8-12-2024 1:30 PM

Impact of fluctuating selection on genetic variation when new mutations are expected to be deleterious

Zahra Shafiei, *Western University*

Supervisor: Bertram, Jason, *The University of Western Ontario*

A thesis submitted in partial fulfillment of the requirements for the Master of Science degree in Applied Mathematics

Abstract

Current research continues to debate the influence of fluctuating selection on genetic diversity within populations. In most previous models of fluctuating selection for studying genetic diversity, the distribution of selection coefficients is assumed to be symmetrical, meaning that the chances of having positive and negative selection coefficients are identical over time. These models predict that selective fluctuations reduce genetic diversity similar to the stochastic influence of genetic drift. Using stochastic simulations and analytical approaches based on diffusion approximations, we analyze the impact of fluctuating selection on genetic diversity when the distribution of selection coefficients over time is not symmetric, but is instead shifted to negative values. This captures the fact that new mutations are more likely to be deleterious. We show that, unlike the symmetric case, selective fluctuations can greatly increase genetic variation when new mutations are deleterious on average. We show that this phenomenon occurs because deleterious mutations that would be kept at low frequency in constant environment are able to transiently attain high frequencies in a changing environment. Our findings suggest that fluctuating selection could be an important force for generating genetic diversity even if it does not lead to long-term coexistence of alternate alleles.

Keywords : Genetic diversity · Selective fluctuation · Biased selection · Deleterious mutations.

Summary for Lay Audience

Current research continues to explore how changing environmental conditions affect genetic diversity within populations. Genetic diversity refers to the variety of different genes within a species. One key question in evolutionary biology is how the changes in the environment impact this diversity. In most previous models that studied genetic diversity, the environmental changes are symmetrical, meaning that positive and negative impacts on genes occurred equally over time. Under this assumption, the models predicted that environmental fluctuations reduce genetic diversity, which is the same behavior observed with the genetic drift (randomness of changes in allele frequencies in a finite population due to the unpredictable nature of reproduction and survival). However, our research takes a different approach where we study what happens when the average of environmental changes is not symmetrical but instead tends to be negative. This means that new mutations are more likely to be harmful (deleterious) rather than beneficial. Our findings show that when environmental changes predominantly cause harm, the results are quite different from the symmetric case as is typically observed in nature. Specifically, environmental fluctuations can actually increase genetic diversity when new mutations are deleterious on average. In essence, our study finds when new mutations are on average harmful, environmental change is a powerful force in creating genetic diversity. Our research highlights the importance of considering asymmetrical environmental impacts in understanding how genetic diversity is maintained and generated in natural populations.

Dedication

To the 12th leader and the last last savior (Imam Mahdi), my beloved family, and my dearest people Ms. Afraz and Ms. Raissi Fard, whose support and wisdom have been a constant source of strength.

Acknowledgements

I would like to express my deepest gratitude to my supervisor, Dr. Jason Bertram. His belief in me, coupled with his patience for my endless questions has meant the world to me throughout this journey. Dr. Bertram's intelligence and insightful guidance have made working with him a truly joyful experience. His kindness has profoundly impacted my academic and personal growth.

I would like to thank my family for their unwavering support and encouragement throughout my academic journey. Your constant understanding has given me the strength to persevere through the toughest times. To my brothers, Saeid and Afshin, my first and best companions, thank you for always being there to cheer me on. Your humor and companionship have been a source of joy and comfort. I could not have accomplished this without your love in my heart.

I would also like to thank those whose light of presence has always been by my side, even from the farthest distances. You have been my hope in darkness, my strength in difficulties, and my greatest source of joy in moments of happiness. Our souls are forever connected.

Table of Contents

List of Figures

Chapter 1

1 Introduction

In evolutionary biology, the concept of selective fluctuation stands out as a reasonable approach since environments are not static; they experience different types of changes due to factors such as climate variability, resource availability, and the presence of predators. In such fluctuating environments, traits that provide a survival advantage at one point may become less advantageous as conditions change.

Despite the dynamic nature of real-world environments, researchers often work with models that assume constant selection. This preference arises primarily due to the significant challenges associated with modeling environmental fluctuations. There are numerous ways to represent these fluctuations without an obvious correct choice of modeling. Additionally, mathematical analysis becomes increasingly complex when we attempt to account for stochasticity of different variable conditions. The difficulty of developing accurate and manageable models and mathematical calculations for fluctuating environments leads many scientists to simplify their studies by assuming environmental constancy. Recognizing the limitations of constant selection models in describing real world scenarios, in this study we focus on working with a model of fluctuating selection.

One of the main reasons for studying fluctuating selection is its significant potential impact on genetic variation because when a population experiences different environmental conditions, it is likely to have a wider range of genes diversity suited to those conditions. The study of heritable variation has a rich history, beginning with Charles Darwin's theory of natural selection in the mid-19th century and Gregor Mendel's foundational work on inheritance patterns in pea plants (Smỳkal et al., 2014). These early discoveries laid the groundwork for the Modern Synthesis of the 1930s and 1940s, which integrated Mendelian genetics with Darwinian evolution. By building on this historical foundation, our research

aims to further our understanding of how fluctuating selection influences genetic diversity.

Most previous studies on the relationship between fluctuating selection and genetic variation have focused on balancing selection (Bertram and Masel, 2019; Wittmann et al., 2017; Svardal et al., 2015; Yi and Dean, 2013; Hedrick, 2006, 1986; Gillespie, 1978; Haldane and Jayakar, 1963). While these studies emphasize the importance of balancing selection, our research intentionally avoids trying to incorporate this mechanism. Balancing selection models, as extensively discussed in the literature, often require a complex set of assumptions to maintain genetic diversity within a population. These assumptions include specific conditions such as frequency-dependent selection, where the fitness of an allele depends on its frequency within the population, or heterozygote advantage, where individuals carrying two different alleles have higher fitness than those with two identical alleles. Additionally, balancing selection often relies on the presence of environmental or temporal fluctuations that favor different alleles at different times, thus preventing any single allele from becoming fixed. While these mechanisms are effective in explaining how polymorphism is maintained under certain conditions, they necessitate a finely-tuned set of parameters and a deep understanding of the specific environmental context in which the population exists. The complexity of balancing selection models, with their dependence on specific assumptions and conditions, can limit their generalizability and applicability across different scenario. In contrast, we aim to explore scenarios where balancing selection assumptions are not a concern. Our approach avoids the complexities of balancing selection by focusing on a model where mutations either go to fixation (mutation replaces wild type allele) or are lost, without the possibility of long-term persistence. This simplicity makes our model easier to analyze and more broadly applicable, allowing us to explore how genetic variation arises in changing environments without relying on the strict assumptions required by balancing selection. Our approach provides clear insights into the evolutionary processes at play, making it a useful tool for understanding genetic diversity in fluctuating environ-

ments. This is crucial because the conditions for balancing selection are often restrictive and highly dependent on specific assumptions.

In the few studies that are not focused on balancing selection such as the works of Huerta-Sanchez et al. (2008) and Takahata and Kimura (1979), selective fluctuation was found to reduce genetic variation. This outcome is intuitive, as selective fluctuation is a stochastic force that can diminish genetic diversity by introducing noise into the model and eventually push alleles to fixation or loss (and hence loss of genetic variation). This is analogous to the diversity-reducing influence of random genetic drift (i.e. the inherent stochasticity of allele frequencies in a finite population due to the unpredictable nature of reproduction and survival).

