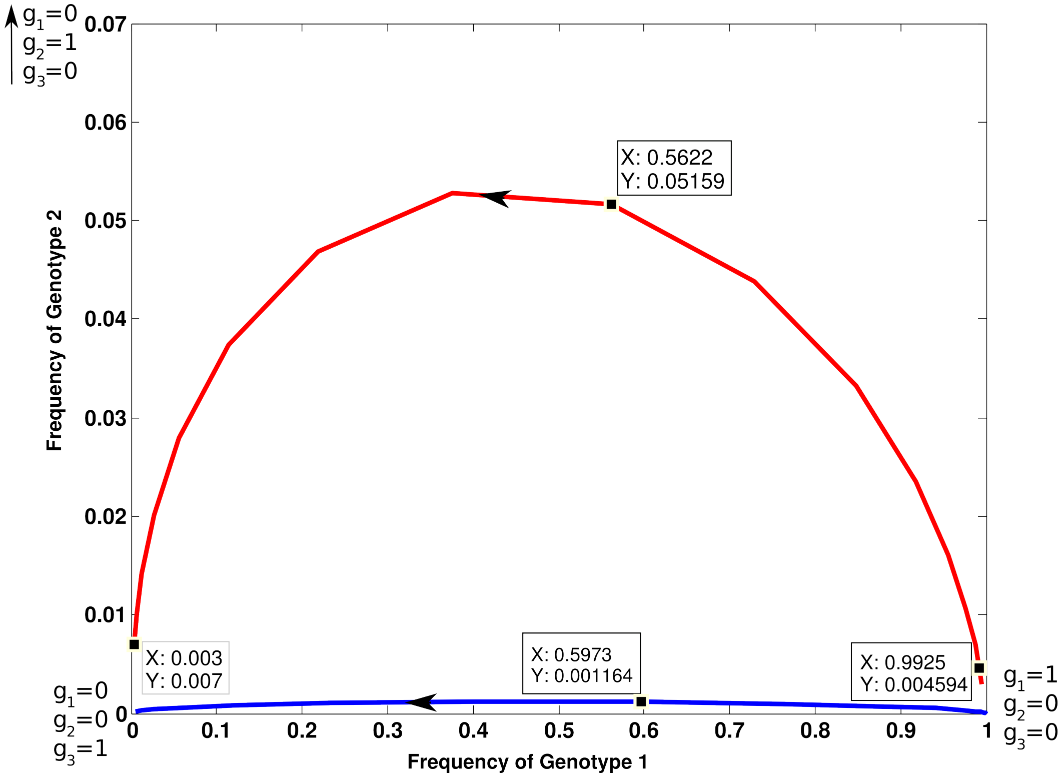

Figure 8.5: Mean Trajectory (Blue) and Rate Minimizing Optimal Trajectory (Red) in Population with $g = 3$ Genotypes.

Mean trajectory has 0 cost. $\mathrm{TR}_{\mathrm{opt}}$ on the other hand (takes 12 minutes to generate) has a cost of 6×10^{-5}. The mean trajectory has 26 steps and large deviations trajectory has 19 steps. Here we observe that the trajectories are not really identical as in the case of 2 genotypes. We compute the mean trajectory using $[1, 0, 0]$ as our initial state and population of genotype 2 does not grow noticeably and stays near boundary (almost 0). However in our computations, we have to use interior points and our tolerance for boundary is 10^{-3}, thus not allowing the trajectory to get too close to boundary like in the mean trajectory.

Next, we present graphs for the rare event trajectory with initial state having frequency of genotype 1, $H_1 \geq 0.99$ and target state having frequency of genotype 2, $G_2 \geq 0.99$. The step by step computation for $\mathrm{TR}_{\mathrm{opt}}$ is shown in following sequence of figures.

Figure 8.6 shows the possible target histograms, which have the frequency of genotype 2, $G_2 \geq 0.99$. In this case we obtain 10 possible target histograms using discretization of 0.2% around [0.0080.990.002].

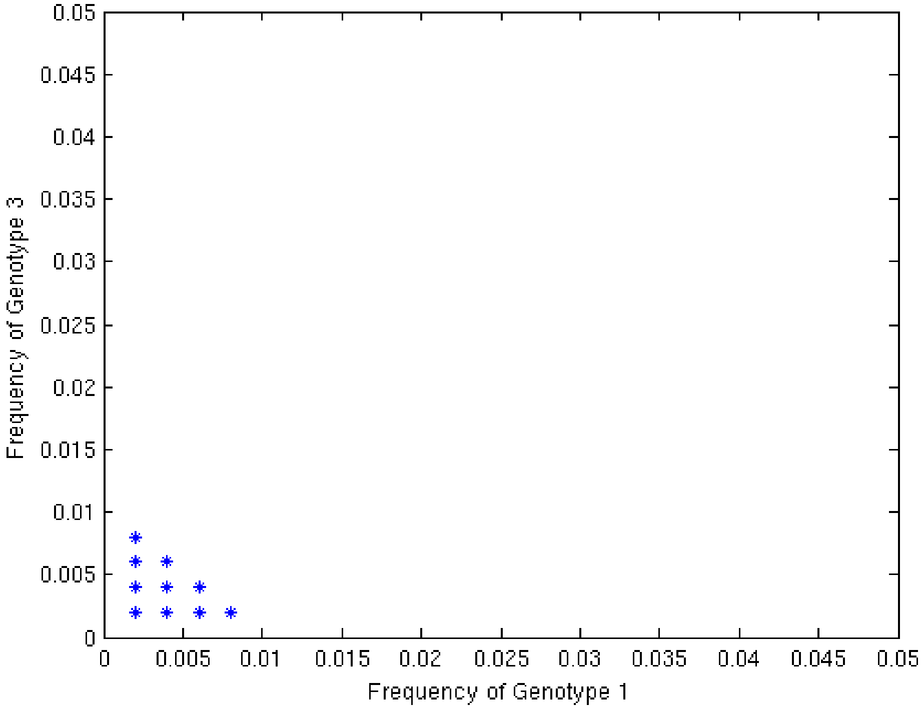

Figure 8.6: Possible Target Histograms in a Population with $g = 3$ Genotypes

Figure 8.7 shows histograms in the neighborhood of a target terminal histogram [0.0020.990.008]. They form the set Ω which is the set of all penultimate steps. In this case we obtain 400 neighbors of terminal state using discretization of 0.2%.

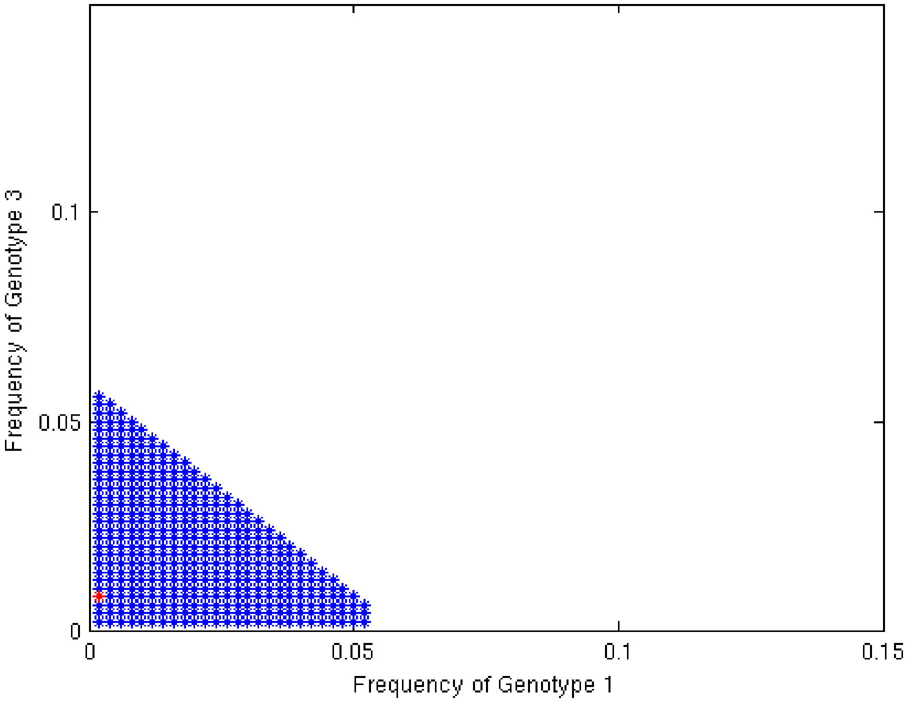

Figure 8.7: Possible Penultimate Histograms in a Population with $g = 3$ Genotypes

The one-stage rate minimizing trajectory which takes approximately 2 minutes to generate and has cost of 0.21 can be seen in Figure 8.8.

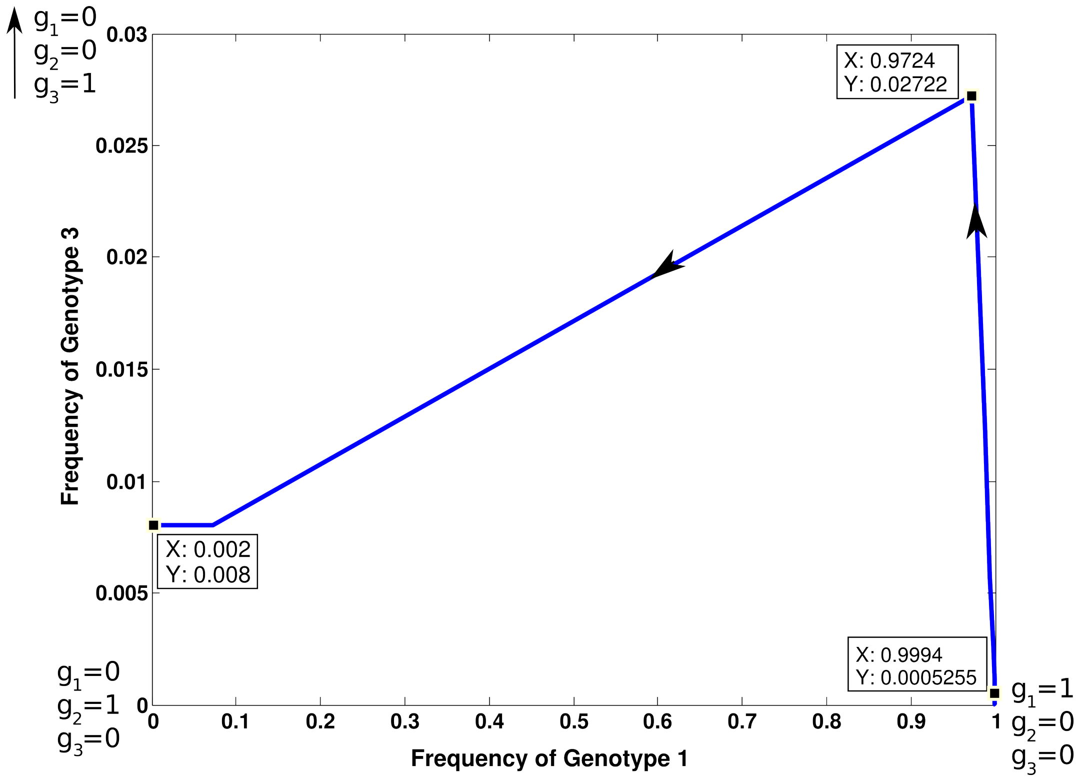

Figure 8.8: Rate Minimizing Optimal Trajectory at Step 1

Some of the incomplete trajectories at first step from the set $\mathrm{TR}_{\mathrm{inc}}$ are shown in Figure 8.9.

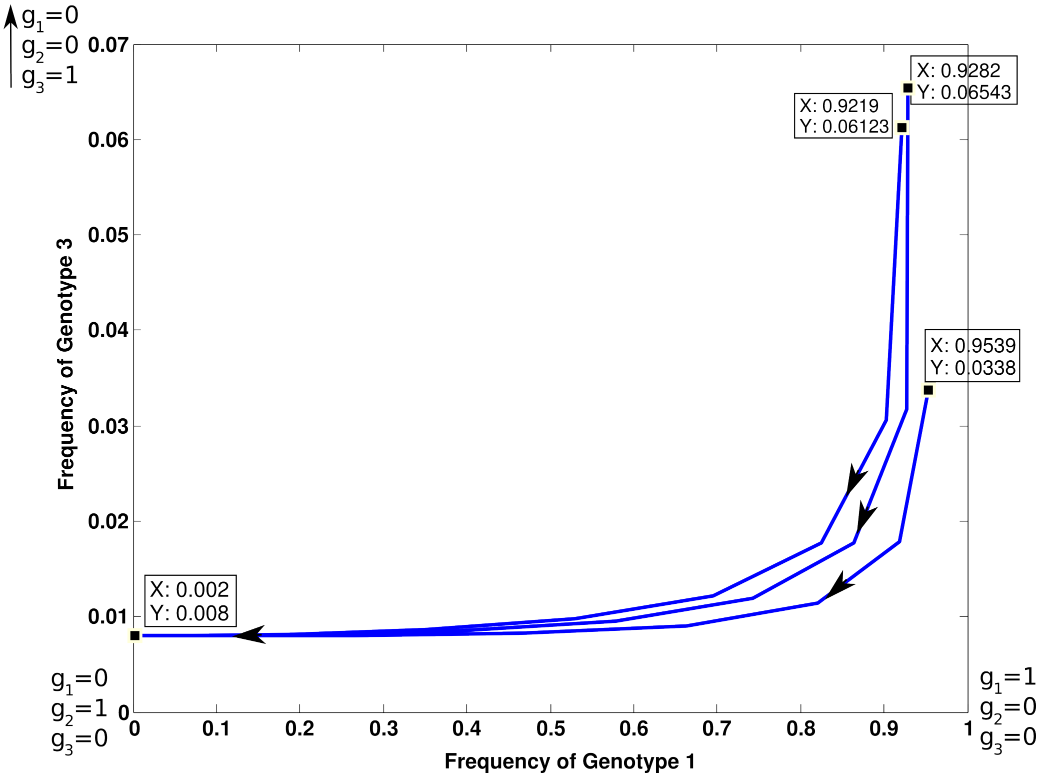

Figure 8.9: Incomplete Trajectories at Stage 1.

The initial point of each of these incomplete trajectories is considered a potential new target point $newH_{\text{tar}}$. Penultimate step histogram at this stage can be chosen from the set $n\Omega_H$, where $n\Omega_H$ is the set of neighborhood histograms around $newH_{\text{tar}}$ and is presented in Figure 8.10.

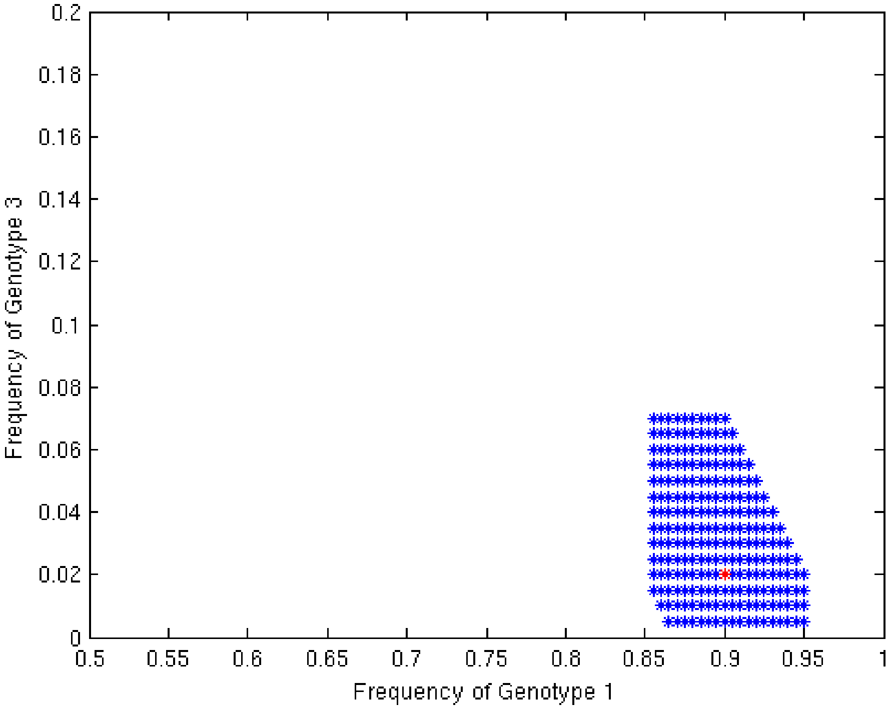

Figure 8.10: State Space $n\Omega_H$ of Possible Penultimate Histogram for the Incomplete Trajectory.

So, figure 8.11 shows the multi-stage optimal rate minimizing trajectory, TR_{opt} obtained.

Rate minimizing optimal trajectory, TR_{opt} in this case has a cost of 0.178.

The costs and number of steps associated to rate minimizing trajectories in this

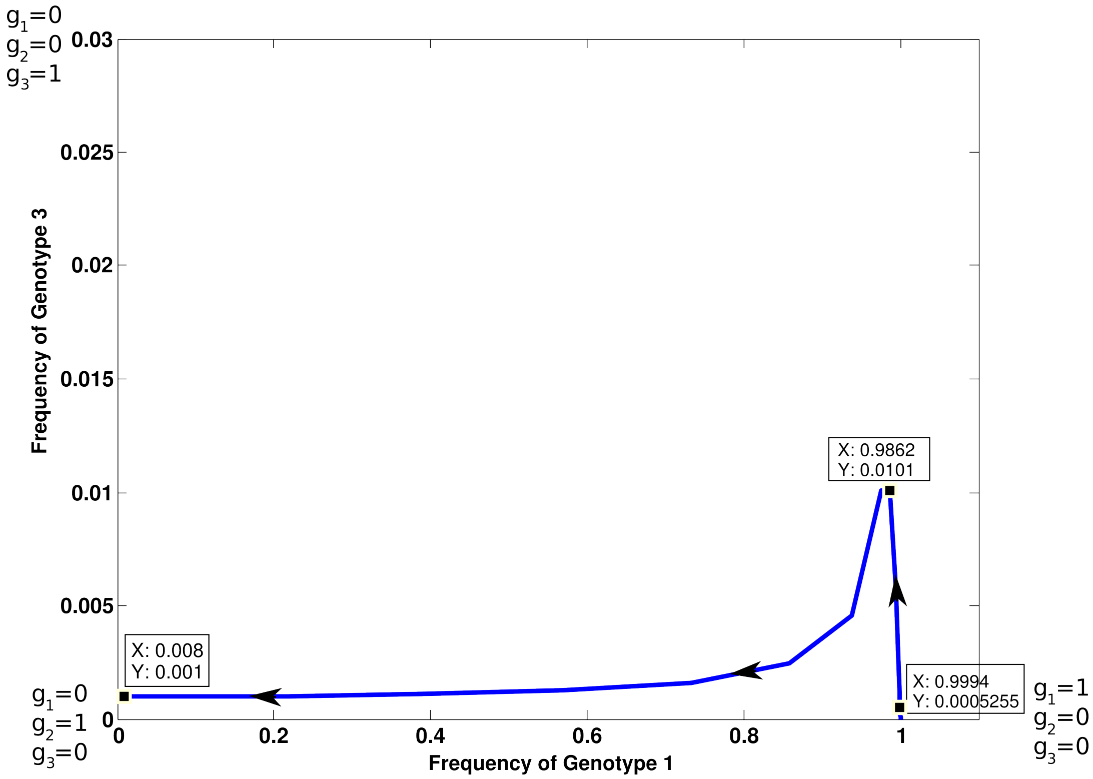

Figure 8.11: Rate Minimizing Optimal Trajectory in Population with $g = 3$ Genotypes.

case are tabulated in Table 8.5.

Initial Genotype	Target Genotype	Number of Steps	Cost
1	2	24	0.178
1	3	19	6×10^{-5}

Table 8.5: Rate Minimizing Trajectories for TC Experiment

8.3 Numerical Results for the Population with 4 Genotypes

Next, we illustrate the case of 4 genotype and study three feasible cases. One is the most likely event of population fixating at fittest genotype 4, second is the rare event of population fixing at a lesser fit genotype 3 and third is the rarest event of population fixing at genotype 2 while starting at population dominant with genotype 1.

The realistic growth factors and mutation rates adopted for this example are derived from [129] and are shown in Table 8.6.

Genotype	Growth Factor	Mutation Rate
1	$200^{1.06}$	0.5×10^{-6}
2	$200^{1.08}$	0.5×10^{-6}
3	$200^{1.1}$	0.5×10^{-6}
4	$200^{1.12}$	0.5×10^{-6}

Table 8.6: Parameters for TC Experiment

Recall that H is the initial and G is the terminal state in the trajectory and H_i, G_i denote the frequency of genotype i in initial and terminal target state respectively.

Figure 8.12 shows the mean trajectory. The initial state is one where genotype 1 is dominant in the population, $H_1 \geq 0.99$ and the target state has genotype 4 dominant, $G_4 \geq 0.99$.

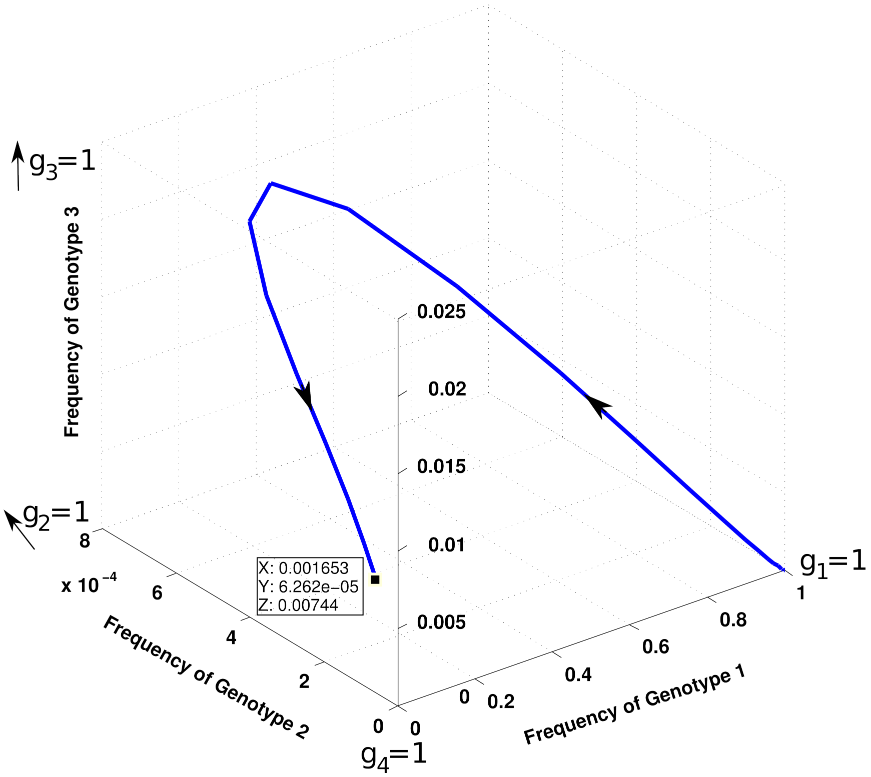

Figure 8.12: Mean Trajectory in Population with $g = 4$ Genotypes.