In this study, we modeled selective fluctuation using the framework of Karlin and Levikson (1974) which assumes white noise, in other words no correlation between selective coefficient parameters over time. In nature, fluctuating selection has temporal autocorrelation over generations, meaning the environment has a memory over time (Takahata et al., 1975). However, temporal autocorrelation makes mathematical analysis complex, which is why we consider white noise in our model (Takahata et al., 1975). This approach, while not entirely reflective of reality, allows for more straightforward and tractable mathematical analysis and makes it possible for us to add one more important assumption.

The studies mentioned above such as Huerta-Sanchez et al. (2008) and Takahata and Kimura (1979), ignored a critical observation that mutations are, on average, deleterious (Eyre-Walker and Keightley, 2007). This arises from the nature of mutations as random changes to the DNA sequence, which are more likely to disrupt than improve. Therefore, mutations are more likely to have harmful effects than beneficial ones, explaining the bias towards deleterious mutations in the distribution of fitness effects of new mutations (Eyre-Walker and Keightley, 2007). Specifically, these studies assumed that fluctuations are symmetric, meaning that the average fitness across the different environments experienced over time is

zero. This assumption amounts to focusing on fluctuations rather than deterministic components (i.e. biases) from their models, concentrating solely on the stochastic aspects of environmental variability. The key difference between our model and previous works in fluctuating selection (Huerta-Sanchez et al., 2008; Takahata and Kimura, 1979) is that we introduce a bias term, namely the negative shift in our selective fluctuation model. The negative shift reflects our main assumption which is new mutations are expected to be deleterious on average.

Chapter 2

2 Methods

2.1 Basic Model Description

We start with the the model of Karlin and Levikson (1974), considering a haploid population with a fixed size which consists of N individuals that has two possible alleles denoted as A and a. We define the fitness of the two alleles as follows:

A	a
$1+\sigma$	$1+\tau$

where σ and τ are random variables representing the presence of selective fluctuation due to environmental changes for type A and a, respectively. We calculate the change in frequency between two successive generations in this population to be:

$$\Delta x = \frac{(\sigma-\tau)}{1+\sigma x+\tau(1-x)}x(1-x) \tag{2.1}$$

where x is the frequency of allele A (i.e. the proportion of individuals that carry allele A in the population). Equation (2.1), is a selection model used frequently in population genetics standard text books such as Crow and Kimura (2009). In the coefficient of $x(1-x)$, which is the selection coefficient, the numerator is the difference in fitness between the two types presented in the population, while the denominator is the mean fitness of the population.

However, unlike the standard selection model, which assumes σ and τ are constant over time, Karlin and Levinkson (1974) assume that σ and τ are statistically independent random variables which take different values each generation. Additionally, σ and τ are uncorrelated over time steps meaning that in any given generation the values of σ and τ are independent of their values in both previous and subsequent generations. These two as-

sumptions result in an extreme form of fluctuations in this model. In nature, there could be correlations between σ and τ, both within a single generation and across multiple generations. This is complex to work with, so for simplicity, we ignore all of these correlations.

Karlin and Levikson (1974) also assume that σ and τ are independent and identically distributed random variables which implies σ and τ have an identical probability distributions and therefore have the same variance and the same expectation. We want to change one of these assumptions. We draw σ and τ, from two distributions with an identical variance v but different expectations. The expectations of these distributions have the same distance from origin. As a result, we have a negative shift in selection coefficient of this model.

To adjust the value of this negative shift in selection coefficient, we define a new parameter called δ and we assume the expectation of σ and τ to be:

$$E(\sigma) = -\frac{\delta}{2}, \quad E(\tau) = \frac{\delta}{2} \tag{2.2}$$

Therefore, the expectation of the difference between σ and τ is:

$$E(\sigma - \tau) \quad = \quad -\delta \tag{2.3}$$

This assumption means that the distribution of selection coefficients over time will be shifted towards negative values.

In the above model, selection causes x to change stochastically. The second source of stochasticity in this model is genetic drift modeled by using the Wright-Fisher model (see Sec. 2.3 for further details of the Wright-Fisher model). Genetic drift also introduces random fluctuations in x, therefore, it will influence the overall genetic variation observed in the population. We model genetic drift by using a binomial sampling method where the success probability in each generation is the updated allele frequency after accounting for

the selective fluctuations and the number of trials is equal to population size. Wright Fisher genetic drift model operates on the principle that each individual produces a substantial number of "potential" offspring each generation and that is why the success probability x in each generation remains constant. Also, the next generation is then formed by randomly sampling a number of individuals equal to the population size from this pool of potential offspring, with the previous generation entirely replaced meaning that we consider non overlapping generations.

We incorporate mutation events in two ways. First, to facilitate comparison with the model proposed by Huerta-Sanchez et al. (2008), our simulation implements mutations exclusively at the lower boundary at $x = 0$. This means that new mutations appear only at ($x = 0$), mirroring the approach taken in their model. The purpose of this consistency is to allow us to directly compare our results with Huerta-Sanchez et al. (2008) results. However, when conducting our analysis in the subsection 3.4.2, we find it more straightforward to use a two-sided symmetric mutation model.

2.2 Quantifying Genetic Variation

In general, genetic variation measures diversity in gene frequencies within and between populations, arising from the effect of new mutations, genetic drift, and natural selection. In particular, within populations, genetic variation is the raw material for evolution, enabling populations to adapt to changing environments or pursue new evolutionary strategies. This variation is crucial for the adaptive potential of populations, enabling them to evolve and respond to environmental changes.

We aim to understand how genetic variation changes under the influence of fluctuating selection and genetic drift when we expect new mutations to be deleterious on average.

There are many ways to quantify genetic variation. In this study, we use the following

formula to calculate genetic variation:

$$V_g = 2E[x(1-x)] \tag{2.4}$$

The logic behind the equation (2.4) is as follows. When we assume a haploid population with two possible alleles denoted as A and a, if we keep x fixed which is the frequency of allele A, then $2x(1-x)$ represents the probability that when we pick two individuals at random, one will have allele A and the other will have allele a (i.e. the two individuals are genetically different). The factor of 2 arises because either allele can be chosen first, since there is no inherent difference between A and a; therefore, we consider both possibilities. This makes $2x(1-x)$ a measure of genetic variation for given x.

We take expectation (E) over a hypothetical infinite number of replicates of this population because in fact x is not fixed but varies over time in any one population. The expectation accounts for different probabilities of x values. This variability in x reflects the random nature of evolution, so we take an expectation to get a typical or representative value of V_g. The following section outlines how this is implemented in practice.

2.3 Simulation

We begin with a vector of replicated allele frequencies x in a population with a fixed size N, with each x initially set to be $\frac{1}{N}$ (i.e. a single individual with the A allele). In each generation, we draw two independent samples of σ and τ from two normal distributions with an identical variance v, and expectations given by equation (2.2).

Genetic drift is modeled with the Wright-Fisher model; therefore, the number of alleles in the next generation is determined by the current allele frequencies after accounting for selection. Specifically, allele frequency in the next generation is sampled from a binomial distribution with N trials with each having the success probability of x which is the updated

frequency after applying Eq. (2.1).

We also assume mutation events occur only at the lower boundary when x hits 0 and not at the upper boundary ($x = 1$), or any other frequency. New mutations are simulated to occur at a specified rate, denoted as μ per locus per individual per generation. When a mutation happens, we set x to be $\frac{1}{N}$ representing a single mutant individual in the population. Conversely, if x reaches the upper boundary ($x = 1$), we set x to 0 to restart the recursion pattern. Without this reset, individuals would eventually accumulate at ($x = 1$); therefore, we would not have any further evolution in the population.

We run this simulation for $10N$ generations and we take a snapshot from the population when we have the steady-state. The steady-state is when the shape of the distribution of replicate frequencies is constant over time if we have infinitely many trajectories. Our simulated distribution will not be exactly steady because it is a finite sample, but we take a big sample with the size of 1 million trajectories so it should be approximately constant over time. We obtain site frequency spectrum of our model at steady-state and we retain only the final frequencies, disregarding intermediate values.

All codes used to perform simulations and produce figures are available at:
https://github.com/zahra-shafiei04/fluctuating-selection.

2.4 Parameter Specification

In this section, we explain our choices for the range of the fluctuating selection parameter v and the negative shift δ in selection coefficient that will be used in Chapter 3. Additionally, we discuss the rationale behind setting the population size $N = 1000$.

To determine the range for v, we need to examine the equation (2.1). Our approach involves choosing a magnitude for the selection coefficients that is representative of typical values seen in each generation empirically. Essentially, this means determining whether

the distribution from which we draw σ and τ is wider or narrower. If the distribution is too wide, it would imply that the per-generation selection coefficient could be very large, which would not be applicable to real-world scenarios. In order to achieve this aim, we approximate the equation (2.1) where the coefficient of $x(1-x)$ is the selection coefficient s. Since σ and τ are much smaller than 1, considering $z = \sigma x + \tau(1-x)$, we can use Taylor series to approximate Eq. 2.1 to be:

$$\frac{(\sigma - \tau)}{1 + \sigma x + \tau(1-x)} x(1-x) \approx (\sigma - \tau)x(1-x)[1 - \sigma x - \tau x(1-x)] \tag{2.5}$$

Since we have $|z| < 1$, therefore $[1 - \sigma x - \tau x(1-x)] \approx 1$. As a result the selection coefficient in this model can be approximated by:

$$s \approx \sigma - \tau \tag{2.6}$$

Which is the selection coefficient. We compare typical values of the selection coefficient, which essentially represent the standard deviation of the selection coefficient distribution over time. Since the standard deviation is simply the square root of the variance, we first start with the variance of selection coefficient:

$$Var(s) = E(s^2) - E((s))^2 \tag{2.7}$$

Our goal is to establish an estimation of the impact of fluctuations on the selection coefficient with considering the portion of the variance in the selection coefficient that is attributable to these fluctuations. We are not concerned with the total amount of variance of the selection coefficient but rather focused on the component driven solely by fluctuations which allows us to quantify the effect of fluctuations on the selection coefficient. To quantify the effect of fluctuations, we exclude the component associated with the negative shift which is the second part of the Eq. 2.7. Based on the Eq. 2.3, we have $(E(s))^2 = \delta^2$

and δ^2 should be omitted from the measurement since δ^2 represents the part not related to fluctuation effects. Based on the approximation in equation (2.6) we can express the effect of selective fluctuations on the selection coefficient as:

$$E(s^2) = E(\sigma - \tau)^2 = E(\sigma)^2 + E(\tau)^2 - 2E(\sigma\tau) \tag{2.8}$$

Given that σ and τ are defined as independent variables, $E(\sigma\tau) = 0$. Since we assume they have identical variances, we set $E(\sigma)^2 = E(\tau)^2 = v$. By substituting these assumptions into equation (2.8), we conclude:

$$E(s^2) = E(\sigma - \tau)^2 = 2v \tag{2.9}$$

We select the range of v to be $[0, 0.1]$ interval, resulting in selection coefficient to be in the range of $0 \leq s \leq \sqrt{0.2} \approx 0.44$. The maximum value is significantly larger than what is typically observed in real-world scenarios which is approximately $s \sim 0.01$. However, this range ensures we see the full range of model behaviors for the parameter v specifically for the genetic variation plot presented in Chapter 3 (see Figure 3.1).

With the choice of the fluctuating selection parameter v addressed, we next focus on determining the appropriate values for the negative shift δ and population size N. In this model, the relative influence of selection and drift on allele frequency dynamics depends on frequency. In Chapter 3, we will analyze this in more detail. For now, we outline a comparison of the strength of the negative shift and genetic drift to justify our choice of parameter values. We recall one well-known result is the relative balance between a random genetic drift and a constant selection pressure with selection coefficient s. The frequency at which these two forces have comparable influence on allele frequencies is given by $x = \frac{1}{2Ns}$. For any given value of x less than this value, genetic drift is the dominant force, while for any given values of x above this point, selection dominates. This principle has been widely used in

numerous studies, including Desai and Fisher (2007b). We also applied this principle in our study.

We leverage this result to define the scale between the stochastic force of genetic drift, parameterized by N and the deterministic force of the negative shift parameterized by δ. The scale is given by:

$$x = \frac{1}{2N\delta} \tag{2.10}$$

To study the impact of the negative shift δ, we must choose N and δ so that $x \ll 1$. Otherwise, genetic drift will dominate over a large proportion of the x domain, making the negative shift irrelevant. As a result, we chose $N = 1000$ and $\delta \sim 0.005$ to achieve a scale of $\frac{1}{10}$ as given in the Eq. (2.10) which means $x = \frac{1}{10}$ is the frequency at which the negative shift starts to matter, allowing us to study the effect of δ in this model.

Chapter 3

3 Results

3.1 Genetic Variation

Our analysis of genetic variation reveals significant insights into how fluctuating selection and genetic drift impact the diversity of allele frequencies within a population in the presence of a negative shift in the expectation of selection coefficients. In this chapter, we will discuss these results.

We calculate and visualize genetic variation across different values of selective fluctuation parameters v and δ. The genetic variation (V_g) is visualized across the parameter space providing a clear visual representation of how genetic variation changes with these parameters which is presented in Fig. 3.1.