Figure 8.13 shows the mean trajectory in blue and optimal rate minimizing trajectory, $\mathrm{TR_{opt}}$ in red.

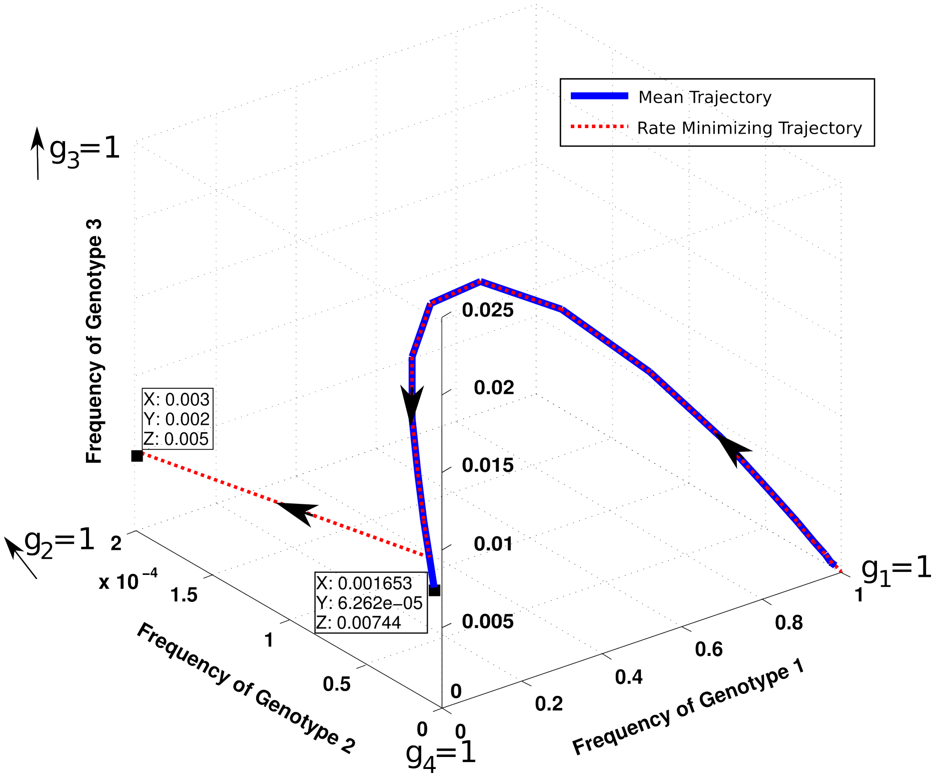

Figure 8.13: Mean Trajectory and Rate Minimizing Optimal Trajectory in Population with $g = 4$ Genotypes.

As before the mean trajectory has 0 cost associated to it. $\mathrm{TR_{opt}}$ on the other hand (takes 50 minutes to generate) has a cost of 0.005. The mean trajectory has 26 steps and large deviations trajectory has 27 steps. Here again we observe see that both the trajectories are almost identical and so our large deviation trajectory provides us with a reliable estimate for the paths of trajectories.

We now illustrate the first rare event in the case of 4 genotypes. We create a large deviation trajectory from initial population with dominant genotype 1, $H_1 \geq 0.99$ to a target population which has frequency of genotype 3, $G_3 \geq 0.99$.

The step by step computation for large deviations trajectory is shown in following sequence of figures.

Figure 8.14 shows the possible target histograms which have $G_3 \geq 0.99$. In this case we obtain 22 neighbors using discretization of 0.5% around $[0.002, 0.003, 0.99, 0.005]$.

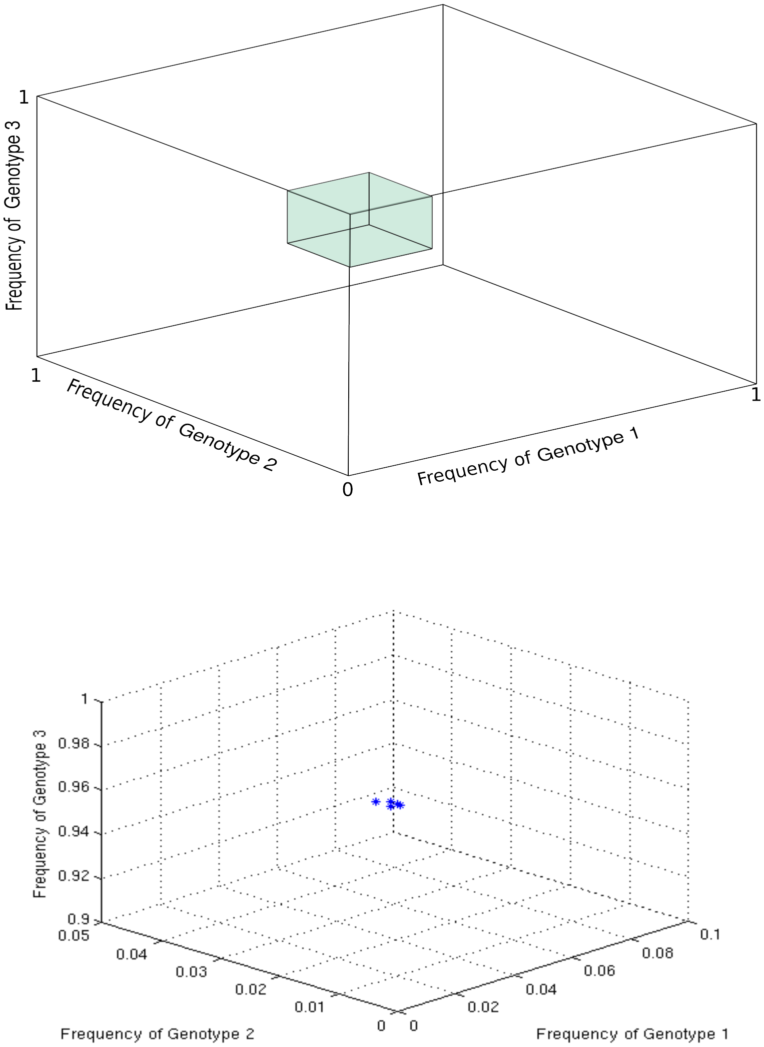

Figure 8.14: Possible Target Histograms in a Population with $g = 4$ Genotypes.

Figure 8.15 shows histograms in the neighborhood of a target terminal histogram [0.0050.0020.990.003]. They form the set Ω which is the set of all penultimate steps. In this case we obtain 166 neighbors using discretization of 0.5%.

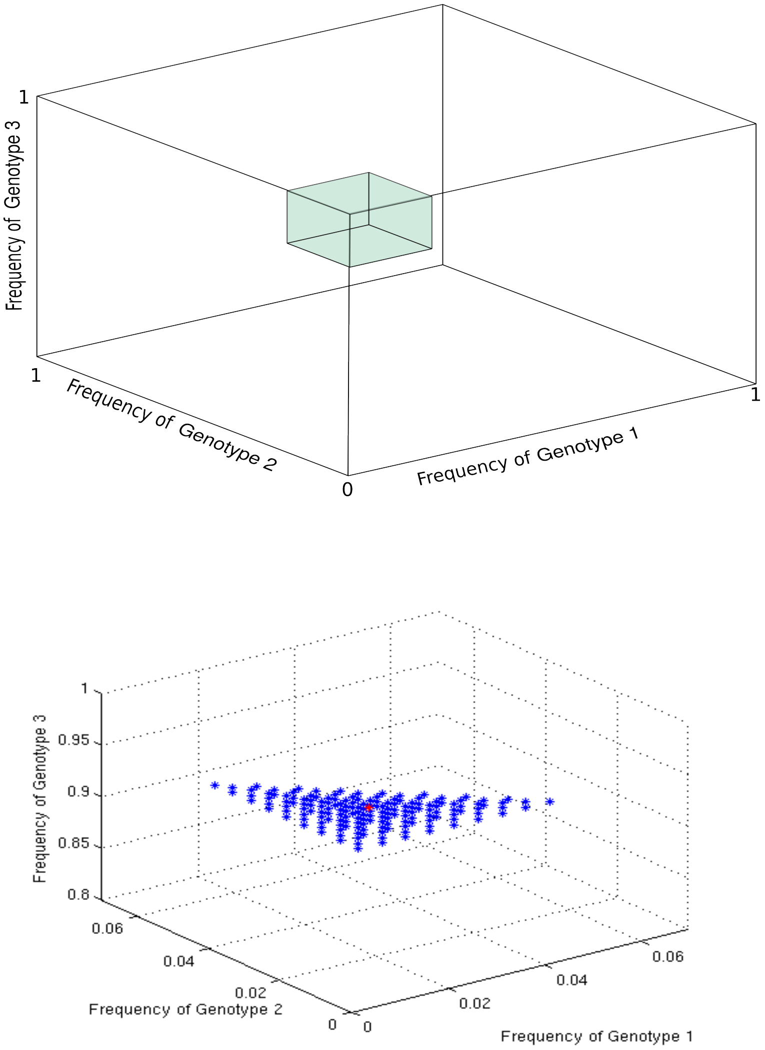

Figure 8.15: Possible Neighbors for Target Histogram.

The one-stage rate minimizing trajectory with 19 steps which takes approximately 30 seconds to generate and has cost of 0.15 can be seen in Figure 8.16.

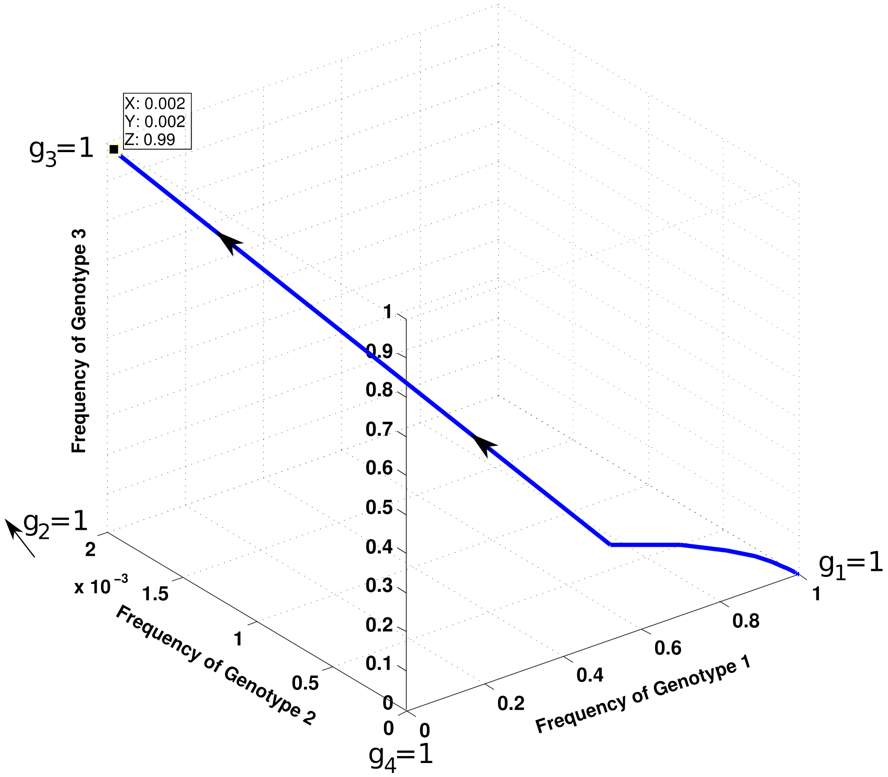

Figure 8.16: Rate Minimizing Optimal Trajectory at Step 1.

Some of the incomplete trajectories out of 20 incomplete trajectories in the set $\mathrm{TR}_{\mathrm{inc}}$ at first step are shown in Figure 8.17.

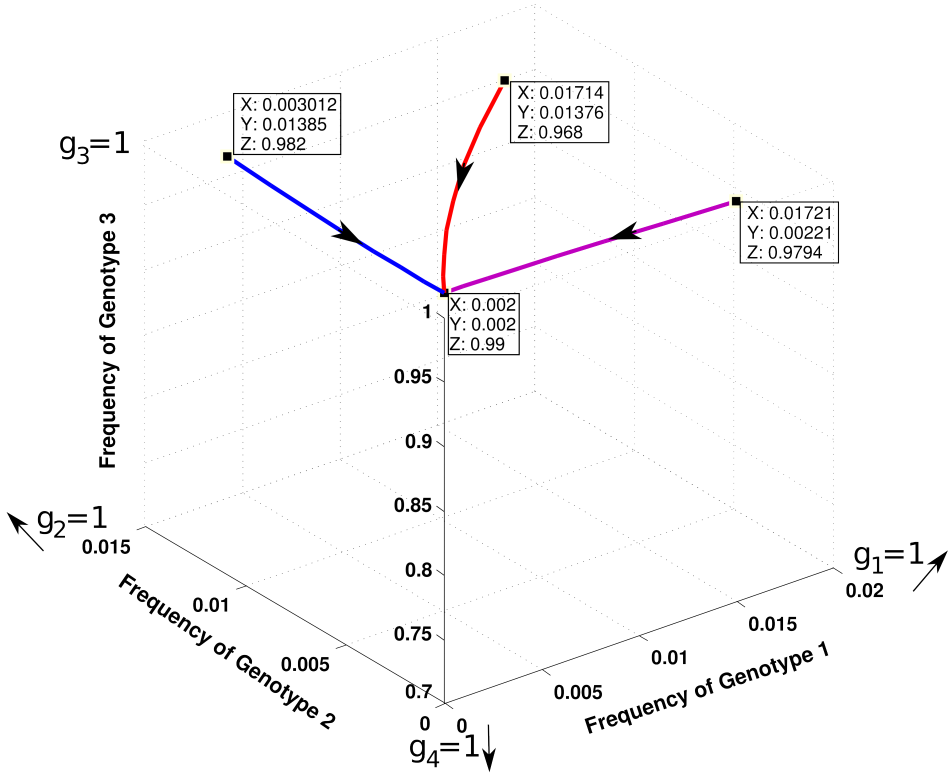

Figure 8.17: Incomplete Trajectories at Step 1

The initial point of each of these incomplete trajectories is considered a potential new target point $newH_{\text{tar}}$. Penultimate step histogram at this stage can be chosen from the set $n\Omega_H$, where $n\Omega_H$ is the set of neighborhood histograms around $newH_{\text{tar}}$ and is presented in figure 8.18.

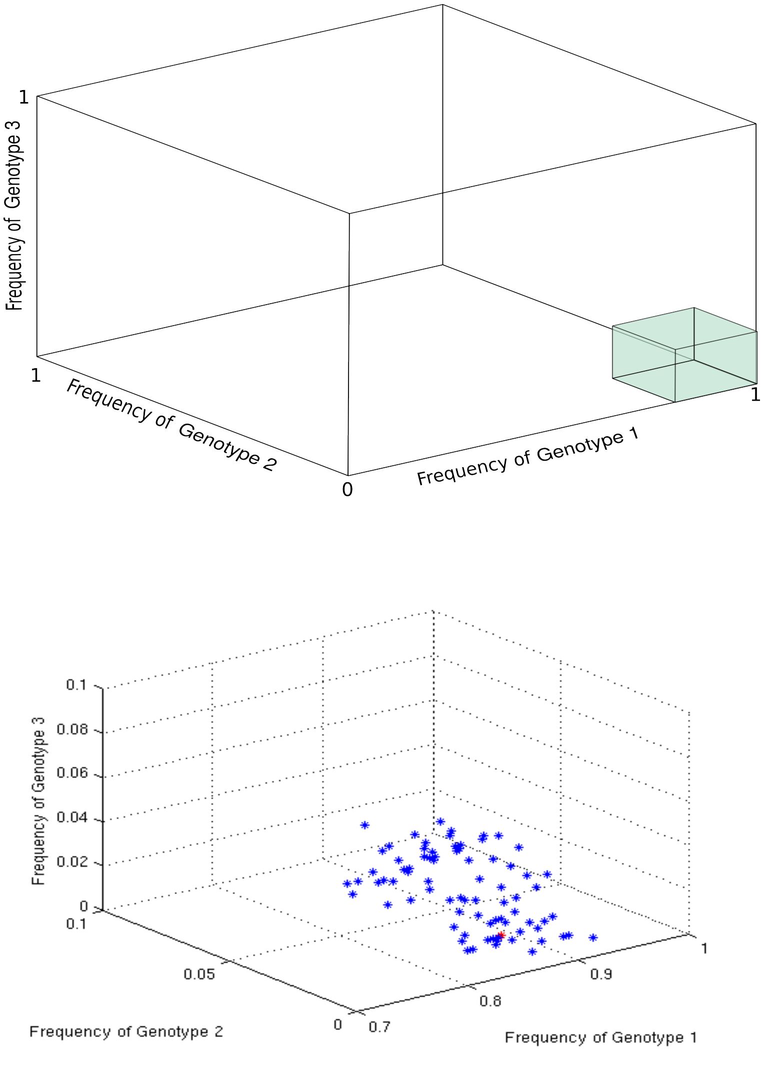

Figure 8.18: State Space $n\Omega_H$ for Possible Target Histogram in the Incomplete Trajectory

So, figure 8.19 shows multi-stage optimal rate minimizing trajectory, $\mathrm{TR}_{\mathrm{opt}}$ obtained.

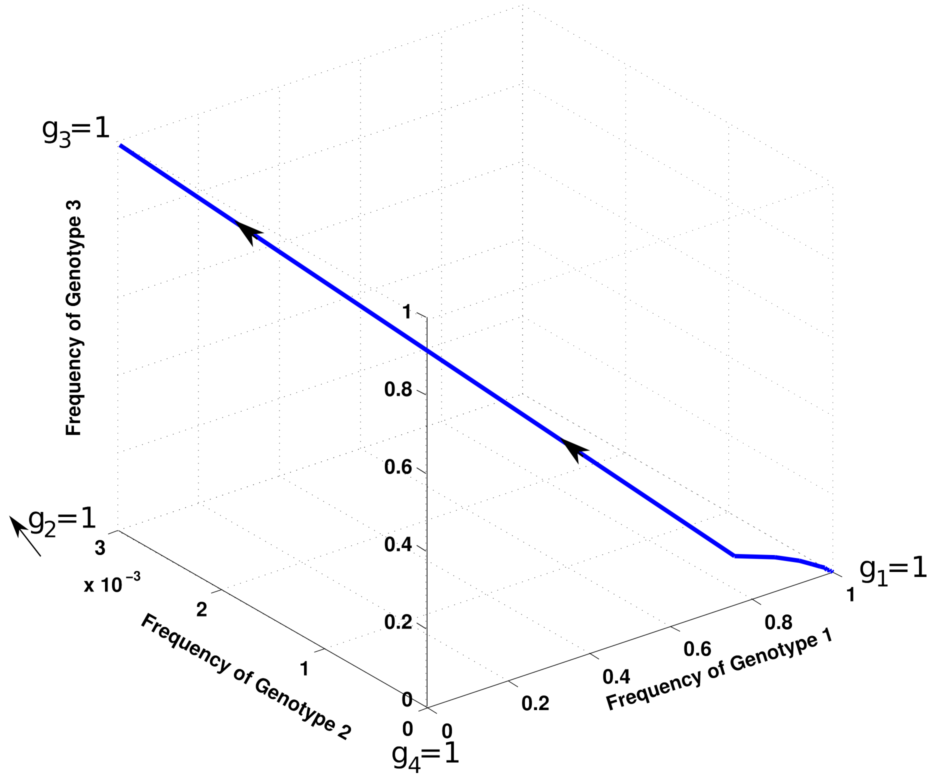

Figure 8.19: Rate Minimizing Optimal Trajectory in Population with $g = 4$ Genotypes.

Rate minimizing optimal trajectory, $\mathrm{TR}_{\mathrm{opt}}$ in this case (takes 81 minutes to generate) has a cost of 0.09.

Next we illustrate the second rare event in the case of 4 genotypes. We create a rate optimizing trajectory from population with dominant genotype 1, $H_1 \geq 0.99$ to target population with dominant genotype 2, $G_2 \geq 0.99$.

The step by step computation is shown in following sequence of figures. Figure 8.20 shows the possible target histograms which have $G_2 \geq 0.99$. In this case we obtain 23 neighbors using discretization of 0.3% around $[0.002, 0.99, 0.003, 0.005]$.

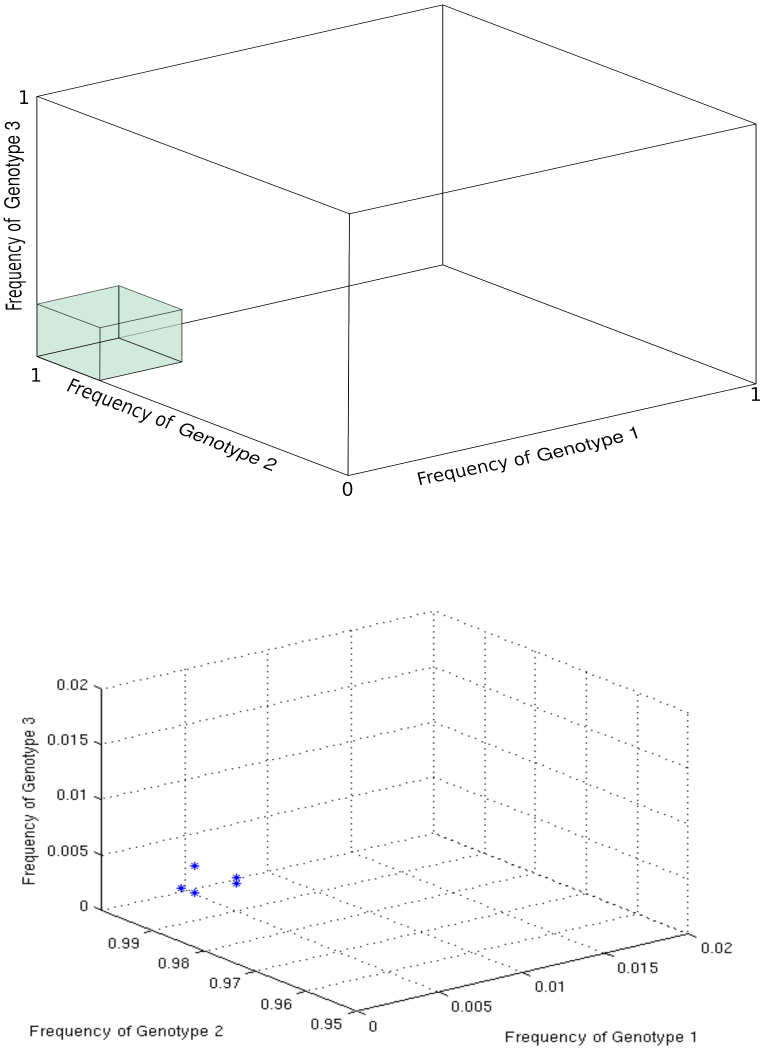

Figure 8.20: Possible Target Histograms in a Population with $g = 4$ Genotypes.

Figure 8.21 shows histograms in the neighborhood of a target terminal histogram [0.0050.0020.990.003]. They form the set Ω which is the set of all penultimate steps. In this case we obtain 1017 neighbors using discretization of 0.3% around $[0.002, 0.99, 0.002, 0.006]$.