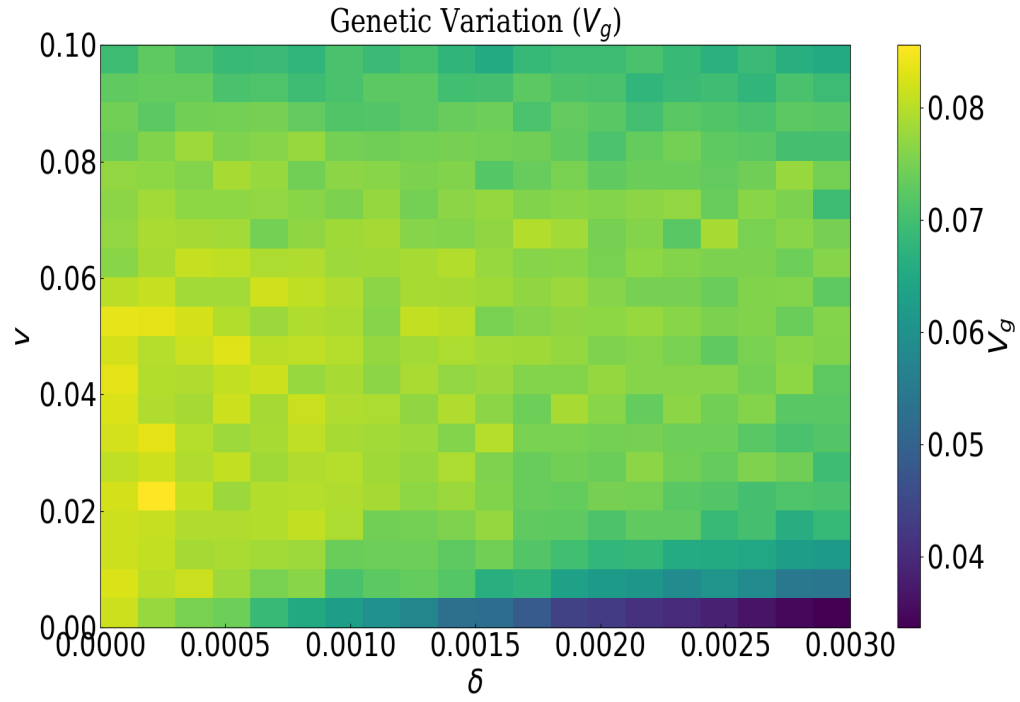

Figure 3.1: Heat map depicting V_g across different negative shift δ and fluctuation values v computed from simulations using equation (2.4). Each cell in the heat map corresponds to the calculated V_g for the specific combination of negative shift and selective fluctuation values. The color intensity indicates the magnitude of V_g, with darker shades representing lower variation and lighter shades representing higher variation. There is a dark region in the lower right part of the plot, where $v \approx 0$ and $\delta > 0$ where the negative shift in selection coefficients depletes variation. Increasing v even slightly in that region restores this variation.

By calculating genetic variation across different values of v and δ, we identify a clear pattern where increasing fluctuation, when $\delta = 0$, leads to a slight reduction in genetic variation. This observation aligns with findings from other studies in population genetics, which suggest that adding a stochastic element, like fluctuating selection, without any bias such as negative shift reduces genetic diversity, corroborating the findings of Huerta-Sanchez et al. (2008) and Takahata and Kimura (1979).

On the other hand, as the negative shift values increase, genetic diversity declines in the absence of selective fluctuations ($v = 0$). If we focus on the rightmost corner of the heat

map, where the negative shift value is at the highest and we have the lowest amount of fluctuation, genetic variation is at minimum. With the same value of δ as fluctuation increases, V_g doubles first and then decreases again with further fluctuation.

In terms of the speed of impact, the negative shift has a significantly more intense effect on genetic variation compared to selective fluctuation. For instance, a very small negative shift $\delta = 0.001$ is enough to significantly influence genetic variation and clearly observe the pattern discussed above. This highlights the importance of accounting for the presence of a bias $\delta > 0$, since even small amount of the negative shift in selection coefficient has a significant impact on genetic variation in comparison with fluctuation and changes the prediction.

Overall, our findings underscore the significant influence of the selective fluctuation on genetic variation in the presence of negative shift. Figure 3.1 illustrates that fluctuations lead to an increase in genetic variation, contrasting with observations made in the absence of the negative shift. This unexpected result prompts a deeper investigation into the underlying mechanisms. In this section, we calculated genetic variation as a numerical relationship between the parameters v and δ, which, while informative, does not provide a comprehensive explanation of the observed phenomena. To clarify the reasons behind this phenomenon, we will calculate the site frequency spectrum in next section, a frequency-dependent analysis that offers deeper insights into the distribution of allele frequencies within a population.

3.2 Site Frequency Spectrum

The site frequency spectrum (SFS) is a fundamental tool in population genetics, providing a summary of the allele frequency distribution within a population. The site frequency spectrum details how many alleles are present at different frequencies in a sample. Mathematically, the site frequency spectrum is a vector that consists of the counts of sites with

frequency x for $0 \leq x \leq 1$.

In this study, x values are used to generate the site frequency spectrum by creating the histogram of the allele frequencies at the final simulated generation. We then compare these results to the probability density $\phi(x)$, which is the continuous analog of the SFS predicted by the diffusion models (after appropriate normalization) that will be introduced in section 3.3 below. We want to ensure that the diffusion approximation is accurately reproducing the simulated SFS, so we plot both the simulated and analytical solutions of the diffusion approximation on the same graph. We have an analytical solution for $\phi(x)$ for the case of no negative shift ($\delta = 0$) from ref. Huerta-Sanchez et al. (2008), which allows us to make direct comparisons.

The SFS helps in understanding the evolutionary dynamics by comparing the observed data with theoretical models, thus providing insights into the historical and ongoing processes shaping the genetic diversity of populations. By examining the distribution over x, which is frequency-dependent, we gain the ability to observe how many mutations are at each specific frequency. On the other hand, the V_g graph depends on the model parameters δ and v, providing a single computed value for V_g. Each approach offers unique insights into the genetic dynamics of populations.

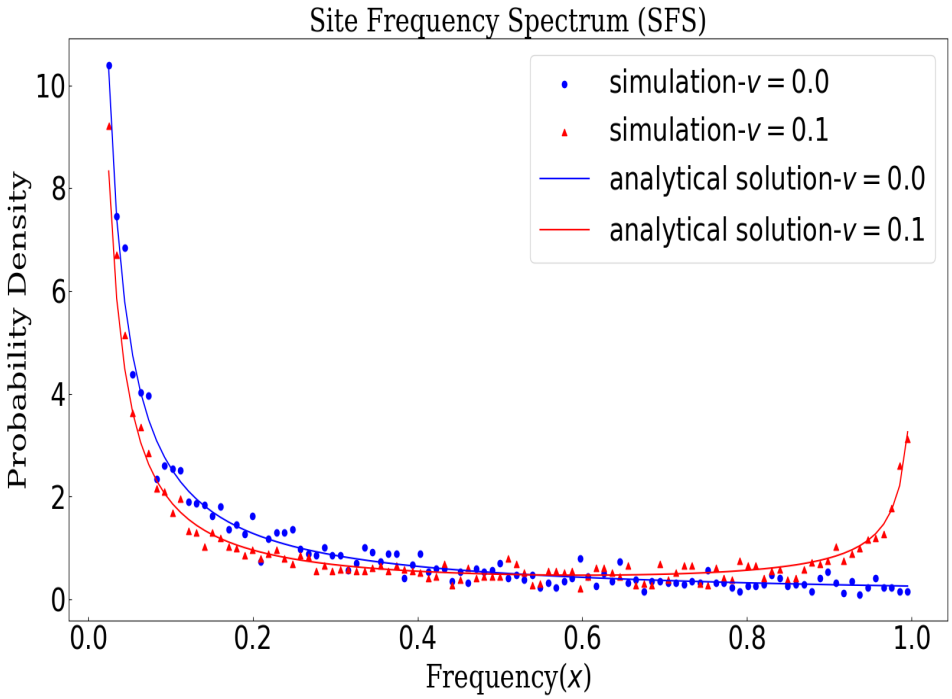

Figure 3.2: The histogram visualizing the site frequency spectrum representing in two scenarios one with no fluctuation ($v = 0$) in blue and one with high fluctuation ($v = 0.1$), both without a negative shift ($\delta = 0$). When $v = 0.1$, the frequency distribution shows a higher concentration at the upper boundary $x = 1$, with a corresponding decrease at the lower boundary $x = 0$ which implies that with increasing fluctuation, more mutations are making it to higher frequencies. The solution of the diffusion approximation, which is the analytical function in solid line, is completely following the behavior expected from simulation curves.

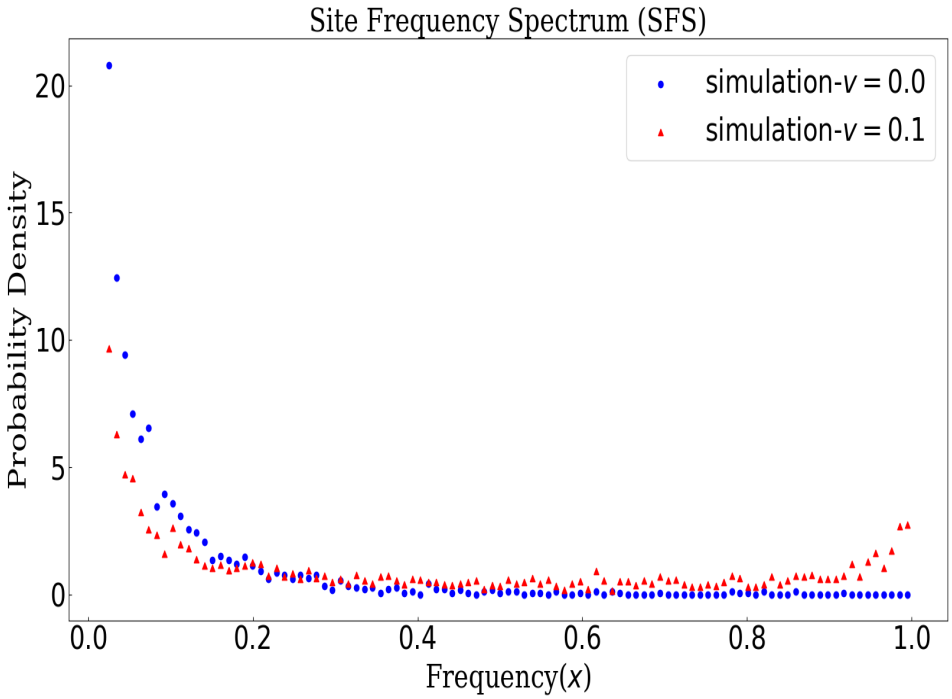

Figure 3.3: The plot represents the site frequency spectrum under two scenarios: no fluctuation ($v = 0$) in blue and high fluctuation ($v = 0.1$) in red, both with a negative shift of $\delta = 0.003$. The red points consistently remain above the blue points after $x \approx 0.25$, indicating higher variation when we have selective fluctuation and negative shift especially at intermediate values of x. Unlike the previous figure, this plot does not include the analytical curve since we are exploring a new space where we have negative shift in our diffusion approximation equation (3.1) for which we do not have an analytical solution.

In comparing Figs. 3.2 and 3.3, we observe that in Fig. 3.3, the red line with $v = 0.1$ remains above the blue line with $v = 0$ for more values of x compared to Fig. 3.2. In Fig. 3.3 the red line intersects the blue line at $x \approx 0.25$, whereas in Fig. 3.2, the intersection occurs at $x \approx 0.6$. This earlier intersection in the second plot highlights that we get more variation sooner with the presence of a negative shift. Therefore, when we have the negative shift ($\delta > 0$), causes a stronger release of mutations to higher frequencies compared to when there is no negative shift.

Note that Fig. 3.3 does not include the analytical curve due to the complexity of introducing negative shift, which leads to complicated terms in the diffusion and directional coefficient

of the equation (3.7) and makes it challenging to derive an analytical solution.

3.3 Diffusion Approximation

To analyze a complicated model such as this, one of the most powerful tools that we can use is the diffusion approximation. The diffusion approximation simplifies the analysis of the dynamics of allele frequency changes over time by converting a discrete model generated with the simulation into a continuous one. This approach allows us to apply calculus, making it possible to derive and solve differential equations that describe the behavior of allele frequencies under evolutionary forces of mutations, genetic drift and neutral selection. (Allen, 2003)

In fact, one of the key applications of the diffusion approximation is in evaluating the site frequency spectrum, which enables us to understand how the dynamics of allele frequency change over time. In the context of population genetics modeling, the probability density function which is the solution of the diffusion approximation serves as a continuous approximation of the site frequency spectrum.

In this section, we use the framework provided by Huerta-Sanchez et al. (2008) to write down the forward Kolmogorov diffusion equation, known as the Fokker-Planck equation. The basic form of the Fokker-Planck equation is:

$$\frac{\partial \phi(x,t)}{\partial t} = -\frac{\partial}{\partial x}[b(x)\phi(x,t)] + \frac{1}{2}\frac{\partial^2}{\partial x^2}[a(x)\phi(x,t)] \tag{3.1}$$

where $\phi(x,t)$ is the probability density function for the allele frequency x at time t given that an allele started at x_0 at $t = 0$. The function $\phi(x)$ that satisfies $\frac{\partial \phi(x)}{\partial t} = 0$ is the steady-state solution of the Fokker-Planck equation.

In equation (3.1), $a(x)$ is the diffusion coefficient describing the stochastic parts of our

model which are drift and selective fluctuation

$$a(x) = \frac{x(1-x)}{2N} + x^2(1-x)^2 E(\sigma - \tau)^2 \tag{3.2}$$

and $b(x)$ describes the directional part of the model including the negative shift in selection coefficient values:

$$b(x) = x(1-x)(2N)\left[E(\sigma - \tau) - E\left(\frac{\sigma^2 - \tau^2}{2}\right) + E(\sigma - \tau)^2\left(\frac{1}{2} - x\right)\right] \tag{3.3}$$

Here $E(\sigma - \tau)^2 = 2v$ based on the Eq. (2.9) and we can impose (2.3) in these equations as well.

To solve the forward Kolmogrov differential equation, an assumption about the initial density is required. We assume that the initial density is concentrated at an initial frequency $x_0 \neq 0$ which means the probability density is given by a Dirac delta distribution, which can be written as:

$$\phi(x, 0) = D(x - x_0). \tag{3.4}$$

where $D(x) = 0$ for all $x \neq 0$ describing the main property of Dirac delta function and we have $\int_{-\infty}^{\infty} D(x)dx = 1$ since it is a distribution.

The boundary conditions of this equation are complicated (Kimura, 1964) and determined by the mutation model. When we assume a mutation model in our analysis below, it will set these boundary conditions implicitly. These conditions ensure that the problem is well-posed and the solutions are meaningful within the context of population genetics.