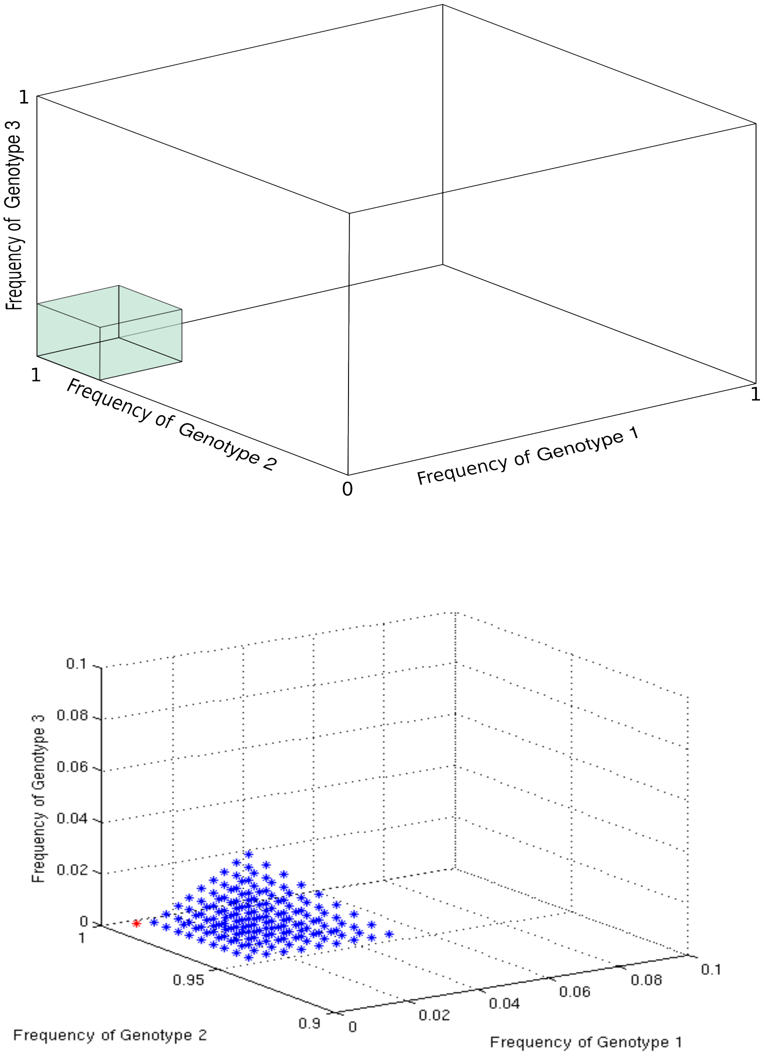

Figure 8.21: Possible Neighbors for Target Histograms

The one-stage rate minimizing trajectory with 21 steps which takes approximately 50 seconds to generate and has cost of 0.41 can be seen in Figure 8.22.

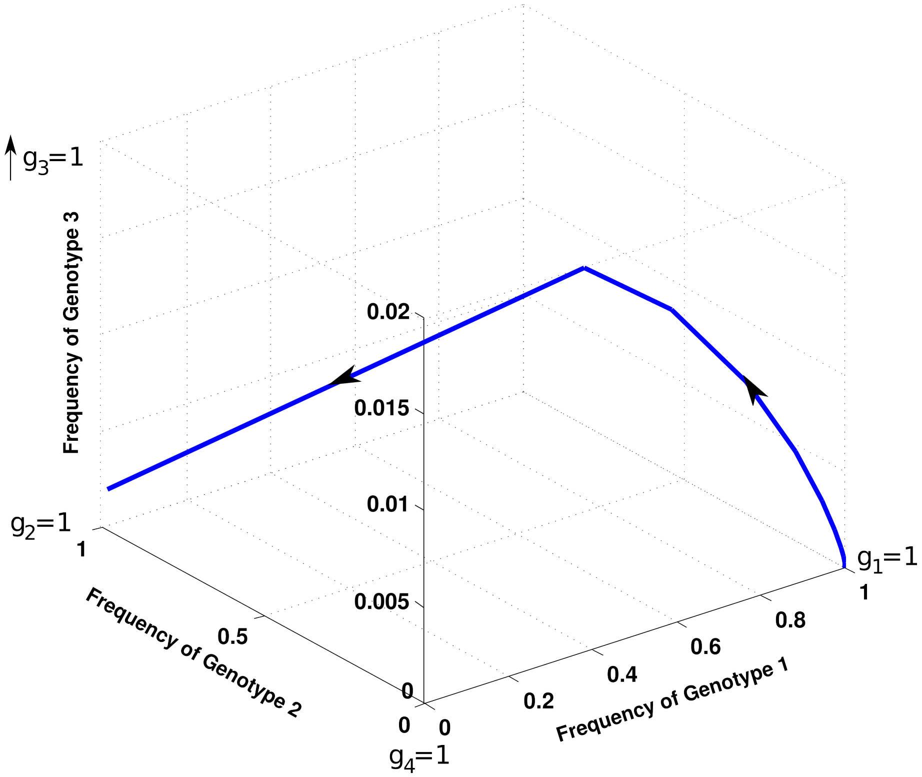

Figure 8.22: Rate Minimizing Optimal Trajectory at Step 1

Some of the incomplete trajectories out of 97 incomplete trajectories in the set $\mathrm{TR}_{\mathrm{inc}}$ at first step are shown in Figure 8.23.

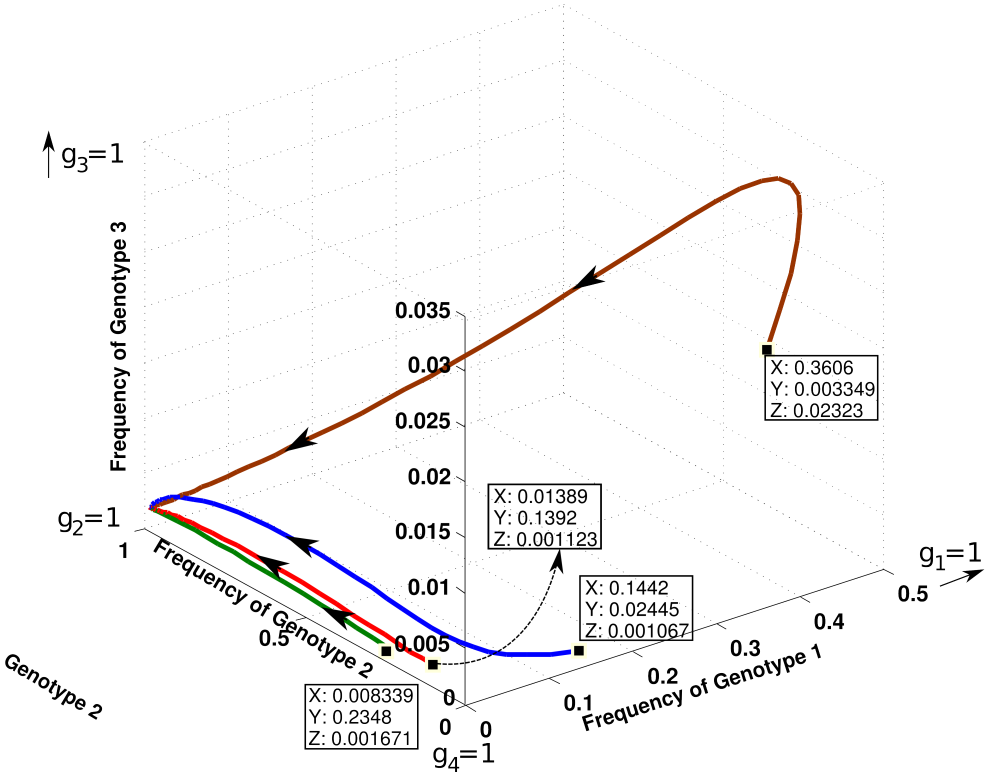

Figure 8.23: Incomplete Trajectories at Step 1

The initial point of each of these incomplete trajectories is considered a potential new target point $newH_{\text{tar}}$. Penultimate step histogram at this stage can be chosen from the set $n\Omega_H$, where $n\Omega_H$ is the set of neighborhood histograms around $newH_{\text{tar}}$ and is presented in figure 8.24.

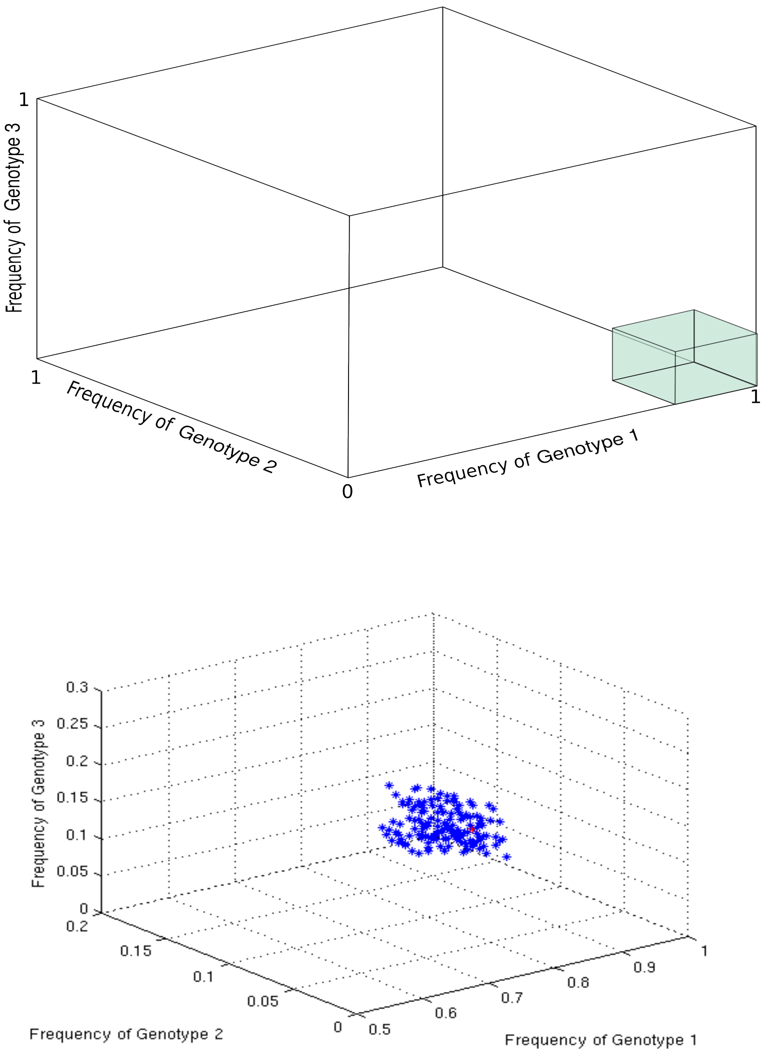

Figure 8.24: State Space $n\Omega_H$ for Possible Target Histogram in the Incomplete Trajectory

So, figure 8.25 shows multi-stage optimal rate minimizing trajectory, $\mathrm{TR}_{\mathrm{opt}}$ obtained.

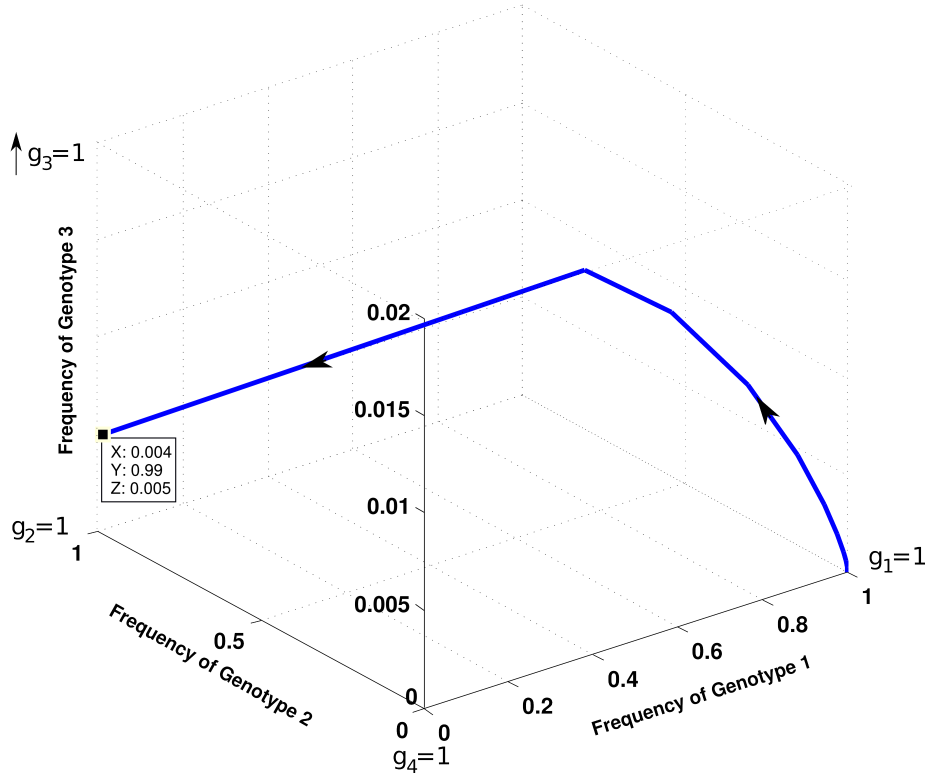

Figure 8.25: Rate Minimizing Optimal Trajectory in Population with $g = 4$ Genotypes.

Rate minimizing optimal trajectory, $\mathrm{TR}_{\mathrm{opt}}$ in this case (takes 16 hours to generate) has a cost of 0.42.

The costs and number of steps associated to rate minimizing trajectories in this case are tabulated in Table 8.7.

Initial Genotype	Target Genotype	Number of Steps	Cost
1	2	18	0.42
1	3	18	0.09
1	4	27	0.005

Table 8.7: Rate Minimizing Trajectories for TC Experiment

CHAPTER 9

Direct Simulation of the Stochastic Evolution Dynamics

In this chapter we simulate and present the results of direct simulation of the population growth model described in chapter 2. The process parameters will be given by TC experiments, as stated in Tables 9.1 and 9.6.

We generate random sets of $K = 10^4$ population histograms trajectories and compute the empirical frequencies of key evolutionary events such as near fixation of specific genotypes. This is done for the cases of 2, 3, and 4 genotypes.

We simulate the population model described in chapter 2 with locked box dynamics. In this experimental stochastic population evolution, the population evolves over generations with daily growth + dilution cycles. During the growth period, the number of cells increases from the initial value N to N_{sat}. Next, random independent mutations, governed by the vector $\vec{m} = (m_j)$ of mutation rates, are implemented in the intermediary population. The number of mutations is assumed to have a Poisson distribution dependent on the size of the population. In the numerical simulation of the model, only forward mutations are allowed (i.e.), mutants can only evolve into fitter genotype. This makes mutation process unidirectional and irreversible.

Next, to generate the new population, one performs a "dilution" of the population by extracting from intermediary population a random sample of size N. For this sub sampling we avoid the direct use of multinomial sampling function in MATLAB since it involves computing factorials. Calculating factorials is not an efficient method with the large population sizes, N considered (N ranging from 50000 to 500000) in our case. So, instead we implement the inbuilt MATLAB subroutine '*mnrnd*' which utilizes the fact that the marginals for multinomial are binomial.

Simulations are done for a population starting with pure concentration of genotype of lowest growth factor (genotype 1 in our case) as initial state.

Given a target genotype, we want to estimate probability for the evolving population to reach a stage where the frequency of target genotype is almost 1. In reality, however one rarely achieves 100% fixation of target genotype in the population. Denote by fTH the threshold fixation frequency of the target genotype in the terminal state of the evolutionary trajectories.

The evolution trajectories are then sorted on the basis of the observed genotype with frequency$\geq fTH$ in the population histogram of the terminal state. We are interested in estimating probability, $P_N(A)$ where A is the event that required target genotype has a frequency$\geq fTH$ in the terminal state of trajectory.

We present results for evolution trajectories in the population with $g = 4$ genotypes and restricted random mutation matrix. The process parameters used are given in Table 9.1.

Genotype	Growth Factor	Mutation Rate
1	$200^{1.06}$	0.5×10^{-6}
2	$200^{1.08}$	0.5×10^{-6}
3	$200^{1.1}$	0.5×10^{-6}
4	$200^{1.12}$	0.5×10^{-6}

Table 9.1: Parameters for Direct Simulation with $g = 4$ Genotypes

Recall that N is the population size and we present results for 3 population sizes, $N = 5 \times 10^4$, 2×10^5 and 5×10^5.

1. For initial histogram $[1, 0, 0, 0]$, the Table 9.2 displays the fixation probabilities of each genotype, for a fixation threshold, $fTH = 90\%$.

Fixation Genotype	$N = 5 \times 10^4$	$N = 2 \times 10^5$	$N = 5 \times 10^5$
1	0	0	0
2	0.03	10^{-4}	0
3	0.25	0.09	0.003
4	0.72	0.9	0.997

Table 9.2: Probability of Reaching Fixation for all Genotypes in Simulated Trajectories with 4 Genotypes and $fTh = 90\%$ as the Fixation Threshold.

2. For initial histogram $[1, 0, 0, 0]$, the Table 9.3 displays the fixation probabilities of each genotype, for a fixation threshold, $fTH = 95\%$.

Fixation Genotype	$N = 5 \times 10^4$	$N = 2 \times 10^5$	$N = 5 \times 10^5$
1	0	0	0
2	0.03	0	0
3	0.27	0.06	0.003
4	0.7	0.94	0.997

Table 9.3: Probability of Reaching Fixation for all Genotypes in Simulated Trajectories with 4 Genotypes and $fTh = 95\%$ as the Fixation Threshold.

3. For initial histogram $[1, 0, 0, 0]$, the Table 9.4 displays the fixation probabilities of each genotype, for a fixation threshold, $fTH = 98\%$.

Fixation Genotype	$N = 5 \times 10^4$	$N = 2 \times 10^5$	$N = 5 \times 10^5$
1	0	0	0
2	0.03	0	0
3	0.25	0.06	0.003
4	0.72	0.94	0.997

Table 9.4: Probability of Reaching Fixation for all Genotypes in Simulated Trajectories with 4 Genotypes and $fTh = 98\%$ as the Fixation Threshold.

Recall that $K = 10^4$ is the total number of simulated trajectories, p is the empirical estimate of fixation probability of a genotype in the simulated trajectories and $fTH$is the fixation threshold fTH.

These probabilities are almost similar for all the three thresholds we selected and the estimation error on these empirical estimates of fixation probabilities estimated is given by

$$\sqrt{\frac{p(1-p)}{K}}$$

where $Kp \geq 5$. However when $Kp < 5$, we compute the estimation error as follows.

We know that for a Binomial distribution, $B(K, p)$ where K is very large and p is very small the appropriate binomial probabilities can be approximated by way of the Poisson probability function with mean Kp [46].

So, for our rare events with very small non-zero probability and 0 empirical probability estimate, we estimate error by computing a one-sided 95% confidence interval for the true mean using the Poisson probability with estimated mean Kp.

Hence we are interested in estimating $\hat{p}$ for which

$$e^{-K\hat{p}} = 0.05.$$

This gives $\hat{p} = 3 \times 10^{-4}$ as the one-sided (upper) bound for the estimation error when $p = 0$ and $K = 10^4$.

The errors are presented in Table 9.5.

Fixation Genotype	$N = 5 \times 10^4$	$N = 2 \times 10^5$	$N = 5 \times 10^5$
1	0	0	0
2	0.002	3×10^{-4}	3×10^{-4}
3	0.004	0.002	5.4×10^{-4}
4	0.004	0.002	5.4×10^{-4}

Table 9.5: Table for Errors in Estimated Probabilities for all Genotypes.

It can be observed that as we increase the population size from $N = 5 \times 10^4$ to $N = 5 \times 10^5$, the fixation probability for the fittest genotype becomes almost 1 and fixation probabilities for other genotypes tend to 0. Hence, fixation of any genotype which is not the fittest becomes a rare event as $N \to \infty$. In other words the fixation probability for all genotypes except for the fittest decrease very quickly with N as $N \to \infty$.

This is the main application of large deviation algorithms implemented in previous chapters. The probability of a rare event A decays, $P_N(A) \to 0$, at an exponential rate given by the rate functional λ where $\lambda = \min_{traj \in A} \text{Cost(traj)}$.

Next, we illustrate results for evolution trajectories in the population with $g = 3$ genotypes and restricted random mutation matrix. The process parameters used are given in Table 9.6.

Genotype	Growth Factor	Mutation Rate
1	200	0.5×10^{-6}
2	$200^{1.08}$	0.5×10^{-6}
3	$200^{1.15}$	0.5×10^{-6}

Table 9.6: Parameters for Direct Simulation with $g = 3$ Genotypes

1. For the initial histogram $[1, 0, 0]$, the Table 9.7 displays the fixation probabilities of each genotype, with a fixation threshold, $fTH = 90\%$.

Fixation Genotype	$N = 5 \times 10^4$	$N = 2 \times 10^5$	$N = 5 \times 10^5$
1	0	0	0
2	0.25	0.06	0.003
3	0.75	0.94	0.997

Table 9.7: Probability of Reaching Fixation for all Genotypes in Simulated Trajectories with 3 Genotypes and $fTh = 90\%$ as the Fixation Threshold.

2. For the initial histogram $[1, 0, 0]$, the Table 9.8 displays the fixation probabilities

of each genotype, with a fixation threshold, $fTH = 95\%$.

Fixation Genotype	$N = 5 \times 10^4$	$N = 2 \times 10^5$	$N = 5 \times 10^5$
1	0	0	0
2	0.26	0.06	0.003
3	0.74	0.94	0.997

Table 9.8: Probability of Reaching Fixation for all Genotypes in Simulated Trajectories with 3 Genotypes and $fTh = 95\%$ as the Fixation Threshold.

3. For the initial histogram $[1, 0, 0]$, the Table 9.9 displays the fixation probabilities of each genotype, with a fixation threshold, $fTH = 98\%$.

Fixation Genotype	$N = 5 \times 10^4$	$N = 2 \times 10^5$	$N = 5 \times 10^5$
1	0	0	0
2	0.25	0.06	0.003
3	0.75	0.94	0.997

Table 9.9: Probability of Reaching Fixation for all Genotypes in Simulated Trajectories with 3 Genotypes and $fTh = 98\%$ as the Fixation Threshold.

Again we observe that the fixation probabilities are almost similar for all the three thresholds we selected and the estimation error on these estimated probabilities estimated is given by Table 9.10. Recall that when $Kp < 5$, we compute the estimation error using Poisson distribution as explained in the case of 4 genotypes.

Fixation Genotype	$N = 5 \times 10^4$	$N = 2 \times 10^5$	$N = 5 \times 10^5$
1	0	0	0
2	0.004	0.002	5.4×10^{-4}
3	0.004	0.002	5.4×10^{-4}

Table 9.10: Table for Errors in Estimated Probabilities for all Genotypes.

Also as we increase the population size from $N = 5 \times 10^4$ to $N = 5 \times 10^5$, the winning probability for the fittest genotype is almost 1 and winning probabilities for other genotypes are almost 0. Hence, fixation of any genotype which is not the fittest becomes a rare event as $N \to \infty$. In other words the fixation probability for all genotypes except for the fittest decrease very quickly with N as $N \to \infty$.