In the following analysis we will make use of a valuable connection between diffusion models and stochastic differential equations. For any given forward Kolmogrov PDE such as a Fokker-Planck equation (3.1) we can write an equivalent stochastic differential equa-

tion (SDE) corresponding to that. As a result the SDE corresponding to the equation (3.1) is (Allen, 2003):

$$dx(t) = b(x)dt + \sqrt{a(x)}dW(t) \tag{3.5}$$

The stochastic process $\{x(t) : t \in [0; 1)\}$ is said to satisfy an Itô stochastic differential equation (SDE), if for $t \geq 0$ it is a solution of the integral equation:

$$x(t) = x(0) + \int_0^t b(x(\zeta), \zeta)\, d\zeta + \int_0^t a(x(\zeta), \zeta)\, dW(\zeta) \tag{3.6}$$

where the first integral is a Riemann integral and the second integral is an Itô stochastic integral (Allen, 2003).

The stochastic process $\{W(t) : t \in [0; \infty)\}$ is a Wiener process (standard Brownian motion) if W(t) depends continuously on t, $W(t) \in (-\infty, \infty)$, and the following three conditions hold:

1. for $0 \leq t_1 < t_2 < \infty$, $W(t_2) - W(t_1)$ is normally distributed with mean zero and variance $t_2 - t_1$; that is, $W(t_2) - W(t_1) \sim N(0, t_2 - t_1)$.
2. for $0 \leq t_1 < t_2 < \infty$, the increments $W(t_1) - W(t_0)$ and $W(t_2) - W(t_1)$ are independent.
3. $Prob\{W(0) = 0\} = 1$

This definition implies that the Wiener process $W(t)$ has stationary and independent increments. Melsa and Sage (2013) relate the Wiener process to the physical process observed by Brown and the concept of Brownian motion. Suppose $W(t)$ is the displacement from the origin at time t of a small particle. The displacement of a particle over the time interval t_1 to t_2 is long compared to the time between impacts. The central limit theorem can be applied to the sum of a large number of these small disturbances so that it can be assumed $W(t_2) - W(t_1)$ has a normal density. The density of the particles' displacement depends on

the length of the time interval and not on the time of observation; therefore, the probability density of the displacement from time t_1 to t_2 is the same as from time $t_1 + t$ to time $t_2 + t$ (Allen, 2003).

3.4 Derivation of Probability Density Functions

We begin by substituting the diffusion coefficient $a(x)$ (3.2) and the directional coefficient $b(x)$ (3.3) into the Fokker-Planck equation (3.1) to obtain:

$$\frac{\partial \phi(x,t)}{\partial t} = -\frac{\partial}{\partial x}\left[x(1-x)\left(-\delta + 2v\left(\frac{1}{2} - x\right)\right)\phi(x,t)\right] + \frac{1}{2}\frac{\partial^2}{\partial x^2}\left[\frac{x(1-x)}{2N} + 2vx^2(1-x)^2\phi(x,t)\right] \tag{3.7}$$

Solving this partial differential equation (PDE) is exceptionally challenging due to the presence of both genetic drift and the selective fluctuation in $a(x)$. No known analytical solution exists, even under the condition where $v = 0$ if $\delta \neq 0$ (Kimura, 1964).

Fortunately, we are only trying to solve this equation at the steady state, which simplifies the problem. By considering the special case where $\phi(x,t) = \phi(x)$, the problem reduces to a second-order ordinary differential equation (ODE) instead of a PDE with difficult boundary conditions.

Eq. (3.7) is not easy to solve even at the steady state; therefore, we choose an approach which simplifies the equation by eliminating terms corresponding to forces that are weak at a given frequency. As a result, we aim to solve Eq. (3.7) for the steady state in two situations:

1. considering only the negative shift δ and selective fluctuation v without genetic drift N which leads to a local solution.This only applies far enough away from $x = 0, 1$ that the influence of drift can be neglected compared to the influence of the negative

shift.

2. considering the negative shift δ and genetic drift N without fluctuating selection v, leading to a well-known global solution.

Therefore, we are essentially activating each stochastic component of our PDE separately, each time in the presence of the negative shift in our model.

We did not pursue numerical approaches because our goal when using the diffusion approximation is to achieve an analytical understanding of the underlying processes. Numerical methods would only produce more simulated curves and not provide the exact analytical insights we seek.

3.4.1 Local Probability Density Function for Negative Shift and Selective Fluctuation

To focus on negative shift and selective fluctuation effects only, we consider the population size to be big enough that the genetic drift effect vanishes from $a(x)$. Therefore, we have the simplified partial differential equation:

$$\frac{\partial \phi(x,t)}{\partial t} = \frac{1}{2}\frac{\partial^2}{\partial x^2}\left[2vx^2(1-x)^2\phi(x,t)\right] - \frac{\partial}{\partial x}\left[x(1-x)\left(-\delta + v\left(1-2x\right)\right)\phi(x,t)\right] \tag{3.8}$$

To find the steady-state solution of the Fokker-Planck equation (3.8), we set $\frac{\partial \phi(x,t)}{\partial t} = 0$ and thus:

$$-\frac{dF}{dx} = 0 \tag{3.9}$$

where

$$F(x) = -\frac{1}{2}\frac{\partial}{\partial x}\left[2vx^2(1-x)^2\phi(x)\right] + \left[x(1-x)\left(-\delta + v\left(1-2x\right)\right)\phi(x)\right] \tag{3.10}$$

is called the probability flux. Intuitively, if we take a snapshot at a given frequency, the density of the trajectories crossing that snapshot in one direction is the probability flux, essentially the overall flow of probability at that point. Equation. (3.9) implies $F(x)$ is equal to a constant number c (i.e. frequency-independent).

We assume mutations occur at both boundaries, $x = 0$ and $x = 1$ at equal rates per individual, with no mutation flux away from the boundaries. As a result of this mutation symmetry, we can set $c = 0$, simplifying the equation and making it solvable, as noted by ref. Kimura (1964). Any non-zero net flux would be the result of a small surplus of mutations reaching to fixation. This flux is so minor that it does not significantly alter the curves, allowing us to approximate $F(x)$ as zero. However, if mutations were only from one side, this approximation may not hold.

Note that in this section, we are not looking for the exact analytical curve but rather studying the overall rate of decay for the probability density functions. Approximating the probability flux as zero allows us to understand the general shape of the solutions and compare them effectively.

As a result, we have a first order ordinary differential equation which is solvable using the integration factor method. Therefore, we can rearrange (3.10) to have:

$$-vx^2(1-x)^2\phi'(x)+[-v2x(1-x)^2+2v(1-x)x^2+ \\ x(1-x)(-\delta+v(1-2x)]\phi(x)=0. \tag{3.11}$$

Therefore we will have:

$$\phi'(x)+\left[\frac{2}{x}-\frac{2}{1-x}-\frac{-\delta+v(1-2x)}{vx(1-x)}\right]\phi(x)=0. \tag{3.12}$$

We evaluate the integral of $g(x)$ which is the coefficient of $\phi(x)$ in Eq. (3.12) and then we

know that $G(x) = \exp \int g(x)\, dx$, as a result $G(x)$ is:

$$G(x) = x^{(1+\frac{\delta}{v})}(1-x)^{(1-\frac{\delta}{v})}. \tag{3.13}$$

We multiply $G(x)$ to both side of the equation (3.12) and we solve for $\phi(x)$ so we have:

$$\phi_1(x) = \frac{k_1}{x(1-x)}\left(\frac{x}{1-x}\right)^{-\frac{\delta}{v}}. \tag{3.14}$$

which to our knowledge is a new result relating selective fluctuation and the negative shift.