As explained before, this is the main application of large deviation algorithms implemented in previous chapters. The probability of a rare event A, $P_N(A) \to 0$ at an exponential rate given by the rate functional λ where $\lambda = \min_{traj \in A} \text{Cost}(\text{traj})$.

Next we present Table 9.11 with the computing times in minutes required to simulate the above trajectories.

Number of Genotypes	$N = 5 \times 10^4$	$N = 2 \times 10^5$	$N = 5 \times 10^5$
3	13	33	67
4	28	78	174

Table 9.11: Table for CPU Times in Minutes for Estimating Probabilities for all Genotypes in Simulated Trajectories with 3 Genotypes.

The following Figure 9.1 shows histogram for length of trajectories starting at $[1, 0, 0, 0]$ required for fixation when we have $g = 4$ genotypes and the genotype 4 fixates first, i.e., $G_4 \geq 0.98$.

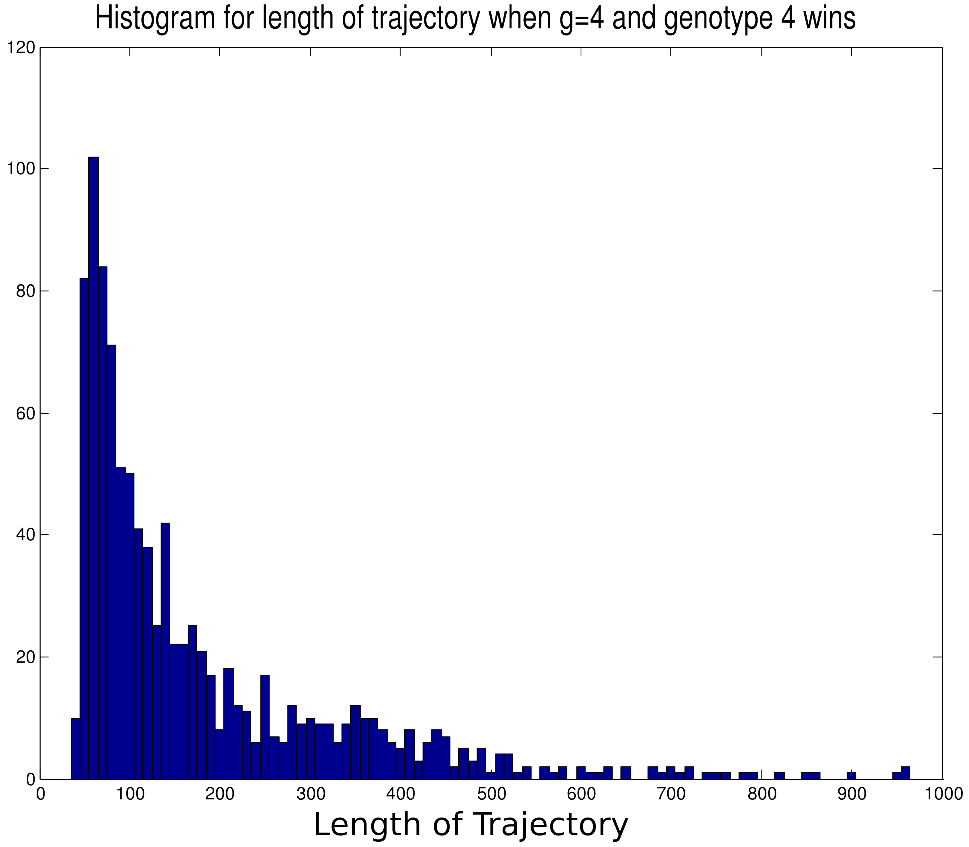

Figure 9.1: Histogram for Length of the Trajectories Observed in Direct Simulation where Initial State is $[1, 0, 0, 0]$ and Genotype 4 Wins.

The following Figure 9.2 shows histogram for length of trajectories starting at $[1, 0, 0, 0]$ required for fixation when we have $g = 4$ genotypes and the genotype 3 fixates first, i.e., $G_3 \geq 0.98$.

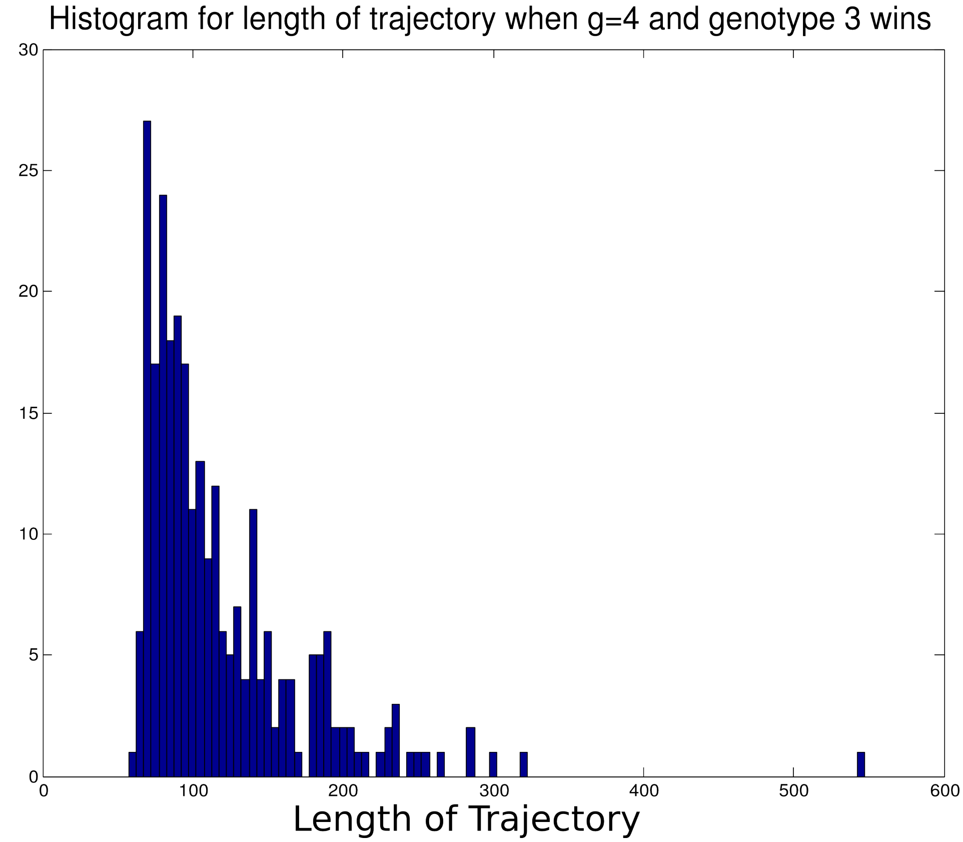

Figure 9.2: Histogram for Length of the Trajectories Observed in Direct Simulation where Initial State is $[1, 0, 0, 0]$ and Genotype 3 Wins.

The following Figure 9.3 shows histogram for length of trajectories starting at $[1, 0, 0, 0]$ required for fixation when we have $g = 4$ genotypes and the genotype 2 fixates first, i.e., $G_2 \geq 0.98$.

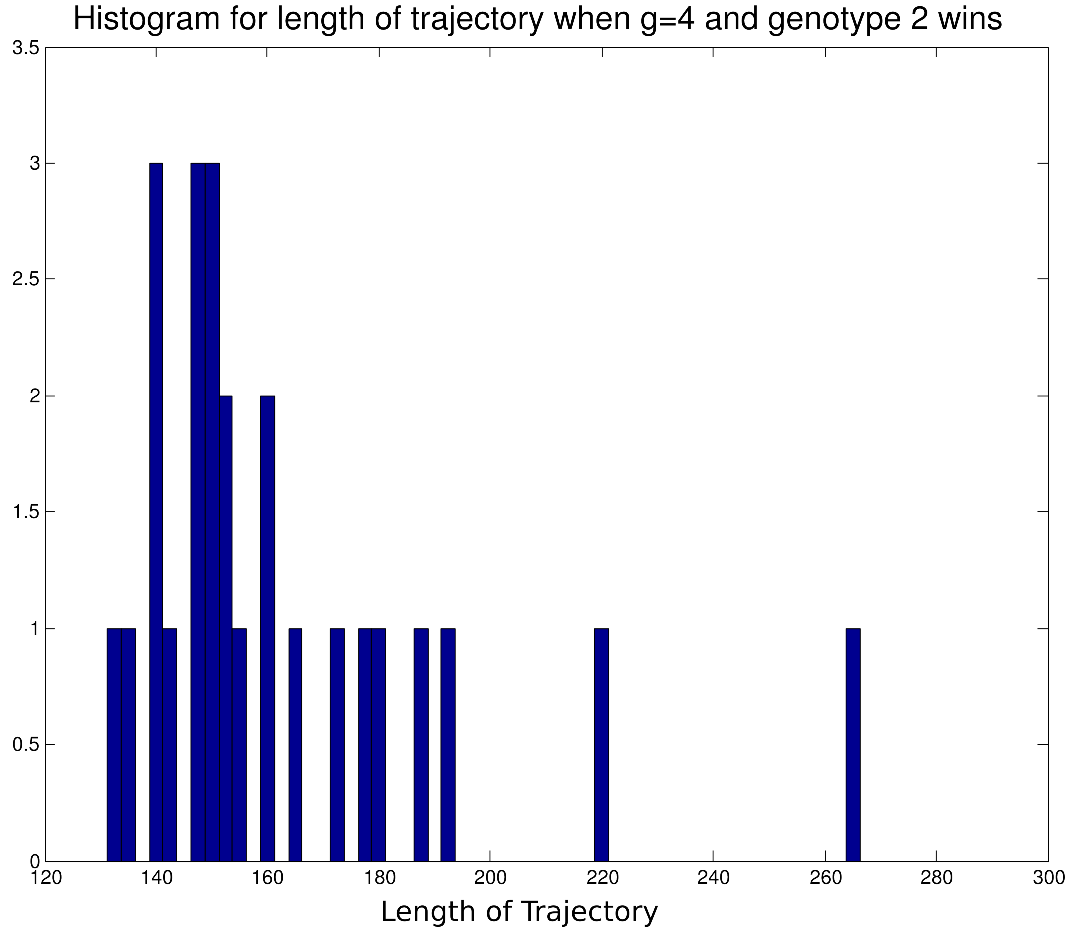

Figure 9.3: Histogram for Length of the Trajectories Observed in Direct Simulation where Initial State is $[1, 0, 0, 0]$ and Genotype 2 Wins.

The following Figure 9.4 shows histogram for length of trajectories starting at $[1, 0, 0]$ required for fixation when we have $g = 3$ genotypes and the genotype 3 fixates first, i.e., $G_3 \geq 0.98$.

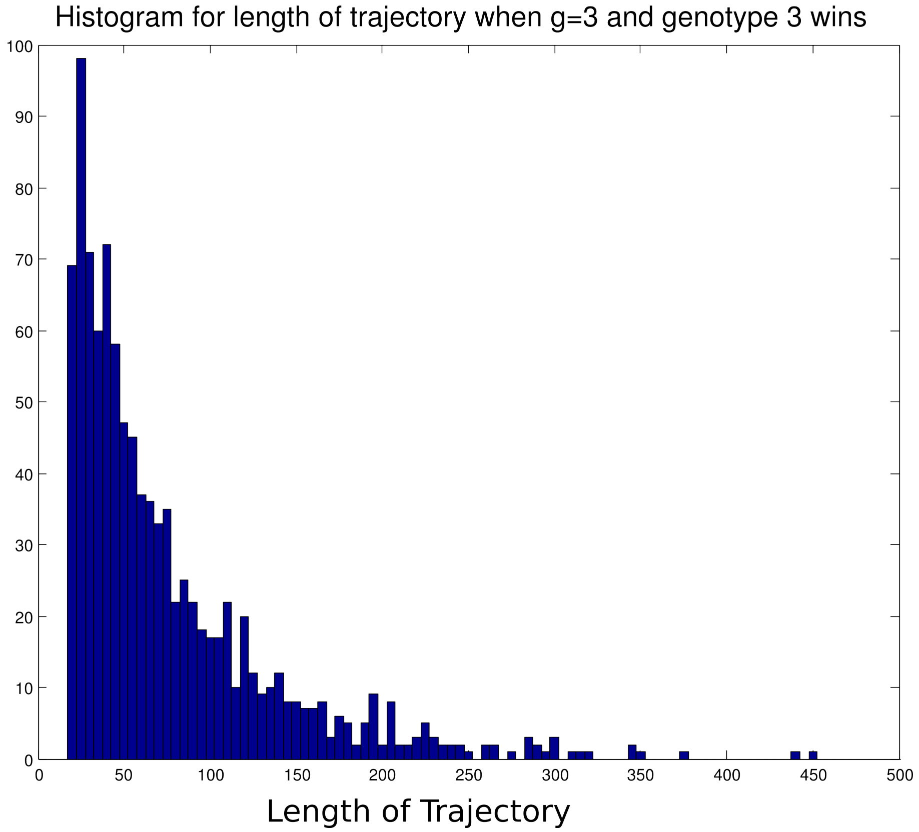

Figure 9.4: Histogram for Length of the Trajectories Observed in Direct Simulation where Initial State is $[1, 0, 0]$ and Genotype 3 Wins.

The following Figure 9.5 shows histogram for length of trajectories starting at $[1, 0, 0]$ required for fixation when we have $g = 3$ genotypes and the genotype 2 fixates first, i.e., $G_2 \geq 0.98$.

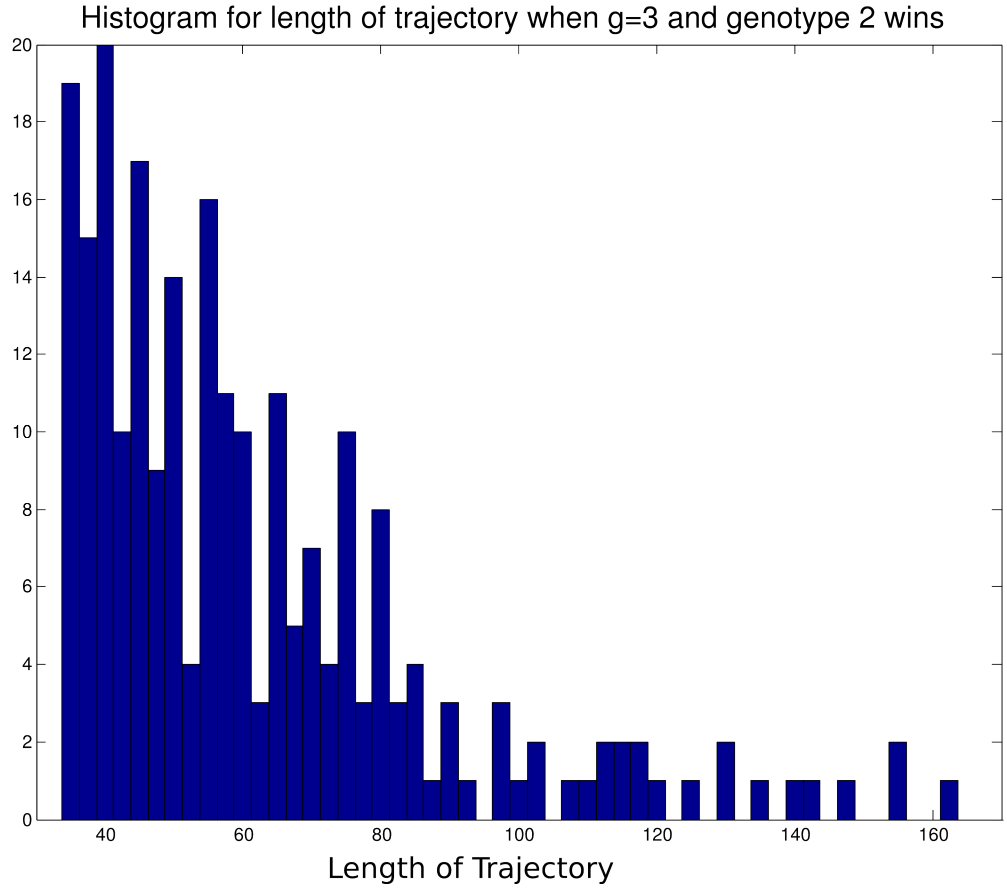

Figure 9.5: Histogram for Length of the Trajectories Observed in Direct Simulation where Initial State is $[1, 0, 0]$ and Genotype 2 Wins.

Using the rate functional formulas derived in previous chapters we compute the costs of these simulated trajectories. For each population size, quantile curve of the costs of all trajectories observed is computed. Hence we obtain 3 quantile curves for each of the 3 population sizes. These quantiles of costs for all three population sizes with 3 genotypes is shown in Figure 9.6.

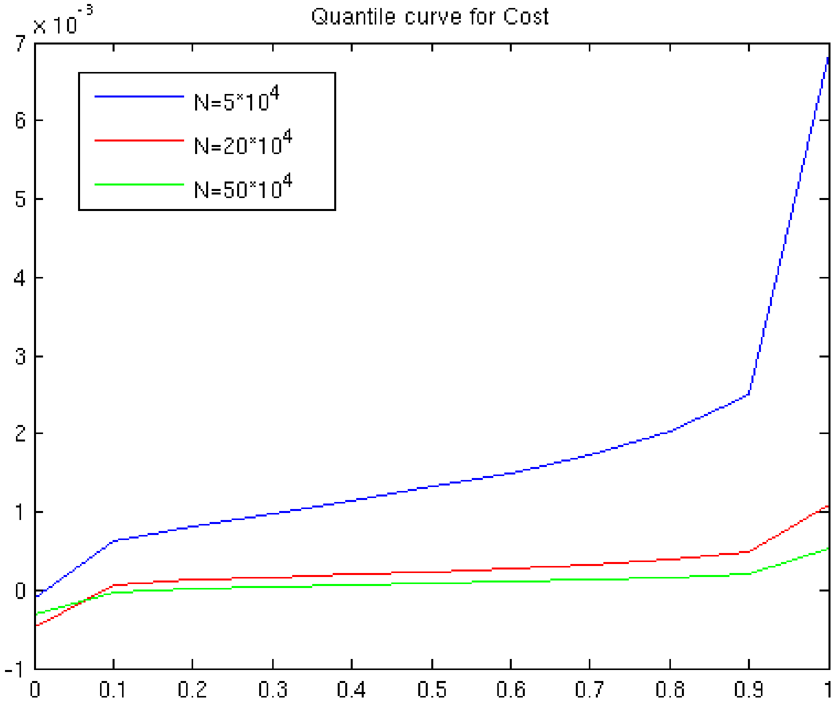

Figure 9.6: Comparison of Cost Quantiles over all Three Population Sizes with 3 Genotypes.

Similar quantiles of costs for all three population sizes with 4 genotypes are shown in Figure 9.7.

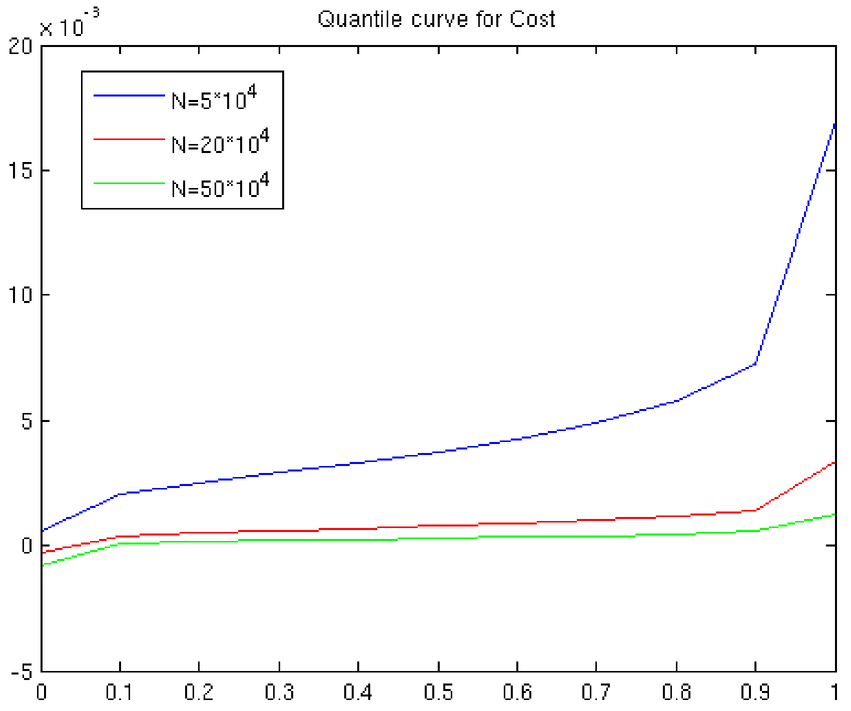

Figure 9.7: Comparison of Cost Quantiles over all Three Population Sizes with 4 Genotypes.

These quantile curves demonstrate that with an increase in population sizes, the overall cost decreases.

Figure 9.8 shows histogram for cost of trajectories observed in direct simulation starting at $[1,0,0,0]$ where the genotype 3 fixates first and $fTH = 0.98$, i.e., $G_3 \geq 0.98$ across population sizes $N = 5 \times 10^4,\ 2 \times 10^5$ and 5×10^5.

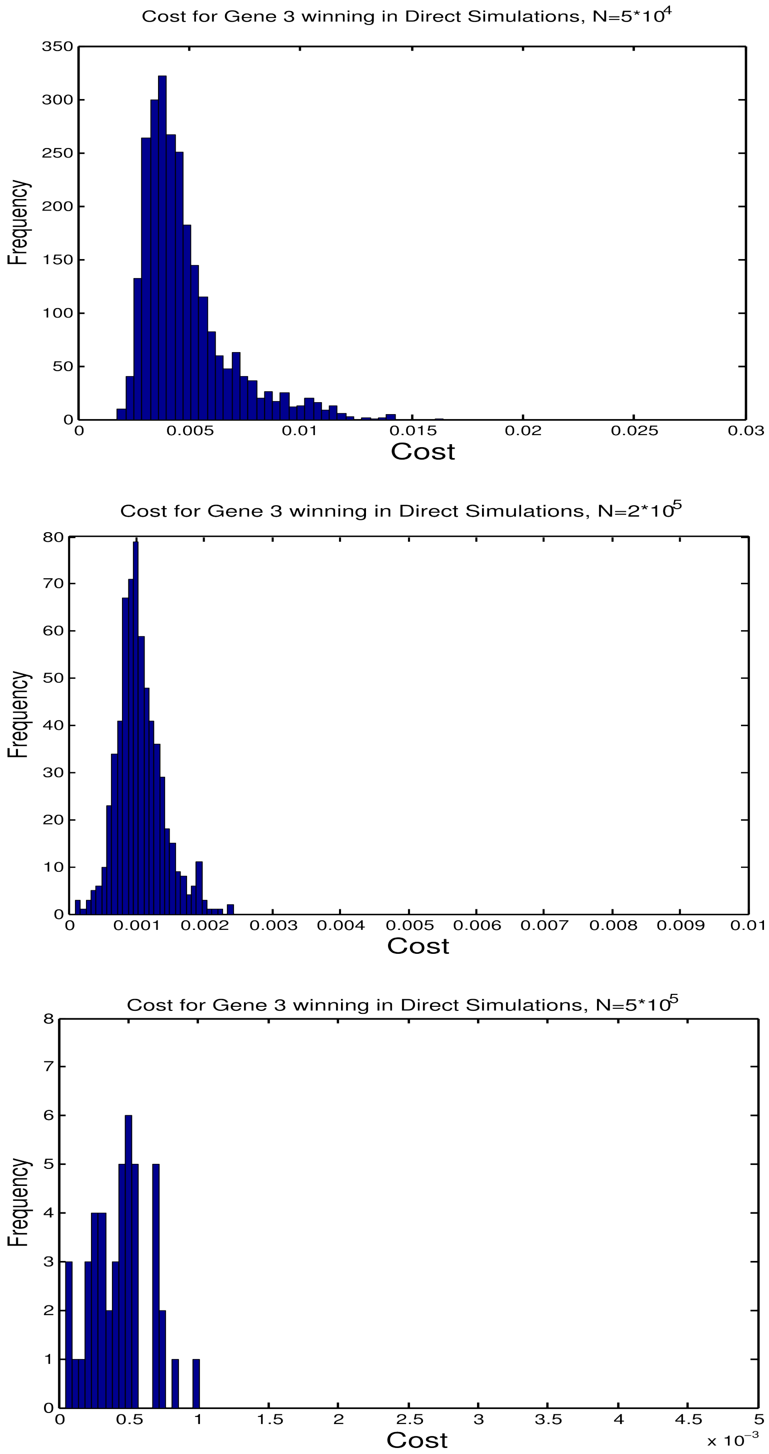

Figure 9.8: The Histograms of Cost of Trajectories Observed in Direct Simulation Starting at $[1,0,0,0]$ where the Genotype 3 Fixates First and $fTH = 0.98$ across Population Sizes $N = 5 \times 10^4,\ 2 \times 10^5$ and 5×10^5.

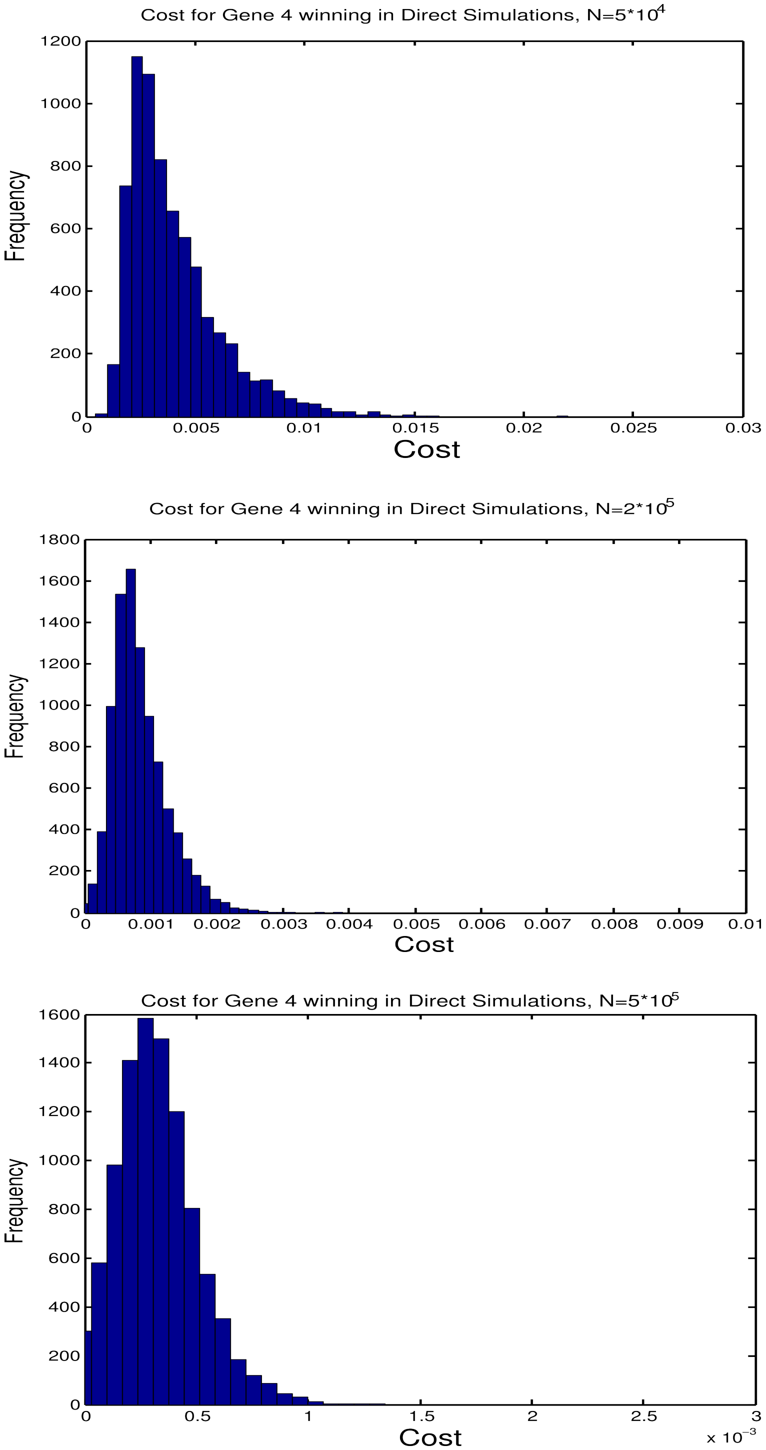

Figure 9.9: The Histograms of Cost of Trajectories Observed in Direct Simulation Starting at $[1, 0, 0, 0]$ where the Genotype 4 Fixates First and $fTH = 0.98$ across Population Sizes $N = 5 \times 10^4,\ 2 \times 10^5$ and 5×10^5.

Similarly Figure 9.9 shows histogram for cost of trajectories observed in direct simulation starting at $[1, 0, 0, 0]$ where the genotype 4 fixates first and $fTH = 0.98$, i.e., $G_4 \geq 0.98$ across population sizes $N = 5 \times 10^4,\ 2 \times 10^5$ and 5×10^5.

It can be observed that in both cases, as N increases from 5×10^4 to 5×10^5 the x-axis is shrinking or the the cost histograms are becoming narrower. Thus as population size increases we expect the evolution to almost always follow the most likely path and have very less deviations.

Next, we present a quantile comparison for length of trajectories required to reach fixation frequency, fTH by target genotype in terminal state.

For the various fixation thresholds, $fTH = 0.98, 0.95, 0.90$ and number of genotypes, $g = 3$ the quantile curves comparing trajectory lengths are shown in Figure 9.10.

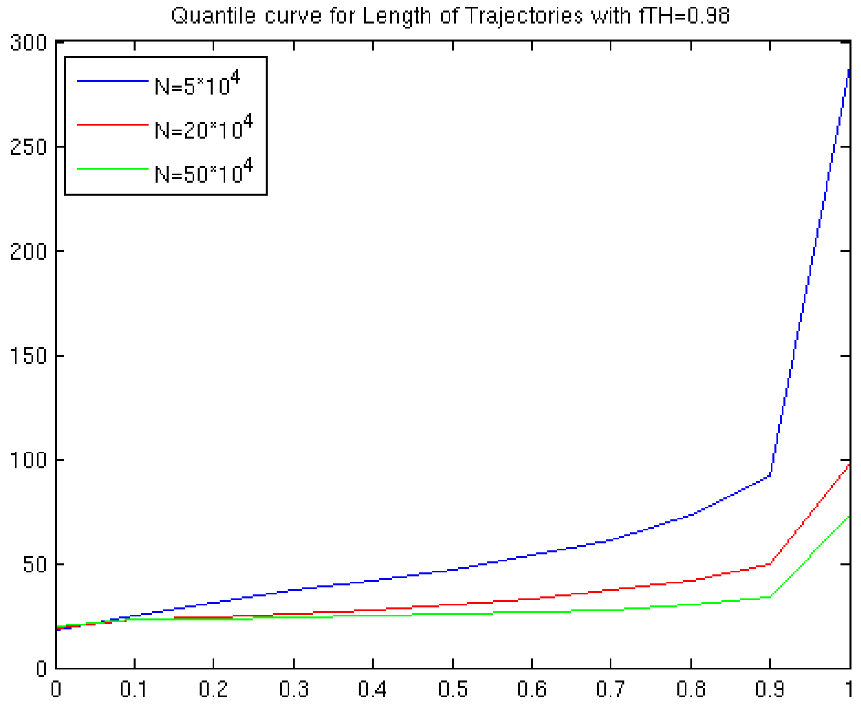

Figure 9.10: Comparison of Trajectory Length Quantiles over all Three Population Sizes with 3 Genotypes and $fTh = 0.98$.

Similarly, for the fixation threshold, $fTH = 0.98$ and number of genotypes, $g = 4$ the quantile curves comparing trajectory lengths are shown in Figure 9.11.

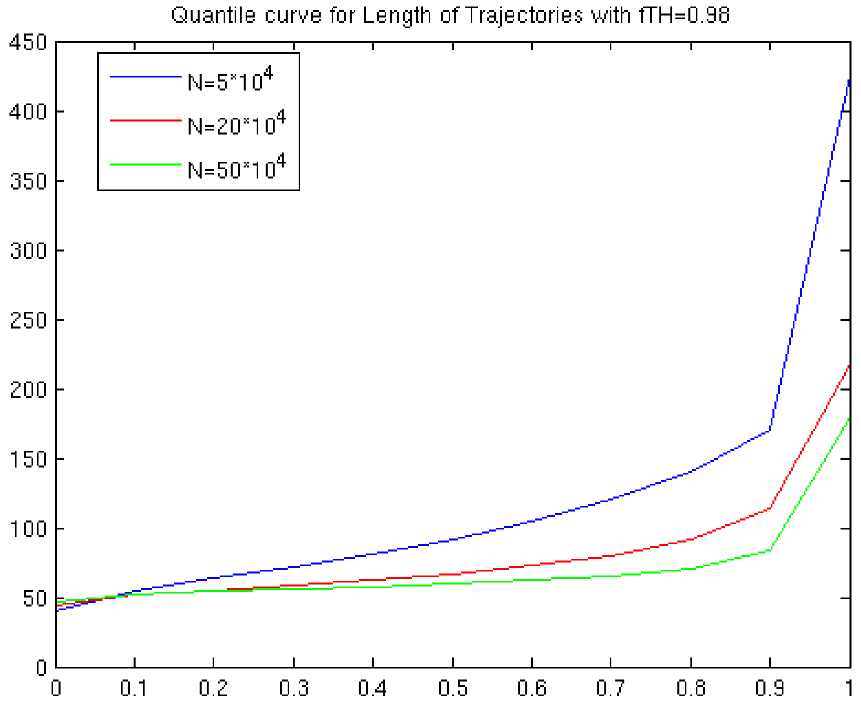

Figure 9.11: Comparison of Trajectory Length Quantiles over all Three Population Sizes with 4 Genotypes and $fTh = 0.98$.

Again we observe the decrease in the trajectory length required for fixation as we increase the population sizes.

CHAPTER 10

Clonal Interference in Mutational Trajectories

This chapter refers to a currently submitted paper prepared by a team of UH biologists Ricardo Azevedo, Tiago Paixão et al. [94]. The paper (yet to be published) discusses the effect of population size, mutational effects and epistasis on the repeatability of evolution on simple adaptive landscapes. We concentrate primarily on applying our large deviations approach to study the repeatability of evolution (hereafter CI model) using parameters from laboratory experiments on populations of bacterium *Escherichia coli* performed by Tim Cooper et al.[129] at the University of Houston Biology Department.

10.1 Adaptive Landscape Model

Wright [127] defined adaptive landscape using a genotype-to-fitness, $(G \rightarrow F)$ map and specification of how genotypes are connected. According to Robert Skipper [114] "Sewall Wright's adaptive landscape is the most influential heuristic in evolutionary biology."

Introduced by Wright, the fitness landscape describes the possible mutational trajectories by which lineages evolve in a step wise manner from genotypes that lie in regions of low fitness to ones of higher fitness [127, 128]. When viewed as a whole, this metaphorical landscape represents a species' possible paths of adaptive evolution towards the optimal genotype in a particular environment.

Thus, fitness landscapes illustrate possible steps adaptive evolution can take to increase the evolutionary fitness of individuals within a population, and the accessibility of the fittest point on the landscape is determined by the shape of the fitness landscape [37].

Evolutionary transitions from one phenotype to another are mediated by mutations to their underlying genotypes. The space of all genotypes can be considered as a mutational network with each genotype as a node and mutations between genotypes as edges. In other words, any two genotypes that differ exactly by one single point mutation are connected by an edge. One can then represent phenotypes (or fitness values) as colors. Thus mutational networks capture patterns of mutational connectivity among genotypes and phenotypes [127, 26].

Historically, mutations have been thought of by evolutionary biologists in terms of distributions of fitness effects and a lot of interest has been shown to measure the fractions of mutations that are typically beneficial, neutral, and deleterious. Although these distributions critically determine the local evolutionary dynamics, they provide little information about processes on larger scale. For this purpose, it is useful to think in terms of mutational paths connecting distant genotypes and, more generally, in terms of the large scale patterns of mutational connectivity within genotype spaces.

In this model we use mutational networks defined using L fitness loci, each with K alleles. There are K^L genotypes since typical genotypes are arbitrary configuration of alleles. We consider $L = 3$ loci (hereafter referred to as A, B and C) and $K = 2$ alleles ($A/a, B/b$ and C/c). This means that a gene, say A is allowed to express itself in $K = 2$ ways as A or a.

The complete network for mutational trajectories [94] to be generated are shown in the Figure 10.1.

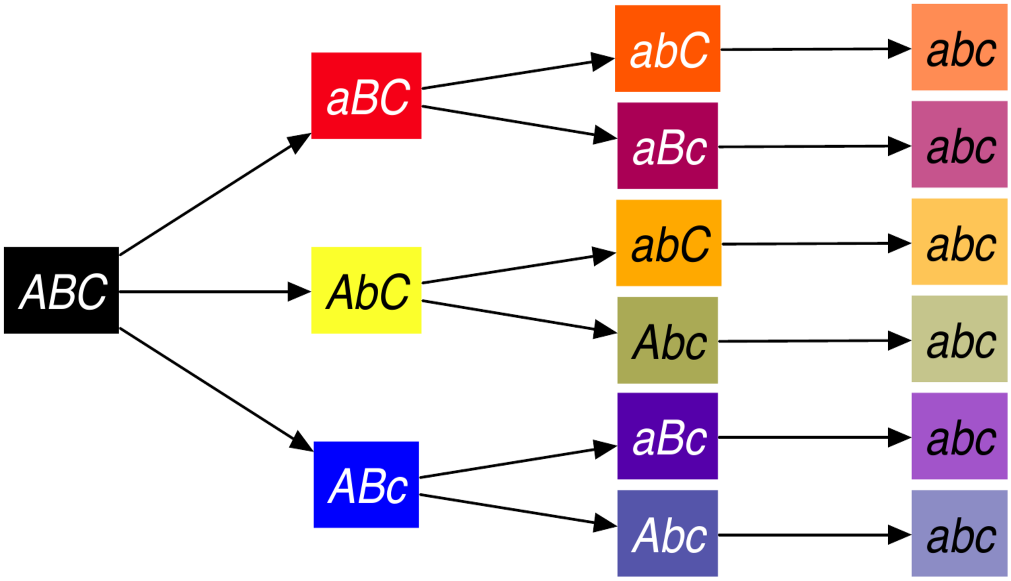

Figure 10.1: Network of Mutational Trajectories in the Model.

We consider the case where we have only irreversible mutation available to the population, and that mutants cannot mutate further. For instance, consider a single fitness locus A, with two alleles, A and a, with A for the ancestor allele and a for the derived allele. Then A alleles mutate irreversibly to a alleles and a cannot mutate back into A. These mutations are independent events. The mutant allele confers a selective advantage $s_a > 0$.

A mutant arising with selective advantage s to the ancestor genotype has by definition a multiplicative growth factor per time interval given by F^{1+s} where F is the growth factor of the ancestor genotype. For details about the range of fitness values estimated experimentally by TC experiments refer to chapter 2.

Selective advantages are essentially the basis for evolution by natural selection. They are important characteristics of organisms enabling them to survive and reproduce better than other organisms in a given environment.

Hence, the relative fitness of the ancestral genotype ABC is set to be $w_{ABC} = 1$ and every derived allele increases the fitness by s_a, s_b, s_c respectively. Without loss of generality, we assume that $s_a > s_b > s_c$.

Table 10.1 lists the relative fitnesses of all genotypes, evaluated in terms of s_a, s_b, and s_c by the same formulas as in the paper [94].

Genotype	Relative Fitness
ABC	1
aBC	$1 + s_a$
AbC	$1 + s_b$
ABc	$1 + s_c$
Abc	$(1 + s_b)(1 + s_c)$
aBc	$(1 + s_a)(1 + s_c)$
abC	$(1 + s_a)(1 + s_b)$
abc	$(1 + s_a)(1 + s_b)(1 + s_c)$

Table 10.1: Relative Fitnesses for all Genotypes.

So the genotype *abc* containing all derived alleles is assumed to be the fittest.

The emergence and fixation of the fittest genotype *abc* can be realized by six mutational trajectories where each trajectory is characterized by the order of successive emergence and fixation of the alleles a, b, c in the population.

1. $a \rightarrow b \rightarrow c$,
2. $a \rightarrow c \rightarrow b$,
3. $b \rightarrow a \rightarrow c$,
4. $b \rightarrow c \rightarrow a$,
5. $c \rightarrow a \rightarrow b$,
6. $c \rightarrow b \rightarrow a$.

corresponding to each possible mutational trajectory (Figure 10.1) based on the order in which derived alleles were acquired to reach *abc*. In the study [94], the authors investigate by intensive simulations the effects of population size on particular mutational trajectory being followed during adaptive evolution.

We consider population sizes from $N = 50000$ to $N = 10^7$ and estimate the probabilities for mutational trajectory "tr", in terms of a corresponding Cost(tr) which we estimate below by computing adequate trajectories minimizing the rate functional introduced earlier in chapter 6.

10.2 Numerical Adaptation of CI Model

We estimate the rate minimizing trajectories in 2 stages. In the first stage, the evolutionary trajectory starts from a pure ancestral genotype ABC to reach a population with only one derived allele. All the possibilities at this stage are listed below.

1. $ABC \rightarrow aBC$,
2. $ABC \rightarrow AbC$,
3. $ABC \rightarrow ABc$.

Then in the second stage, a second derived allele emerges and fixates in the population. Again the possibilities are

1. $aBC \rightarrow abC$,
2. $aBC \rightarrow aBc$,
3. $AbC \rightarrow abC$,
4. $AbC \rightarrow Abc$,
5. $ABc \rightarrow aBc$,
6. $ABc \rightarrow Abc$.

Finally we stop when all ancestral alleles have mutated into derived alleles, so that the genotype abc has then reached fixation.

Due to the large population sizes being considered, one can consider as biologically significant fixation events, the situations where a specific genotype reaches a moderately high threshold frequency, such as 70%. Indeed in large populations, multiple genotypes can coexist side by side for a fairly long time once both genotypes have reached sufficiently high frequencies. We present the results for fixation threshold of 90% . However, from a biological point of view a fixation threshold of 70% would already be quite meaningful, and our computations and concepts could also be applied to these moderate fixation thresholds.

In the CI model [94], back or reverse mutations are not allowed, mutations are only allowed from ancestral alleles to derived alleles. So a derived allele is not allowed to mutate into any other allele, for instance, $a \rightarrow bAbc \rightarrow ABc$ is not allowed. So, this gives us a sparser mutation matrix leading to some changes in the formulas derived for rate functional minimizing trajectories.

Without loss of generality, at every stage, set genotype 1 to be the genotype with the most ancestral alleles and the other genotypes are arranged in the increasing order of their relative fitnesses.

In the model we introduced before in chapter 2, the growth factor was given by

$$F_g = F_{anc}^{1+S_g}$$

where F_{anc} is the growth factor of an ancestor and S_g is the selective advantage of genotype g over the ancestor.

In our numerical computations below, the values of growth factors and mutation parameter are the experimental values estimated for the *E. coli* experiments of Tim

Cooper [129]. These values are tabulated in Table 10.2.

In the evolution model we consider here, we assume as before that mutations emerge at random, according to Poisson distributions, but that regressive mutations never occur.

Table 10.2 presents the values of selection coefficient and mutation rates used in our computations.

Genotype Index	Genotype	GrowthFactor = $F^{1+\text{SelectionCoefficient}}$	Mutation Rate
1	ABC	200	0.5×10^{-6}
2	ABc	$200^{1.06}$	0.5×10^{-6}
3	AbC	$200^{1.1}$	0.5×10^{-6}
4	aBC	$200^{1.15}$	0.5×10^{-6}

Table 10.2: Parameters from TC [129] Experiment used in CI Model.

In Stage I, as explained above, we start from a pure population of genotype ABC, and study what are the most likely process trajectories leading to the fixation of a single derived allele.

For Stage I, we have ABC as ancestor and 3 mutational trajectories to aBC, AbC, and ABc giving us a total of 4 genotypes. The 4×4 mutations matrix, r will have $r_{i,j} = 0$, for all $i = 2, 3, 4$ with $j = 1, ..., 4$ and $r_{1,1} = 0$. So the mutation matrix has non zero entries in the first row only.

We have labeled ABC, ABc, AbC, and aBC as genotype 1, 2, 3, and 4 respectively in the increasing order of their growth factors, which is also the increasing order of their fitnesses.

Similarly for Stage II, we have 3 possible ancestors: ABc, Abc, and aBc. So we subdivide this stage into 3 substages, each starting from different ancestor.

For instance, Stage II.1 starts from aBC and has mutational trajectories to abC and aBc. The mutations matrix, r will now have $r_{i,j} = 0$, for all $i = 2, 3$ with $j = 1, 2, 3$ and $r_{1,1} = 0$.

Stage II.2 starts from AbC and has mutational trajectories to abC and Abc. The mutations matrix, r will now have $r_{i,j} = 0$, for all $i = 2, 3$ with $j = 1, 2, 3$ and $r_{1,1} = 0$.

Similarly, Stage II.3 starts from ABc and has mutational trajectories to aBc and Abc. The mutations matrix, r will now have $r_{i,j} = 0$, for all $i = 2, 3$ with $j = 1, 2, 3$ and $r_{1,1} = 0$.

Again we observe that the mutation matrix has non zero entries in the first row only. Now we implement the techniques derived in chapter 6 for building reverse optimal rate minimizing trajectories.

10.3 Large Deviations Application: Most Likely Mutational Trajectory

We derive the formulas like in chapter 6 before for the trajectory $x \to y \to z$. Let $\bar{r}, \bar{p}$ and r, p be the respective random matrix of restricted mutations and the intermediary population histogram before dilution for x and y respectively. Also, let $\vec{m} = (m_i)$ be the vector of mutation rates where $m_i = m$, $\forall i$, i.e., we consider that the rates of mutation are equal regardless of the species and the number of genotypes in system be g .