3.4.2 Local Probability Density Function for Negative Shift and Genetic Drift

To concentrate exclusively on the effects of negative shift and genetic drift, we assume that there is no fluctuation in the model $v = 0$. This assumption results in genetic drift being the sole factor in $a(x)$ and simplifies $b(x)$ by eliminating the second term $2v(\frac{1}{2} - x)$. Consequently, the partial differential equation becomes:

$$\frac{\partial \phi(x,t)}{\partial t} = \frac{1}{2}\frac{\partial^2}{\partial x^2}\left[\frac{x(1-x)}{N}\phi(x,t)\right] - \frac{\partial}{\partial x}[(-\delta)x(1-x)\phi(x,t)] \tag{3.15}$$

To determine the steady-state solution of the Fokker-Planck equation (3.15), we have the infinite time limit same as before which leads to have the probability flux function:

$$F(x) = -\frac{1}{2N}\frac{\partial}{\partial x}\left[x(1-x)\phi(x)\right] + \left[(-\delta)x(1-x)\phi(x)\right] \tag{3.16}$$

Where $-\frac{dF}{dx} = 0$. As before, mutations are considered to be symmetric in our analytical discussion, there is no flux in this analysis and the probability flux function equals zero:

$$\left[-\frac{x(1-x)}{2N}\right]\phi'(x) + \left[-\frac{(1-2x)}{2N} + (-\delta)x(1-x)\right]\phi(x) = 0 \tag{3.17}$$

As a result, we obtain a first-order ordinary differential equation which is solvable using the integration factor method. The first step in the integration factor method is rearranging the equation so that the coefficient of the first order derivative $\phi'(x)$ equals to 1. Therefore we will have:

$$\phi'(x) + \left[\frac{(1-2x)}{2N} + (-\delta)x(1-x)\right]\phi(x) = 0 \tag{3.18}$$

Therefore, the integration factor of Eq. (3.18) is:

$$G(x) = x(1-x)e^{2N\delta x} \tag{3.19}$$

We multiply $G(x)$ to both side of the equation (3.18) and we solve for $\phi(x)$:

$$\phi_2(x) = \frac{k_2}{x(1-x)}e^{-2N\delta x} \tag{3.20}$$

In constant selection models, the Eq. 3.20 is a well known result where δ could be the selection coefficient.

3.4.3 Analytical Solutions Discussion

In this section, we compare the analytical solutions for local probability density functions derived in the previous sections 3.4.1 and 3.4.2 compare their decay rates.

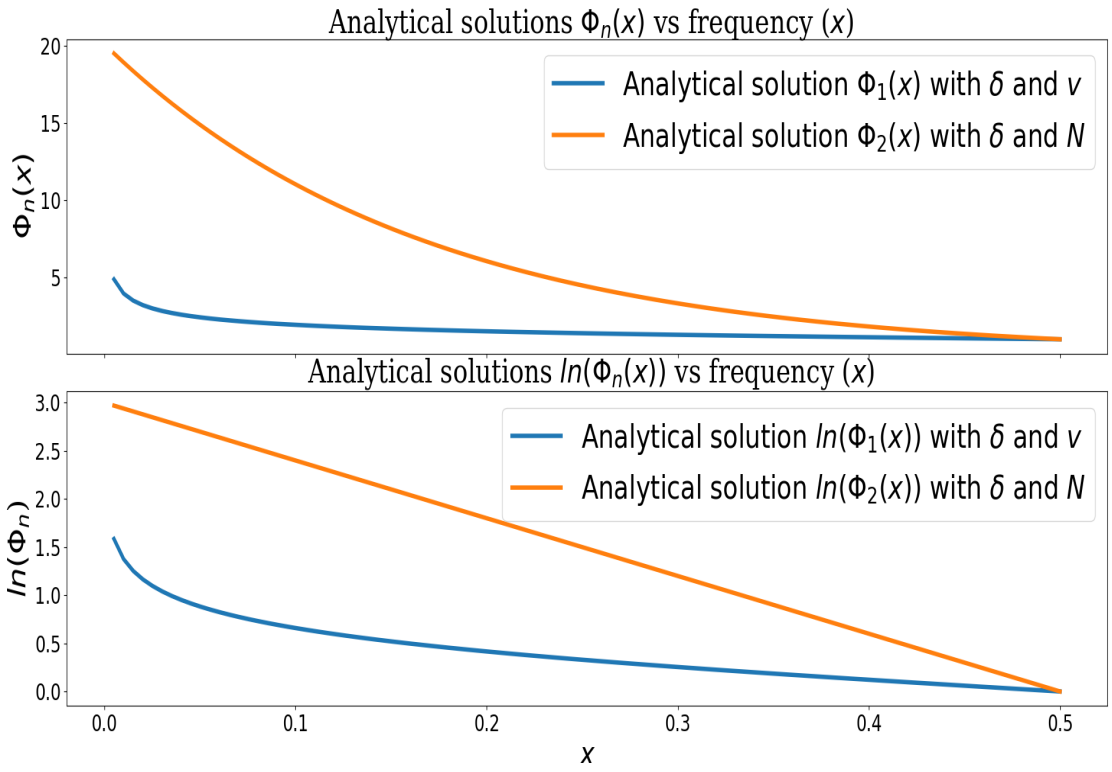

Figure 3.4: The figure illustrates the analytical solutions for probability density functions, $\phi_n(x)$. Both panels depict frequencies ranging from 0 to 0.5 on the x-axis. In the second panel, the natural logarithm of the y-axis values from the first panel is visualized. The purpose of this plot is to compare the decay rates of the analytical solutions. As anticipated, exponential probability density function $\phi_2(x)$ exhibits a much faster decay at intermediate frequencies due to the steeper negative slope in both panels.

The calculation for the exact values of the normalization constants k_1 and k_2 is not possible due to the fact that calculating k_1 needs an integral over whole domain but we know that $\phi_1(x)$ is only known locally. Therefore, we have to find the same starting or ending coordinate for both functions for comparative plotting. Since we are focusing on the lower frequencies, we aligned the final points of both functions at $x = 0.5$. As a result, we analyze the relative changes of $\phi_1(x)$ and $\phi_2(x)$ to understand how quickly $\phi_1(x)$ decays compared to $\phi_2(x)$ (see Fig. 3.4).

The plot confirms our expectation that $\phi_2(x)$, which has an exponential factor, decays faster than the rational function $\phi_1(x)$. This means that in $\phi_2(x)$ scenario when we have the negative shift and genetic drift, fewer mutations are reaching to higher frequencies and most

of them are sticking in the lower boundary at $x = 0$. In this scenario the negative shift is preventing mutations to reach higher frequencies much more than $\phi_1(x)$ where we have selective fluctuation. We can conclude that selective fluctuation is indeed the stochastic force that pushes mutations to higher frequencies in presence of the negative shift and makes the slope of the decay of the probability density function slower so that we have more mutations at higher frequencies. The second panel, utilizing natural logarithms for y-values, transforms the exponential decay into a linear function, highlighting the contrast between the two analytical solutions.