For now, we use the rate functional approximation as derived before using zero order approximation for r, ρ, p. There is one main difference in the expression for cost or rate functional in this case. Since the mutation matrix has non-zero entries in the first row only, we have

$$\begin{aligned}\mathrm{RF}(x,y) = {} & m\sum_{j=1}^{g}(g-j)F_jx_j\rangle + m\sum_{k=2}^{g}F_1x_1\left(\frac{y_k}{F_kx_k} - \frac{y_1}{F_1x_1} - 1\right)\exp\left(\frac{y_k}{F_kx_k} - \frac{y_1}{F_1x_1}\right)\\ & + \sum_j y_j \log\left(\frac{y_j\langle F, x\rangle}{F_jx_j}\right).\end{aligned}$$

and

$$\begin{aligned}\mathrm{RF}(y,z) = {} & m\sum_{j=1}^{g}(g-j)F_jy_j + m\sum_{k=2}^{g}F_1y_1\left(\frac{z_k}{F_ky_k} - \frac{z_1}{F_1y_1} - 1\right)\exp\left(\frac{z_k}{F_ky_k} - \frac{z_1}{F_1y_1}\right)\\ & + \sum_j z_j \log\left(\frac{z_j\langle F, y\rangle}{F_jy_j}\right).\end{aligned}$$

Thus,

$$\frac{\partial \mathrm{RF}(x,y)}{\partial y_1} = 1 + \log\left(\frac{y_1}{F_1 x_1}\right) - m \sum_{k=2}^{g} \left(\frac{y_k}{F_k x_k} - \frac{y_1}{F_1 x_1}\right) \exp\left(\frac{y_k}{F_k x_k} - \frac{y_1}{F_1 x_1}\right)$$

and

$$\frac{\partial \mathrm{RF}(x,y)}{\partial y_i} = 1 + \log\left(\frac{y_i}{F_i x_i}\right) + m \frac{F_1 x_1}{F_i x_i} \left(\frac{y_i}{F_i x_i} - \frac{y_1}{F_1 x_1}\right) \exp\left(\frac{y_i}{F_i x_i} - \frac{y_1}{F_1 x_1}\right)$$

for $i = 1, ..., g$.

Thus,

$$\frac{\partial \mathrm{RF}(x,y)}{\partial y_1} = 1 + \log\left(\frac{y_1}{F_1 x_1}\right) + m \sum_{k=2}^{g} a_{1,k} \exp(a_{k,1})$$

and

$$\frac{\partial \mathrm{RF}(x,y)}{\partial y_i} = 1 + \log\left(\frac{y_i}{F_i x_i}\right) + m \frac{F_1 x_1}{F_i x_i} a_{i,1} \exp(a_{i,1})$$

for $i = 1, ..., g$, where $a_{i,j} = \frac{y_i}{F_i x_i} - \frac{y_j}{F_j x_j}$.

Similarly,

$$\frac{\partial \mathrm{RF}(y,z)}{\partial y_1} = (g-1) m F_1 + \frac{F_1}{\langle F, y \rangle} - \frac{z_1}{y_1} \tag{10.1}$$

$$+ m \sum_{k=2}^{g} \left(\frac{F_1 z_k}{F_k y_k} - \frac{z_1}{y_1} - F_1 + \frac{z_1 z_k}{F_k y_k y_1} - \frac{z_1^2}{F_1 y_1^2}\right) \exp(b_{k,1})$$

and

$$\frac{\partial \mathrm{RF}(y,z)}{\partial y_i} = (g-i) m F_i + \frac{F_i}{\langle F, y \rangle} - \frac{z_i}{y_i} - m \frac{F_1 y_1 z_i}{F_i y_1^2} b_{i,1} \exp(b_{i,1})$$

for $i = 1, ..., g$, where $b_{i,j} = \frac{z_i}{F_i y_i} - \frac{z_j}{F_j y_j}$.

Now, using Lagrange optimality conditions we have

$$\lambda = \frac{\partial \mathrm{RF}(x,y)}{\partial y_i} + \frac{\partial \mathrm{RF}(y,z)}{\partial y_i},\ \forall i$$

$$\begin{aligned}\lambda = 1 + \log\left(\frac{y_1}{F_1 x_1}\right) + (g-1)mF_1 + \frac{F_1}{\langle F, y\rangle} - \frac{z_1}{y_1} + m\sum_{k=2}^{g} a_{1,k}\exp(a_{k,1})\\ + m\sum_{k=2}^{g}\left(\frac{F_1 z_k}{F_k y_k} - \frac{z_1}{y_1} - F_1 + \frac{z_1 z_k}{F_k y_k y_1} - \frac{z_1^2}{F_1 y_1^2}\right)\exp(b_{k,1})\end{aligned}$$

and

$$\begin{aligned}\lambda = 1 + \log\left(\frac{y_i}{F_i x_i}\right) + (g-i)mF_i + \frac{F_i}{\langle F, y\rangle} - \frac{z_i}{y_i}\\ + m\frac{F_1 x_1}{F_i x_i} a_{i,1}\exp(a_{i,1}) - m\frac{F_1 y_1 z_i}{F_i y_1^2} b_{i,1}\exp(b_{i,1})\end{aligned}$$

The above equations give an implicit system in g variables x_i along with the condition that $\sum_i x_i = 1$.

Again we assume our histograms to be interior points, and with Poisson mutations we have following system of implicit equations in x given y, z

$$\begin{aligned}\lambda = 1 + \log\left(\frac{y_1}{F_1 x_1}\right) + (g-1)mF_1 + \frac{F_1}{\langle F, y\rangle} - \frac{z_1}{y_1} + m\sum_{k=2}^{g} a_{1,k}\exp(a_{k,1})\\ + m\sum_{k=2}^{g}\left(\frac{F_1 z_k}{F_k y_k} - \frac{z_1}{y_1} - F_1 + \frac{z_1 z_k}{F_k y_k y_1} - \frac{z_1^2}{F_1 y_1^2}\right)\exp(b_{k,1})\end{aligned}$$

and

$$\begin{aligned}\lambda = 1 + \log\left(\frac{y_i}{F_i x_i}\right) + (g-i)mF_i + \frac{F_i}{\langle F, y\rangle} - \frac{z_i}{y_i}\\ + m\frac{F_1 x_1}{F_i x_i} a_{i,1}\exp(a_{i,1}) - m\frac{F_1 y_1 z_i}{F_i y_1^2} b_{i,1}\exp(b_{i,1})\end{aligned}$$

Proceeding exactly like in chapter 6, we solve the system for value of x as follows. Thus

$$X_i = -x_i(0)\sum_{j\neq i} x_j(0)(V_j(x(0),y)+\tilde{V}_j(y,z)) + x_i(0)(1-x_i(0))(V_i(x(0),y)+\tilde{V}_i(y,z)).$$

where $x(0)$ is given by equation 6.9 and

$$V_1(x,y) = \sum_{k=2}^{g} a_{1,k}\exp(-a_{1,k}), \tag{10.2}$$

$$V_i(x,y) = \frac{F_1 x_1}{F_i x_i} a_{i,1}\exp(a_{i,1}),\ i = 2,..,g \tag{10.3}$$

$$\tilde{V}_1(y,z) = (g-1)F_i + \sum_{k=2}^{g}\left(\frac{F_1 z_k}{F_k y_k} - \frac{z_1}{y_1} - F_1 + \frac{z_1 z_k}{F_k y_k y_1} - \frac{z_1^2}{F_1 y_1^2}\right)\exp(b_{k,1})$$

$$\tilde{V}_i(y,z) = (g-i)F_i - \frac{F_1 y_1 z_i}{F_i y_1^2} b_{i,1}\exp(b_{i,1})\ i = 2,...,g. \tag{10.4}$$

Here

$$a_{i,j} = \frac{y_i}{F_i x_i} - \frac{y_j}{F_j x_j}$$

and

$$b_{i,j} = \frac{z_i}{F_i y_i} - \frac{z_j}{F_j y_j}$$

So we have

$$x_i(1) = x_i(0) + mX_i$$

as the solution.

10.4 Most Likely Evolution Trajectories

As explained in previous section, we build the rate optimizing trajectory in stages.

There are many possible process trajectories that start at ABC and for which $g_3 = AbC$ reaches near fixation before all other genes g_1, g_2, or g_4 fixate. For large N the probability of such an event Γ is of the order of

$$\exp(-N\Lambda(\Gamma))$$

where

$$\Lambda(\Gamma) = \min_{traj \in \Gamma} RF(traj).$$

The cost of a trajectory is given by the value of associated optimal rate functional(RF) as explained in detail previously in chapters 4 and 5.

We first present results for generating optimal large deviation trajectory for the case of $g = 4$ genotypes, where

Genotype Index	Associated Genotype
1	ABC
2	ABc
3	AbC
4	aBC

Table 10.3: Genotypes in the Mutational Trajectories at Stage I

We start from an almost pure population in the initial state with genotype frequency $\geq 99\%$ and for terminal histogram we assume fixation threshold frequency to be $\geq 90\%$.

1. At Stage I, among all trajectories starting from ABC (genotype 1) and reaching fixation at aBC (genotype 4), the rate minimizing trajectory is the mean trajectory since the mean trajectory has cost 0 and ends up at genotype 4. Hence this event Γ has cost $\Lambda(\Gamma) = 0$, and its probability will become close to 1 when N is large as shown in chapter 4. The mean trajectory is shown in Figure 10.2.

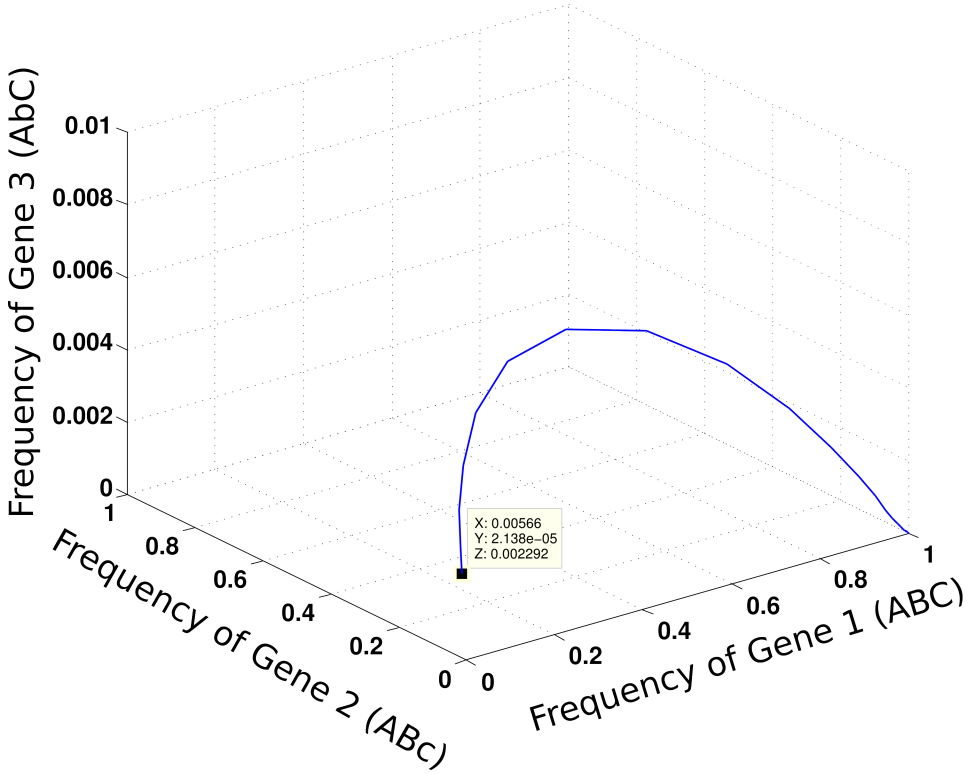

Figure 10.2: CI Model-Stage 1: Mean Trajectory is the Rate Minimizing Trajectory from ABC (Genotype 1) to aBC (Genotype 4).

2. At Stage I, the rate minimizing trajectory from *ABC* (genotype 1) to *AbC* (genotype 3) is presented in the Figure 10.3. The rate minimizing trajectory has 14 steps and an associated cost of 0.018.

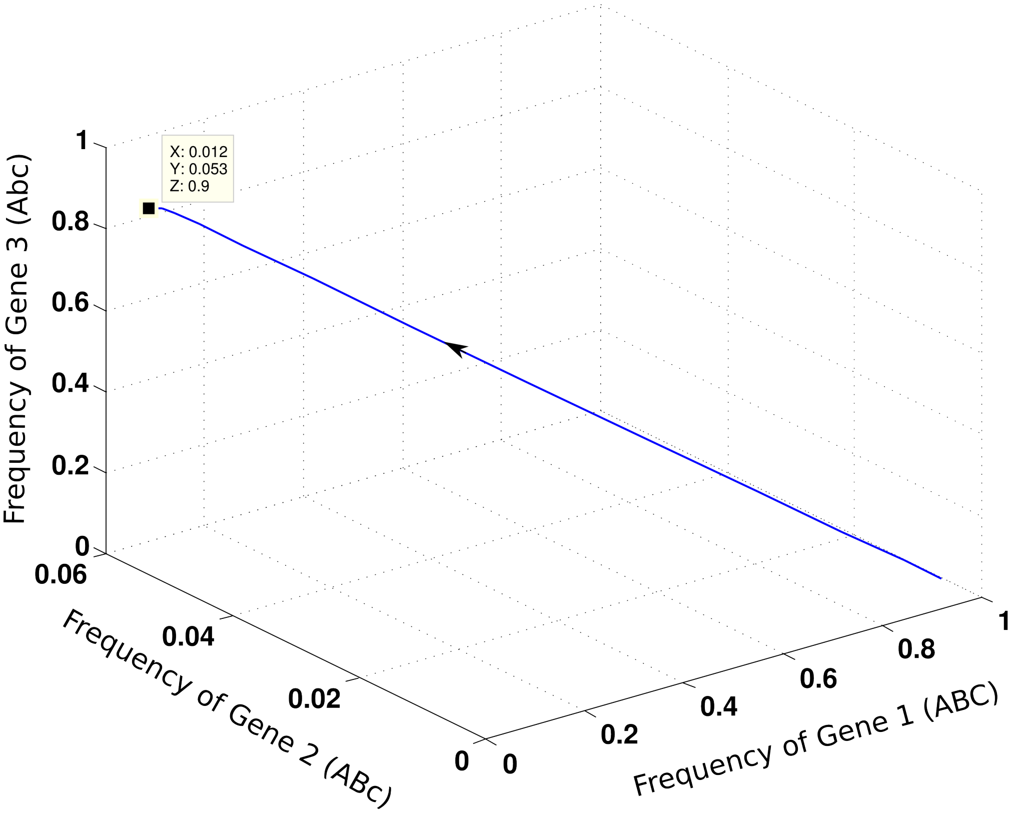

Figure 10.3: CI Model-Stage 1: Rate Minimizing Trajectory from *ABC* (Genotype 1) to *AbC* (Genotype 3).

3. At Stage I, the rate minimizing trajectory for the rare event from ABC (genotype 1) to ABc (genotype 2) is presented in the Figure 10.4. The rate minimizing trajectory has 19 steps and an associated cost of 0.65.

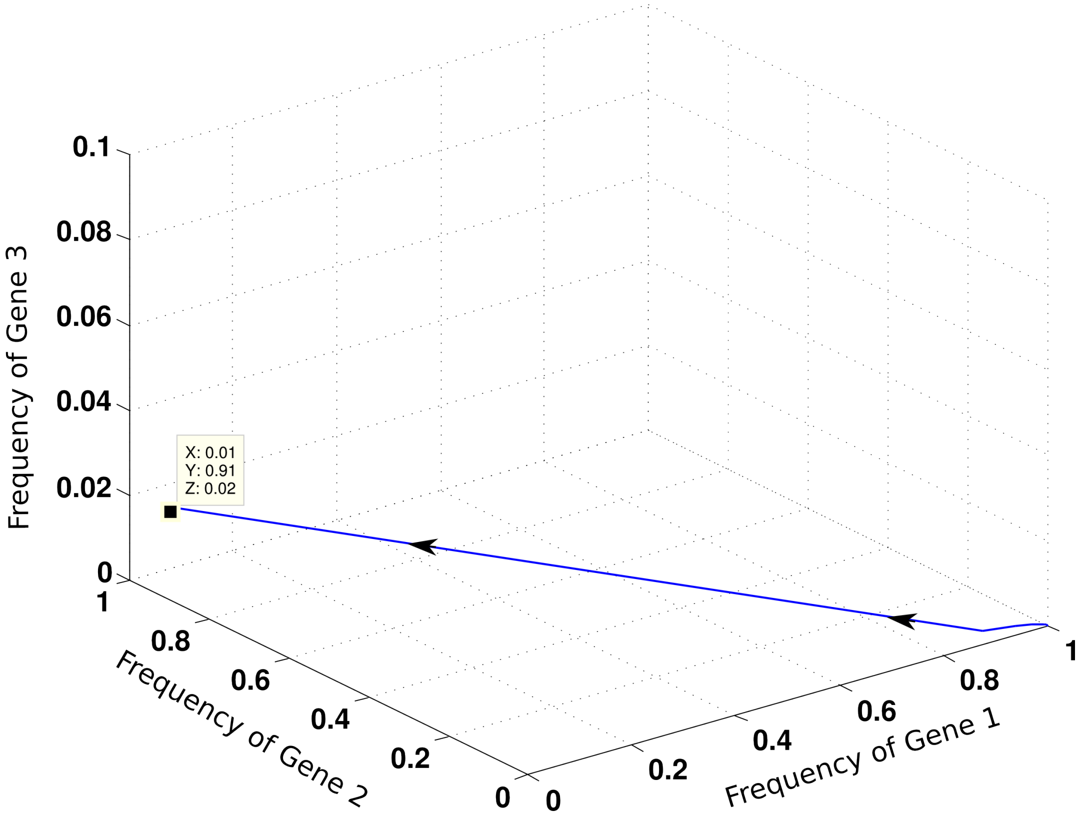

Figure 10.4: CI Model-Stage 1: Rate Minimizing Trajectory from ABC (Genotype 1) to ABc (Genotype 2).

At Stage II we build optimal evolutionary trajectories for the second stage in which a population with one derived allele evolves into a population with two derived alleles. We subdivide Stage II into 3 stages depending on the one derived allele already present in the intermediate population reached at the end of Stage I. Stage II.1 has mutational trajectories involving the genotypes listed in Table 10.4.

Genotype Index	Associated Genotype
1	aBC
2	aBc
3	abC

Table 10.4: Genotypes Involved in Stage II.1

1. At Stage II.1, the rate minimizing trajectory from genotype aBC (Genotype 1) to abC (Genotype 3) is the mean trajectory starting from aBC, which has zero cost. The mean trajectory with 37 steps is presented in Figure 10.5.

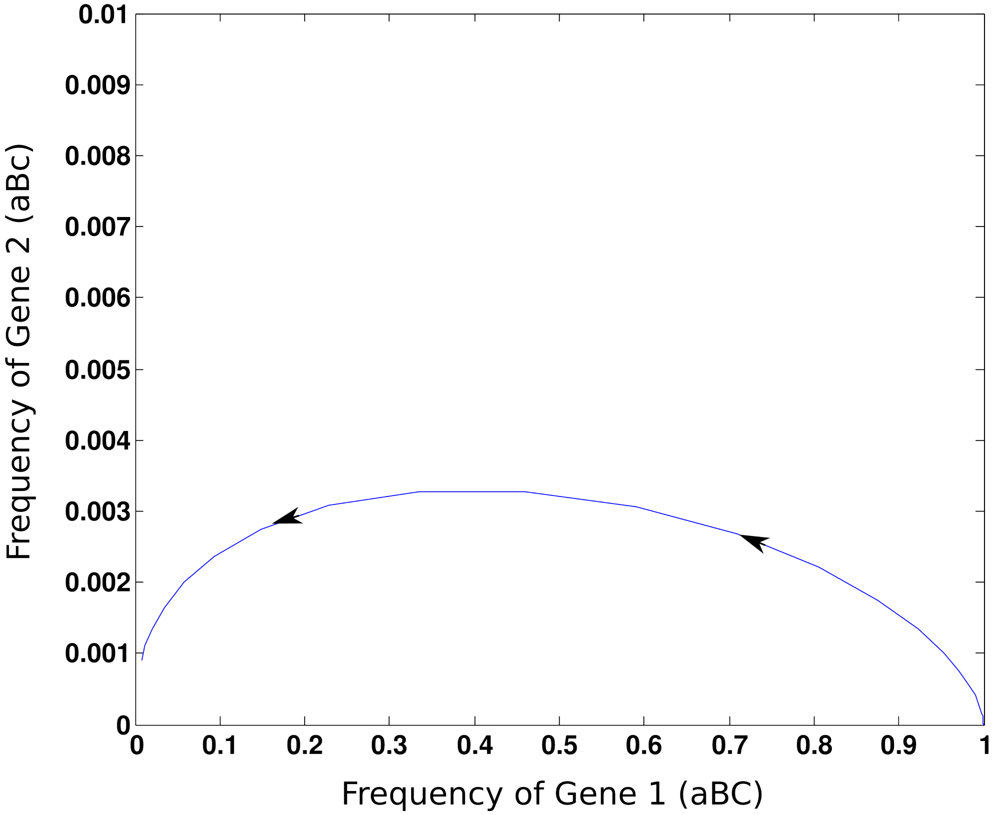

Figure 10.5: CI Model-Stage II.1: Mean Trajectory as the Rate Minimizing Trajectory from aBC (Genotype 1) to abC (Genotype 3).

2. At Stage II.1, the rate minimizing trajectory from genotype aBC (Genotype 1) to aBc (Genotype 2) is shown in Figure 10.6. It has 21 steps and an associated cost of 0.037.

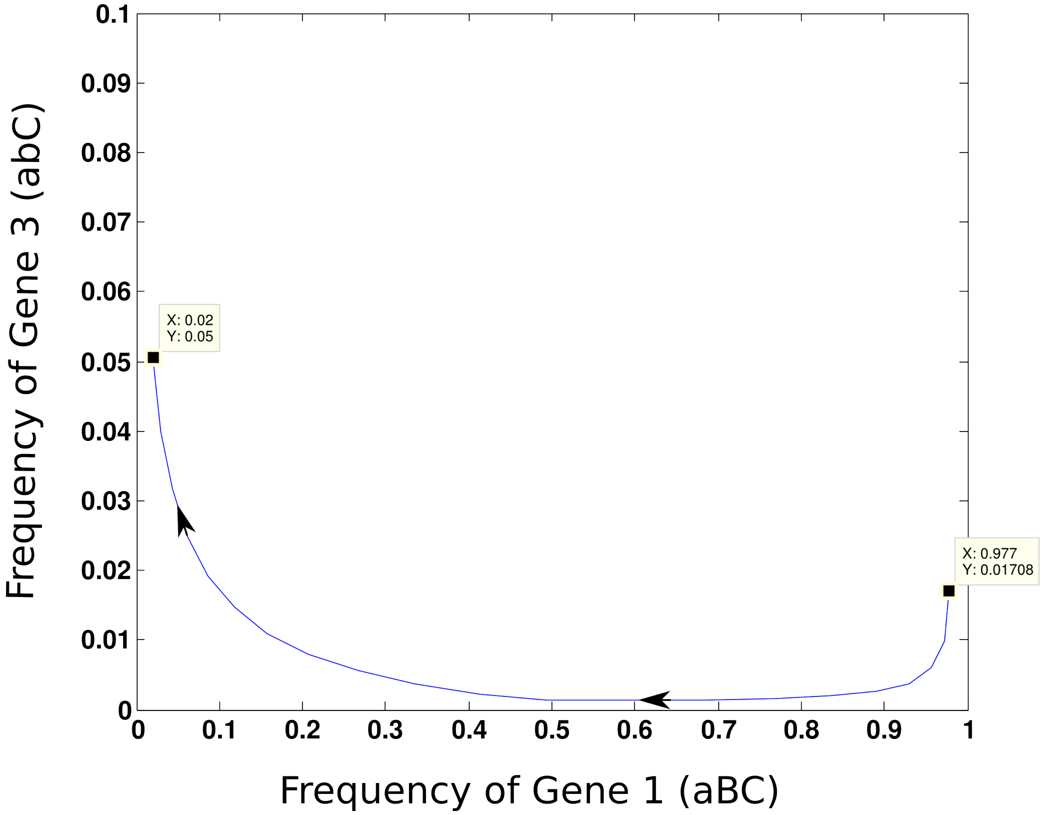

Figure 10.6: CI Model-Stage II.1: Rate Minimizing Trajectory from aBc (Genotype 1) to aBc (Genotype 2).

Stage II.2 has mutational trajectories involving the genotypes listed in Table 10.5

Genotype Index	Associated Genotype
1	AbC
2	Abc
3	abC

Table 10.5: Genotypes Involved in Stage II.2

3. At Stage II.2, the rate minimizing trajectory from genotype *AbC* (Genotype 1) to *abC* (Genotype 3) is the mean trajectory starting from *AbC*, which has zero cost. It is shown in Figure 10.7 and has 26 steps.

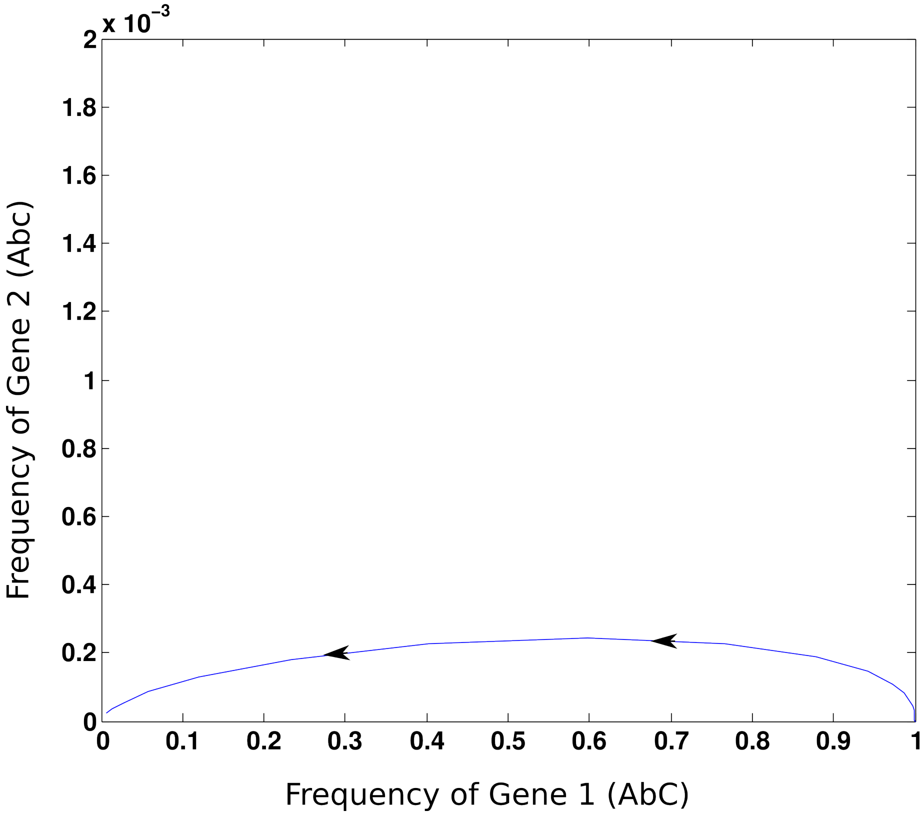

Figure 10.7: CI Model-Stage II.2: Mean Trajectory is the Rate Minimizing Trajectory from *AbC* (Genotype 1) to *abC* (Genotype 3).

4. At Stage II.2, the rate minimizing trajectory from genotype *AbC* (Genotype 1) to *Abc* (Genotype 2) is shown in Figure 10.8. It has 18 steps and an associated cost of 0.19.

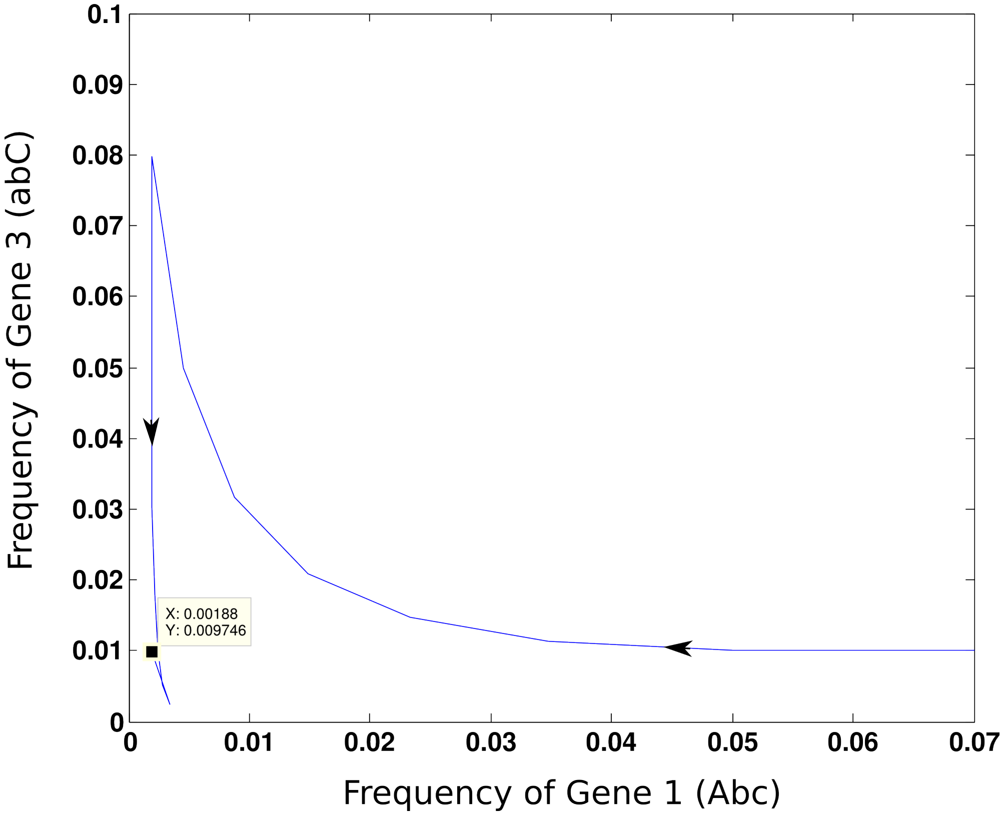

Figure 10.8: CI Model-Stage II.2: Rate Minimizing Trajectory from *Abc* (Genotype 1) to *Abc* (Genotype 2).

Stage II.3 has mutational trajectories involving the genotypes listed in Table 10.6

Genotype Index	Associated Genotype
1	ABc
2	Abc
3	aBC

Table 10.6: Genotypes Involved in Stage II.3

5. At Stage II.3, the rate minimizing trajectory from genotype ABc (Genotype 1) to aBc (Genotype 3) is again the mean trajectory starting at ABc and has zero cost. Mean trajectory with 26 steps is presented in the Figure 10.9.

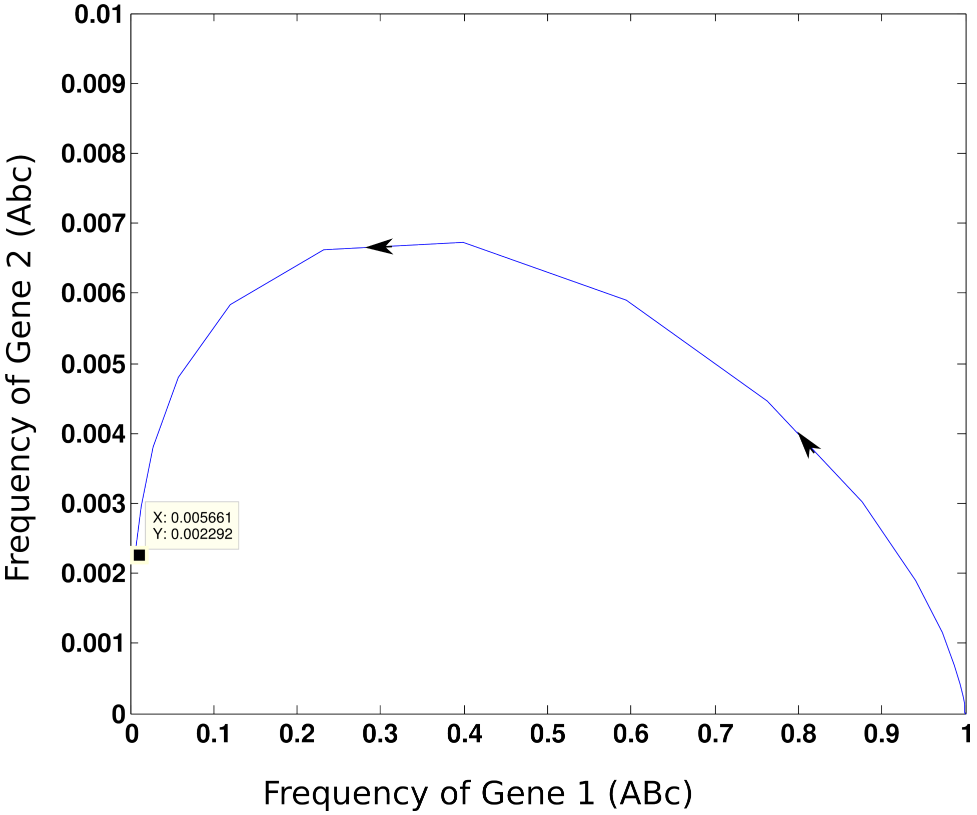

Figure 10.9: CI Model-Stage II.3: Mean Trajectory is the Rate Minimizing Trajectory from ABc (Genotype 1) to aBc (Genotype 3).

6. At Stage II.3, the rate minimizing trajectory from genotype *ABc* (Genotype 1) to *Abc* (Genotype 2) is presented in the Figure 10.10. It has 15 steps and an associated cost of 0.029.

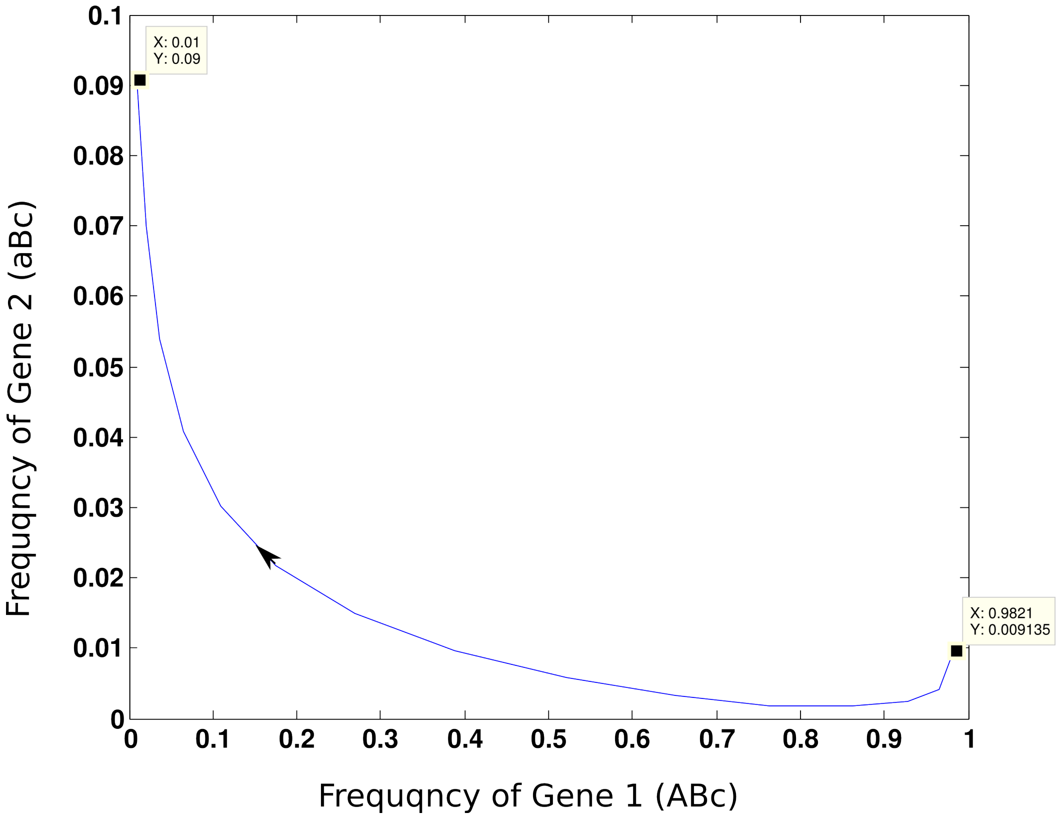

Figure 10.10: CI Model-Stage II.3: Rate Minimizing Trajectory from *ABc* (Genotype 1) to *Abc* (Genotype 2).

So the network of all mutational trajectories with corresponding costs is shown in Figure 10.11

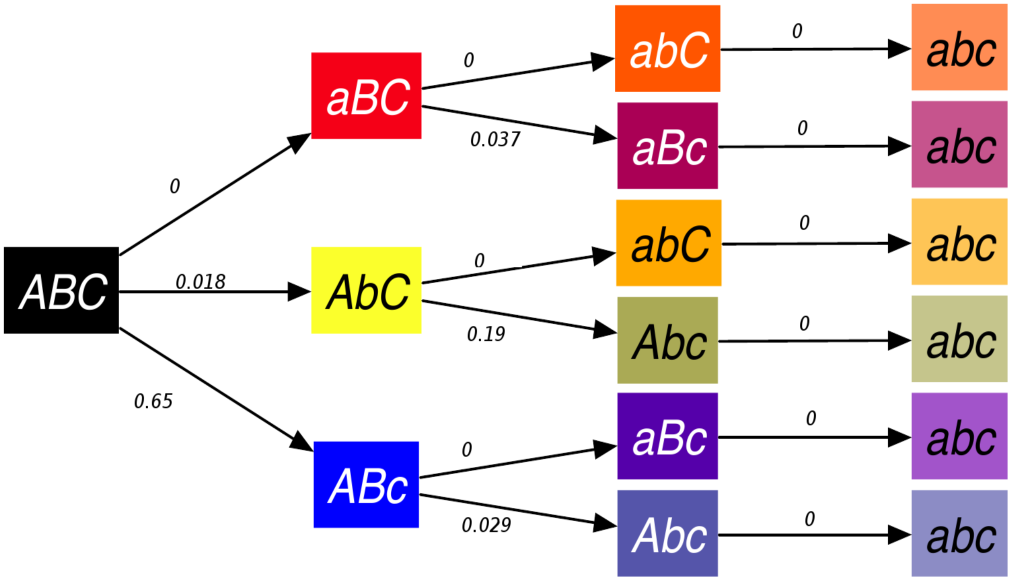

Figure 10.11: CI Model: All Mutational Trajectories with the Rate Functional Values Associated to each Transition.

10.5 Complete Costs for Mutational Trajectories

As explained in chapters 3 and 4 the complete cost of a three-stage mutation trajectory will be the sum of the individual costs of all the three trajectories involved.

Since rate functional (cost) for each part of trajectory is optimal, the sum of these optimal rate functionals gives an optimal rate functional for the complete mutational trajectory.

We now rank the mutational trajectories in order of their increasing associated cost in Table 10.7

Number of Trajectory	Mutational Trajectory	Associated Cost
1	$ABC \rightarrow aBC \rightarrow abC \rightarrow abc$	0
2	$ABC \rightarrow AbC \rightarrow abC \rightarrow abc$	0.018
3	$ABC \rightarrow aBC \rightarrow aBc \rightarrow abc$	0.037
4	$ABC \rightarrow AbC \rightarrow Abc \rightarrow abc$	0.2
5	$ABC \rightarrow ABc \rightarrow aBc \rightarrow abc$	0.65
6	$ABC \rightarrow ABc \rightarrow Abc \rightarrow abc$	0.67

Table 10.7: Associated Costs for the Complete Mutational Trajectories.

The repeatability is mainly the fact that for larger values of population sizes, N the probabilities of mutational trajectories become more and more like $\exp(-N(\text{cost}))$.

10.6 Direct Simulations

Recall that, fTH denotes the threshold fixation frequency of target genotype in the terminal state of the evolutionary trajectories. Also, N is the population size and we present results for 3 population sizes, $N = 5 \times 10^4$, 2×10^5 and 5×10^5.

For Stage I, Table 10.8 shows the distribution of genotypes in terminal state of all the trajectories where $fTH = 0.90$.

Fixation Genotype	$N = 5 \times 10^4$	$N = 2 \times 10^5$	$N = 5 \times 10^5$
1. ABC	0	0	0
2. ABc	0.07	0.003	0
3. AbC	0.3	0.11	0.01
4. aBC	0.63	0.89	0.99

Table 10.8: Probability of Reaching Fixation (Stage I) for all Genotypes in Simulated Trajectories with 4 Genotypes and $fTh = 90\%$ as the Fixation Threshold.

For Stage II.1, Table 10.9 shows the distribution of genotypes in terminal state of all the trajectories where $fTH = 0.90$.

Fixation Genotype	$N = 5 \times 10^4$	$N = 2 \times 10^5$	$N = 5 \times 10^5$
1. aBC	0	0	0
2. aBc	0.26	0.06	0.004
3. abC	0.74	0.94	0.996

Table 10.9: Probability of Reaching Fixation (Stage II.1) for all Genotypes in Simulated Trajectories with 3 Genotypes and $fTh = 90\%$ as the Fixation Threshold.

For Stage II.2, Table 10.10 shows the distribution of genotypes in terminal state of all the trajectories where $fTH = 0.90$.

Fixation Genotype	$N = 5 \times 10^4$	$N = 2 \times 10^5$	$N = 5 \times 10^5$
1. AbC	0	0	0
2. Abc	0.2	0.02	10^{-4}
3. abC	0.8	0.98	0.9999

Table 10.10: Probability of Reaching Fixation (Stage II.2) for all Genotypes in Simulated Trajectories with 3 Genotypes and $fTh = 90\%$ as the Fixation Threshold.

Finally for Stage II.3, Table 10.11 shows the distribution of genotypes in terminal state of all the trajectories where $fTH = 0.90$.

Fixation Genotype	$N = 5 \times 10^4$	$N = 2 \times 10^5$	$N = 5 \times 10^5$
1. ABc	0	0	0
2. Abc	0.33	0.12	0.02
3. aBc	0.67	0.88	0.98

Table 10.11: Probability of Reaching Fixation (Stage II.3) for all Genotypes in Simulated Trajectories with 3 Genotypes and $fTh = 90\%$ as the Fixation Threshold.

The following Figure 10.12 shows probabilities for all the mutational trajectories.

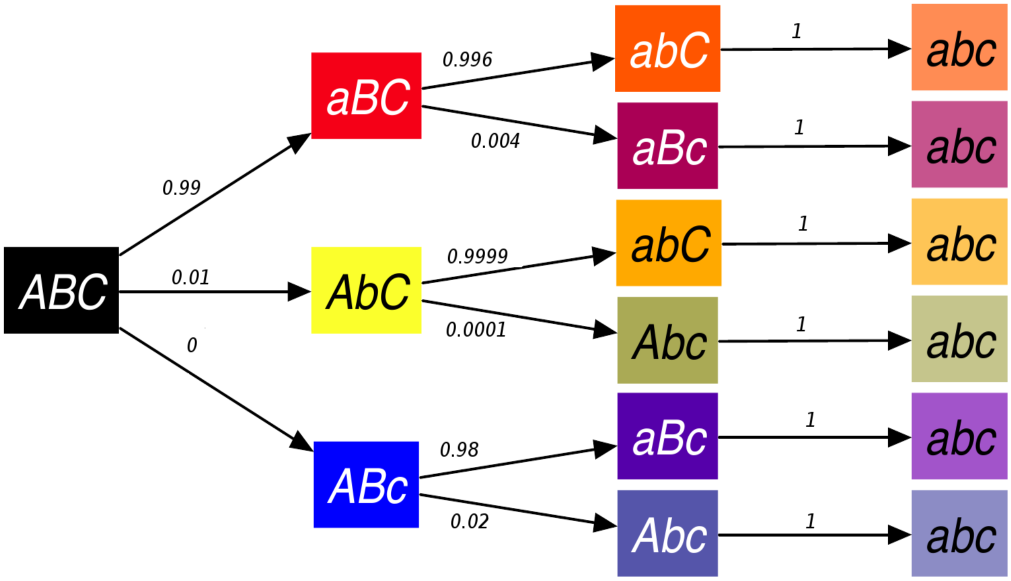

Figure 10.12: CI Model: All Mutational Trajectories with Simulated Transition Probabilities.

Table 10.12 shows the empirical probabilities estimated using direct simulations with $N = 5 \times 10^5$ for all 6 mutational trajectories.

Number of Trajectory	Mutational Trajectory	Probability
1	$ABC \rightarrow aBC \rightarrow abC \rightarrow abc$	0.986
2	$ABC \rightarrow AbC \rightarrow abC \rightarrow abc$	0.01
3	$ABC \rightarrow aBC \rightarrow aBc \rightarrow abc$	0.004
4	$ABC \rightarrow AbC \rightarrow Abc \rightarrow abc$	10^{-6}
5	$ABC \rightarrow ABc \rightarrow aBc \rightarrow abc$	0
6	$ABC \rightarrow ABc \rightarrow Abc \rightarrow abc$	0

Table 10.12: Empirical Probabilities Simulated for the Full Mutational Trajectories.

Note that the ordering of these 6 trajectories by decreasing likelihood is the same for the empirical probabilities $P(traj)$ evaluated by simulations in Table 10.12 and for the minimal costs associated to these 6 mutational trajectories given in the Table 10.7.

CHAPTER 11

Conclusions and Future Work

We have shown that large deviation rate minimizing trajectories provide a computationally efficient way of predicting rare events in the evolution models for bacteria *Eschericia coli*. We have implemented shooting algorithms to efficiently compute reverse large deviation optimal trajectories instead of brute force optimization.

It is often impossible to visualize rare event trajectories during direct simulation and also under laboratory experimental conditions. To that effect we can not only predict the paths for rare trajectories but also obtain rough estimates of their probabilities.

We would like to implement importance sampling to obtain accurate simulations of rare events by forcing the simulated random evolutions to follow the most likely trajectories obtained by minimization of large deviation rate functionals.

We also extended our model to the set up in paper on clonal interference and are able to illustrate similar results using different growth factor and mutation restrictions.

We have shown that large deviation approaches provide new tools, implementable numerically, to study the genetic evolution of large populations of bacteria or viruses, and specifically when one focuses on comparing rare evolutionary events which may never emerge in direct simulations. In future work, we intend to collaborate with experimental biologists to apply these techniques to genetic evolution data for other microbial and bacterial populations.

Bibliography

[1] J.G. Arjan, M. De Visser, C.W. Zeyl, P.J. Gerrish, J.L. Blanchard, and R.E. Lenski. Diminishing returns from mutation supply rate in asexual populations. *Science*, 283(5400):404–406, 1999.

[2] K.C. Atwood, L.K. Schneider, and F.J. Ryan. Periodic selection in *E. coli.*. *Genetics*, 37:146–155, 1951.

[3] R. Azencott. Grandes deviations et applications. *Lecture notes in Mathematics*, 774:1–176, 1980.

[4] Robert Azencott, Mark I. Freidlin, and Srinivasa R.S Varadhan. *Large Deviations at Saint-Flour*. Springer-Verlag New York, LLC, 2012.

[5] R. R. Bahadur. *Some Limit Theorems in Statistics.* SIAM : Society for Industrial and Applied Mathematics, 1971.

[6] R.R. Bahadur and S.L. Zabell. Large deviations of the sample mean in general vector spaces. *Annals of Probability*, 7(4):587–621, 1979.

[7] J. M. Bahi and C. J. Michel. A stochastic gene evolution model with time dependent mutations. *Bulletin of Mathematical Biology*, 66:763–778, 2004.

[8] J.E. Barrick and Lenski R.E. Genome-wide mutational diversity in an evolving population of *Escherichia coli*. *Cold Springs Harbor Symposia on Quantitative Biology*, 74:119–129, 2009.

[9] J.E. Barrick, C.C. Strelioff, R.E. Lenski, and M.R. Kauth. *Escherichia coli* rpob mutants have increased evolvability in proportion to their fitness defects. *Molecular Biology and Evolution*, 27(6):1338–1347, 2010.

[10] Jose Blanchet and Henry Lam. State-dependent importance sampling for rare-event simulation: An overview and recent advances. *Surveys in Operations Research and Management Science*, 17(1):38–59, 2012.

[11] Zachary D. Blount, Christina Z. Borland, and Richard E. Lenski. Historical contingency and the evolution of a key innovation in an experimental population of *Escherichia coli*. *Proceedings of the National Academy of Sciences USA*, 105(23):7899–7906, 2008.

[12] Paul C. Bressloff. Stochastic neural field theory and the system-size expansion. *SIAM Journal on Applied Mathematics*, 70(5):1488–1521, 2009.

[13] James A. Bucklew. *Large Deviation Techniques in Decision, Simulation, and Estimation*. John Wiley and Sons, 1990.

[14] James A. Bucklew. *Introduction to Rare Event Simulation*. Springer, 2004.

[15] Amarjit Budhiraja and Arka Prasanna Ghosh. A large deviations approach to asymptotically optimal control of crisscross network in heavy traffic. *The Annals of Applied Probability*, 15(3):1887–1935, 2005.

[16] Paulo R. A. Campos and Lindi M. Wahl. The effects of population bottlenecks on clonal interference, and the adaptation effective population size. *Evolution*, 63(4):950–958, 2009.

[17] Paulo R. A. Campos and Lindi M. Wahl. The adaptation rate of asexuals: Deleterious mutations, clonal interference and population bottlenecks. *Evolution*, 64(7):1973–83, 2010.

[18] Alain Cercueil and Olivier Franq̧is. Sharp asymptotics for fixation times in stochastic population dynamics with low mutation probabilities. *Journal of Statistical Physics*, 110:311–332, 2003.

[19] Nicolas Champagnat. A microscopic interpretation for adaptive dynamics trait substitution sequence models. *Stochastic Processes and their Applications*, 116(8):1127–1160, 2006.

[20] Nicolas Champagnat, Pierre-Emmanuel Jabin, and Sylvie Méléard. Adaptation in a stochastic multi-resources chemostat model. *ArXiv e-prints :1302.0552*, 2013.

[21] Nicolas Champagnat and Sylvie Méléard. Polymorphic evolution sequence and evolutionary branching. *Centre De Mathématiques Appliquées*, 244(1):112–120, 2008.

[22] Nicolas Champagnat, Sylvie Méléard, and Régis Ferriére. Unifying evolutionary dynamics: from individual stochastic processes to macroscopic models. *Theoretical Population Biology.*, 69(3):297–321, 2006.

[23] H. Chernoff. Asymptotic efficiency for tests based on the sum of observations. *Annals of the Institute of Statistical Mathematics*, 23:493–507, 1952.

[24] V.S. Cooper and R.E. Lenski. The population genetics of ecological specialization in evolving *E. coli* populations. *Nature*, 407:736–739, 2000.

[25] V.S. Cooper, D. Schneider, M. Blot, and R.E. Lenski. Mechanisms causing rapid and parallel losses of ribose catabolism in evolving populations of *Escherichia coli*. *Journal of Bacteriology*, 9(183):2834–2841, 2001.

[26] M.C. Cowperthwaite and L.A.Meyers. How mutational networks shape evolution: Lessons from rna models. *Annual Review of Ecology, Evolution and Systematics*, 38:203–230, 2007.

[27] H. Cramér. Sur un nouveau théorème-limite de la théorie des probabilités. *Colloquium on theory of probability.*, 736:5–23, 1937.

[28] C. Darwin. *The Origin of Species by Means of Natural Selection.* John Murray, 1859.

[29] Amir Dembo and Ofer Zeitouni. *Large Deviations Techniques and Applications.* Springer, 1998.

[30] Lloyd Demetrius, Volker Matthias Gundlach, and Gunter Ochs. Complexity and demographic stability in population models. *Theoretical Population Biology*, 65:211–225, 2004.

[31] Frank den Hollander. *Large Deviations.* AMS : American Mathematical Society, 2000.

[32] M.M. Desai and D.S. Fisher. Beneficial mutation selection balance and the effect of linkage on positive selection. *Genetics*, 176:1759–1798, 2007.

[33] M.M. Desai, D.S. Fisher, and A.W.Murray. The speed of evolution and maintenance of variation in asexual populations. *Current Biology*, 17:385–394, 2007.

[34] U. Dieckmann and R. Law. The dynamical theory of coevolution: A derivation from stochastic ecological processes. *Journal of Mathematical Biology*, 34:579–612, 1996.

[35] U. Dieckmann and R. Law. Relaxation projections and the method of moments. *In: Dieckmann, U., Law, R., Metz, J.A.J. (Eds.), The Geometry of Ecological Interactions: Simplifying Spatial Complexity. Cambridge University Press, Cambridge,*, pages 412–455, 2000.

[36] A. B. Dieker and M. Mandjes. On asymptotically efficient simulation of large deviation probabilities. *2000 Mathematics Subject Classiffication: primary 65C05; secondary 60F10, 60K10.*, 2000.

[37] Kvitek D.J. and Sherlock G. Reciprocal sign epistasis between frequently experimentally evolved adaptive mutations causes a rugged fitness landscape. *PLoS Genetics*, 7(4):e1002056, 2011.

[38] M. Donsker and S. Varadhan. Asymptotic evaluation of certain markov process expectations for large time. i, ii, iii. *Communications on Pure and Applied Mathematics*, pages 28 (1975) p.1–47, 28 (1976) p. 279–301, 29 (1976) p. 389–461.

[39] M. Donsker and S. Varadhan. *Large deviations for Markov processes and asymptotic evaluation of certain expectations for large time in "Probabilistic Methods in Differential Equations". Lecture Notes in Math.* Springer, 1975.

[40] Paul Dupuis and Hui Wang. Importance sampling, large deviations, and differential games. *Stochastics and Stochastics Reports*, 76:481–508, 2004.

[41] Michael Ermakov. Asymptotic minimaxity of tests of kolmogorov and omega-square types. *Theory of Probability and its Applications*, 40:54–67, 1995.

[42] Michael Ermakov. Importance sampling for large and moderate large deviation simulation of tests and estimators. *1991 MSC: primary 62E25, secondary 60F10, 65C05*, 1998.

[43] Benjamin VanderSluis et al. Genetic interactions reveal the evolutionary trajectories of duplicate genes. *Molecular Systems Biology*, 6:429, 2010.

[44] W.J. Ewens. The probability of survival of a new mutant in a fluctuating environment. *Heredity*, 43:438–443, 1967.

[45] WJ. Ewens. *Mathematical Population Genetics.* Springer-Verlag New York, LLC, 2004.

[46] W. Feller. *An Introduction to Probability Theory and its Applications.* John Wiley and Sons Inc, 1968.

[47] Jin Feng and Thomas G. Kurtz. *Large Deviations for Stochastic Processes.* AMS : American Mathematical Society, 2006.

[48] R.A. Fisher. The distribution of gene ratios for rare mutations. *Contributions to mathematical statistics*, 50:205–220, May 1930.

[49] Stephen S. Fong, Andrew R. Joyce, and Bernhard Palsson. Parallel adaptive evolution cultures of *Escherichia coli* lead to convergent growth phenotypes with different gene expression states. *Genome Research*, 15:1365–1372, 2005.

[50] Drew Fudenberg, Martin Novak, Christine Taylor, and Lorens A. Evolutionary game dynamics in finite populations with strong selection and weak mutations. *Theoretical Population Biology*, 70(3):352–363, 2006.

[51] P.J. Gerrish and R.E. Lenski. The fate of competing beneficial mutations in an asexual population. *Genetica*, 102(103):127–144, 1998.

[52] J.H. Gillespie. *The Causes of Molecular Evolution.* Oxford University Press, 1991.

[53] Paul Glasserman and S. Kou. Analysis of an importance sampling estimator for tandem queues. *ACM Transactions on Modeling and Computer Simulation*, 4:22–42, 1995.

[54] Paul Glasserman and Yashan Wang. Counterexamples in importance sampling for large deviations probabilities. *The Annals of Applied Probability*, 7(3):731–746, 1997.

[55] D. Gresham, M. Desai, Tucker C.M., H.T. Jenq, and D.A. et al Pai. The repertoire and dynamics of evolutionary adaptations to controlled nutrient-limited environments in yeast. *PLoS Genetics*, 4(12):e1000303, 2008.

[56] P. Groeneboom, J. Oosterhoff, and F.H. Ruymgaart. Large deviation theorems for empirical probability measures. *Annals of Probability*, 7:553–586, 1979.

[57] J.B.S. Haldane. A mathematical theory of natural and artificial selection part V: Selection and mutation. *Proceedings of the Cambridge Philosophical Society*, 23(7):838–844, July 1927.

[58] J.M. Heffernan and L.M. Wahl. The effects of genetic drift in experimental evolution. *Theoretical Population Biology*, 62:349–356, July 2002.

[59] M. Hegreness, N. Shoresh, D. Hartl, and R. Kishony. An equivalence principle for the incorporation of favorable mutations in asexual populations. *Science*, 311:1615–1617, 2006.

[60] R.B. Helling, C.N. Vargas, and J. Adams. Evolution of *E. coli* during growth in a constant environment. *Genetics*, 116:349–358, 1987.

[61] K. Henle and et al. The role of density regulation in extinction processes and population viability analysis. *Biodiversity and Conservation*, 13:9–52, 2004.

[62] Matthew D. Herron and Michael Doebeli. Parallel evolutionary dynamics of adaptive diversification in *Escherichia coli*. *PLoS Biology*, 11(2), 2013.

[63] Lorens A. Imhof and Drew Fudenberg. Imitation processes with small mutations. *Journal of Economic Theory*, 131:251–262, 2005.

[64] Marianne Imhof and Christian Schlötterer. *E. coli* microcosms indicate a tight link between predictability of ecosystem dynamics and diversity. *PLoS Genetics 2*, 103(7):e103, 2006.

[65] Kavita Jain, Joachim Krug, and Su-Chan Park. Evolutionary advantage of small populations on complex fitness landscapes. *Evolution*, 65(7):1945–1955, 2011.

[66] A. Johansonn and D.J.T. Sumpter. From local interactions to population dynamics in site-based models of ecology. *Theoretical Population Biology*, 64:497–517, 2003.

[67] S.B. Joseph and D.W. Hall. Spontaneous mutations in diploid *Saccharomyces cerevisiae*: More beneficial than expected. *Genetics*, 168:1499–1504, 2004.

[68] K.C. Kao and G. Sherlock. Molecular characterization of clonal interference during adaptive evolution in asexual populations of *Saccharomyces cerevisiae* . *Nature Genetics*, 40(12):1499–1504, 2008.

[69] Samuel Karlin and Howard Taylor. *An Introduction to Stochastic Modeling.* Academic Press, 1998.

[70] Yuseob Kim and H Allen Orr. Adaptation in sexuals vs. asexuals: clonal interference and the fisher-muller model. *Genetics*, 171(3):1377–1386, 2005.

[71] M. Kimura. Some problems of stochastic processes in genetics. *Annals of Mathematical Statistics*, 28:882–901, 1957.

[72] M. Kimura. On the probability of fixation of mutant genes in a population. *Genetics*, 47(6):713–719, 1962.

[73] M. Kimura. Model of effectively neutral mutations in which selective constraint is incorporated. *Proceedings of the National Academy of Sciences of the United States of America*, 76(7):3440–3444, July 1979.

[74] M. Kimura. *The Neutral Theory of Molecular Evolution*. Cambridge Univeristy Press, 1983.

[75] F. C. Klebaner and R. Liptser. Asymptotic analysis and extinction in a stochastic lotka volterra model. *The Annals of Applied Probability*, 11(4):1263–1291, 2001.

[76] Solomon Kullback. *Information Theory and Statistics*. Dover Publications, 1997.

[77] R.E. Lenski. Experimental studies of pleiotropy and epistasis in *Escherichia coli*. I. variation in competetive fitness among mutants resistant to virus T4. *Evolution*, 42:425–433, 1988.

[78] R.E. Lenski, M.R. Rose, S.C. Simpson, and S.C. Tadler. Long-term experimental evolution in *Escherichia coli*. I. adaptation and divergence during 2000 generations. *American Naturalist*, 138(6):1315–1341, Dec 1991.

[79] R.E Lenski and M. Travisano. Dynamics of adaptation and diversification: a 10000 generation experiment with bacterial populations. *Proceedings of the National Academy of Sciences USA*, 91:6808–6814, 1994.

[80] B.R. Levin, F.M. Stewart, and L. Chao. Resource limited growth, competition, and predation: a model and experimental studies with bacteria and bacteriophage. *American Naturalist*, 111:3–24, 1977.

[81] Di Liu. A numerical scheme for optimal transition paths of stochastic chemical kinetic systems. *Journal of Computational Physics*, 227:8672–8684, 2008.

[82] B.A. Melbourne and A. Hastings. Extinction risk depends strongly on factors contributing to stochasticity. *Nature*, 454:101–103, 2008.

[83] Sylvie Méléard and N. Fournier. A microscopic probabilistic description of a locally regulated population and macroscopic approximations. *The Annals of Applied Probability*, 14:1880–1919, 2004.

[84] Sylvie Méléard and Sylvie Roelly. Evolutive two-level population process and large population approximations. *Preprints des Instituts für Mathematik der Universität Potsdam*, 2013.

[85] Sylvie Méléard and Denis Villemonais. Quasi-stationary distributions and population processes. *Centre De Mathématiques Appliquées*, UMR CNRS 7641, 2011.

[86] J.A.J. Mertz, S.A.H. Geritz, G. Meséna, F.A.J. Jacobs, and J.S. van Heerwaarden. Adaptive dynamics, a geometrical study of the consequences of nearly faithful reproduction. *In: van Strien, S.J., Verduyn Lunel, S.M. (Eds.), Stochastic and Spatial Structures of Dynamical Systems. North Holland, Amsterdam*, pages 183–231, 1996.

[87] Jacques Monod. *Chance and necessity : an essay on the natural philosophy of modern biology.* Knopf, New York, 1971.

[88] P.A.P. Moran. *The Statistical Processes of Evolutionary Theory.* Oxford University Press, 1962.

[89] Gregory J. Morrow and Stanley Sawyer. Large deviation results for a class of markov chains arising from population genetics. *The Annals of Probability*, 17(3):1124–1146, 1989.

[90] H. J. Muller. Some genetic aspects of sex. *American Naturalist*, 66(703):118–138, 1932.

[91] J.R. Munkres. *Analysis on Manifolds.* Addison-Wesley, 1991.

[92] T. Ohta. Extension of the neutral mutation drift hypothesis. *Proceedings of the Second Taniguchi International Symposium on Biophysics*, pages 148–167, 1977.

[93] Otso Ovaskainen and Baruch Meerson. Stochastic models of population extinction. *arXiv:*1008.1162, 2010.

[94] Tiago Paixão, Dirk M. Lorenz, Jason Songhurst, Michael W.Deem, Robert Azencott, Tim F.Cooper, and Ricardo B.R.Azevedo. Clonal interference can lead to evolutionary farsightedness. *Working Paper.*

[95] Su-Chan Park and Joachim Krug. Clonal interference in large populations. *Proceedings of the National Academy of Sciences USA*, 104(46):18135–40, Nov.2007.

[96] L.L. Fernandez Perfeito, C. Mota, and I.Gordo. Adaptive mutations in bacteria: high rate and small effects. *Science*, 317:813–815, 2007.

[97] A.D. Peters and S.P. Otto. Liberating genetic variance through sex. *BioEssays*, 25:533–537, 2003.

[98] Huŷen Pham. Large deviations in mathematical finance. *Laboratoire de Probabilités et Modèles Aléatoires.*, CNRS, UMR 7599, 2010.

[99] E. Pollack. Fixation probabilities when the population size undergoes cyclic fluctuations. *Theoretical Population Biology*, 57:51–58, 2000.

[100] A. Purvis and et al. Predicting extinction risk in declining species. *Proceedings of Royal Society London*, 267(1465):1947–1952, 2000.

[101] Sean H. Rice. *Evolutionary Theory: Mathematical and Conceptual Foundations*. Sinauer Associates Inc Publishers, 1961.

[102] Alejandra Rodríguez-Verdugo, Brandon S Gaut, and Olivier Tenaillon. Evolution of *Escherichia coli* rifampicin resistance in an antibiotic-free environment during thermal stress. *Evolutionary Biology*, 13(50):1471–2148, 2013.

[103] D.R. Rokyta, C.J. Beisel, P. Joyce, M.T. Ferris, C.L. Burch, and H.A. Wichman. Beneficial fitness effects are not exponential for two viruses. *Journal of Molecular Evolution*, 67:368–376, 2008.

[104] D.R. Rokyta, P. Joyce, S.B. Caudle, and H.A. Wichman. An empirical test of the mutational landscape model of adaptation using a single stranded dna virus. *Nature Genetics*, 37:441–444, 2005.

[105] D. Rozen, J. de Visser, and P. Gerrish. Fitness effects of fixed beneficial mutations in microbial populations. *Current Biology*, 12:1040–1045, 2002.

[106] Daniel E. Rozen, Michelle G. J. L. Habets, Andreas Handel, and J. Arjan G. M. de Visser. Heterogeneous adaptive trajectories of small populations on complex fitness landscapes. *PLoS One*, 3(3):e1715, 2008.

[107] Daniel E. Rozen, J.Arjan, G.M. de Visser, and Philip J. Gerrish. Fitness effects of fixed beneficial mutations in microbial populations. *Current Biology*, 12(12):1040–1045, 2002.

[108] Walter Rudin. *Functional Analysis*. McGraw-Hill, 1991.

[109] J. S. Sadowsky. On the optimality and stability of exponential twisting in monte carlo estimation. *IEEE Transactions of Information Theory*, 39:119–128, 1993.

[110] J. S. Sadowsky and J. A. Bucklew. On large deviations theory and asymptotically efficient monte carlo estimation. *IEEE Transactions of Information Theory*, 36:579–588, 1990.

[111] I. Sanov. On the probability of large deviations of random variable. *Selected Translations in Mathematical Statistics and Probability*, 1:213–244, 1961.

[112] Vasudha Sehgal. *Estimation of Mutation Rates and Selective Advantages in Cell Population Dynamics.* PhD thesis, Univeristy of Houston, 2011.

[113] D. Siegmund. Importance sampling in the monte carlo study of sequential tests. *Annals of Statistics*, 4:673–684, 1976.

[114] Robert Skipper. The heuristic role of Sewall Wright's 1932 adaptive landscape diagram. *Philosophy of Science*, 71:1176–1188, 2004.

[115] Eric Smith. Large-deviation principles, stochastic effective actions, path entropies, and the structure and meaning of thermodynamic descriptions. Technical Report arXiv:1102.3938, 2011.

[116] J. Stoer and R. Bulirsch. *Introduction to Numerical Analysis.* Springer, 2002.

[117] Julien Tailleur and Vivien Lecomte. Simulation of large deviation functions using population dynamics. *arXiv*:0811.1041*v*1, 2008.

[118] Barbara E. Taylor. Analyzing population dynamics of zooplankton. *Limnology and Oceanography*, 33(6):1266–1273, 1988.

[119] Viet Chi Tran. Large population limit and the time behavior of a stochastic particle model describing an age-structured population. *ESAIM: Probability and Statistics*, 12:345–386, 2008.

[120] M van Hoek and P. Hogeweg. The effect of stochasticity on the lac operon: an evolutionary perspective. *PLoS Computational Biology*, 3(6), 2007.

[121] S. R. S. Varadhan. *Large Deviations and Applications.* SIAM: Society for Industrial and Applied Mathematics, 1984.

[122] L.M. Wahl and P.J. Gerrish. The probability that beneficial mutations are lost in populations with periodic bottlenecks. *Evolution*, 55(12):2606–2610, December 2001.

[123] A.D. Wentzell and M.I. Freidlin. On small random perturabtions of dynamical systems. *Russian Mathematical Surveys*, 25:1–55, 1970.

[124] A.D. Wentzell and M.I. Freidlin. Some problems concerning stability under small random perturbations. *Theory of Probability and its Applications*, 17:269–283, 1972.

[125] C.O. Wilke. The speed of adaptation in large asexual populations. *Genetics*, 167:2045–2053, 2004.

[126] Noah Williams. Small noise asymptotics for a stochastic growth model. *Working Paper 10194*, 2003.

[127] Sewall Wright. The roles of mutation, inbreeding, crossbreeding and selection in evolution. *Proceedings of the Sixth Annual Congress of Genetics*, 1:356–366, 1932.

[128] Sewall Wright. Surfaces of selective value revisited. *The American Naturalist*, 131:115–123, 1988.

[129] Wei Zhang, Vasudha Sehgal, Duy M. Dinh, Ricardo B. R. Azevedo, Tim Cooper, and Robert Azencott. Estimation of the rate and effect of new beneficial mutations in asexual populations. *Theoretical Population Biology*, 81(2):168–178, 2012.