3.5 Dominant Force Analysis

The site frequency spectrum described in the previous section provides the general functional form of the genetic variation observed in Fig.3.1. While we demonstrated that the shapes of the probability density functions differ significantly, this does not explain the specific patterns seen in Fig.3.1 or the parameter values at which these patterns occur. In this section, we compare the negative shift in selection coefficient, selective fluctuation, and genetic drift forces in pairs. By identifying intervals where each force is dominant and defining the boundaries where the dominant force changes, we aim to explain the genetic variation plot.

3.5.1 Selective Fluctuation vs. Genetic Drift

We begin with comparing selective fluctuation and genetic drift which are two stochastic components of our model to identify the relationship between their parameters v and N, respectively.

To achieve this aim, we take a deeper look to diffusion coefficient $a(x)$ in equation (3.2) where we can see the balance between these two stochastic forces.

By factoring $\frac{1}{2N}$ from the Eq. (3.2) we will have:

$$a(x) = \frac{1}{2N}\left[x(1-x) + x^2(1-x)^2 2NE(\sigma-\tau)^2\right] \tag{3.21}$$

where we can see the ratio between v and N, ignoring their x dependence, to be (Huerta-Sanchez et al., 2008):

$$\beta = 2NE(\sigma-\tau)^2 = 4vN \tag{3.22}$$

We want to compare selective fluctuation and genetic drift together. We rearrange Eq. (3.21) and impose β which yields to:

$$a(x) = x(1-x)[1+\beta x(1-x)] \tag{3.23}$$

We solve (3.23) for $\beta x(1-x) = 1$ to establish a comparison between these two forces and conclude that $\beta = 1$ is the threshold of this comparison; therefore, we have 3 situations:

1. $\beta \ll 1$: genetic drift dominates for all x. This occurs because x values range between 0 and 1, where selective fluctuation, being quadratic, exerts influence at a higher order compared to the linear influence of drift.

2. $\beta \sim 1$: selective fluctuation starts to matter and be comparable with genetic drift.

3. $\beta \gg 1$: $a(x)$ is a fourth-order polynomial having four real roots where $x_1 = 0$ and $x_4 = 1$ (the maximum and minimum possible values of x) and also we solve:

$$\beta x(1-x) = 1 \tag{3.24}$$

As a result, we will have:

$$x_2 = \frac{1 - \sqrt{1 + \frac{4}{\beta}}}{2}, \quad x_3 = \frac{1 + \sqrt{1 + \frac{4}{\beta}}}{2} \tag{3.25}$$

As we can see, these are four intersections of the genetic drift and selective fluctuation curves where genetic drift dominates from x_1 to x_2, selective fluctuation dominates from x_2 to x_3, and then drift dominates again from x_3 to x_4.

Therefore, for selective fluctuations to matter (compared to drift) we need β to be at least ~ 1. Even if $\beta \gg 1$, genetic drift still predominates near the boundaries $x = 0, 1$, whereas selective fluctuation predominates at intermediate frequencies.

3.5.2 Genetic Drift vs. Negative Shift

In the previous section we compared the two stochastic components of this model. Now, it is time to take a step forward and compare a stochastic force with a directional part for the other two pairs of forces, which is more challenging.

We use stochastic differential equations since they allow us to directly model how a variable changes over time due to different parts of our equation. Unlike Kolmogorov equations such as equation (3.1), which describe the distribution of the variables rather than the variables themselves, SDEs can provide a more intuitive approach by explicitly detailing the frequency changes with specific forces through the term $dx(t)$ on the left-hand side.

We start with comparing genetic drift with parameter (N) and the negative shift in selection coefficient (δ), setting $v = 0$. The Kolmogorov PDE describing the dynamics is:

$$\frac{\partial \phi(x,t)}{\partial t} = -\frac{\partial}{\partial x}[\delta x(1-x)\phi(x)] + \frac{\partial^2}{\partial x^2}\left[\frac{x(1-x)}{2N}\phi(x)\right] \tag{3.26}$$

Since we consider $v = 0$, the stochastic part of the equation (3.26) only carries genetic drift and eliminates selective fluctuation and the directional part only represents the negative

shift. We write the SDE corresponding to the equation 3.26:

$$dx(t) = \delta x(1-x)dt + \sqrt{\frac{x(1-x)}{2N}}dW(t) \tag{3.27}$$

The first step is to find the duration required for a small increase in frequency due to the negative shift in the selection coefficient. The change in frequency due to only the negative shift (i.e. eliminating genetic drift) is the integral of $dx(t)$ over a small time interval $[0, \epsilon]$, we have:

$$\Delta_\delta x = \int_0^\epsilon dx(t)dt = \int_0^\epsilon \delta x(1-x)dt \tag{3.28}$$

Since we assume that we are considering the change in allele frequency in a small time interval, the change in x is not big over $[0, \epsilon]$ and we have $\Delta_\delta x \ll 1$ (i.e. $x \approx constant$ over $[0, \epsilon]$) and we approximate Eq. (3.28) to be:

$$\Delta_\delta x = \int_0^\epsilon \delta x(1-x)dt \approx \delta x(1-x)\epsilon \tag{3.29}$$

This is the change in frequency due to the negative shift which is basically a flat curve from 0 to ϵ which can be rearrange as follows

$$\frac{\Delta_\delta x}{x} = \delta(1-x)\epsilon \tag{3.30}$$

We call the small relative change in frequency due to the negative shift to be equal to η_1:

$$\frac{\Delta_\delta x}{x} = \eta_1 \tag{3.31}$$

Looking for the time interval at which we have a small change in frequency due to negative

shift, we solve Eq. (3.30) for ϵ imposing Eq. (3.31) leading to have

$$\epsilon = \frac{\eta_1}{\delta(1-x)} \tag{3.32}$$

With having ϵ calculated, now we know how much time does it take for the negative shift to make a small change in frequency. Now, we need to evaluate the change in frequency due to genetic drift during $[0, \epsilon]$. Since genetic drift is the stochastic part of the equation (3.27), we need to use a property called Itô isometry property.

Suppose that $f(t)$ is a random function satisfying:

$$E\left[\left(\int_a^b f(t)dW(t)\right)^2\right] = \int_a^b E\left(f^2(t)\right)dt \tag{3.33}$$

This equation is called Itô isometry property based on ref. Allen (2003). In essence, the Itô isometry property is analogous to the standard isometry property for deterministic integrals, but adjusted for the stochastic nature of Itô calculus. It shows how the variance of the stochastic integral $\int_a^b f(t)dW(t)$ relates to the integral of the square of the function $f(t)$ itself. We want to use this property for genetic drift. As a result we can write:

$$\Delta_N x = E\left[\left(\int_0^\epsilon \sqrt{\frac{x(1-x)}{2N}}dW(t)\right)^2\right]^{\frac{1}{2}} = \left[\left(\int_0^\epsilon E\left(\frac{x(1-x)}{2N}\right)dt\right)\right]^{\frac{1}{2}} \tag{3.34}$$

The reason that we have an expectation in this formula is that we have a stochastic integral so we have a stochastic function when we solve the integral and we need the average typical change of that function and the expectation gives us the average. We use expectation squared to be able to use Itô isomtery property and in next step we take a square root to be able to use standard deviation to adjust the scale which now is a variance in Eq. (3.34). Therefore, Eq. (3.34) is our definition for the change in frequency due to the genetic drift. We solve this equation and since $[0, \epsilon]$ is a small interval, the average change in frequency is

approximately equal to the frequency itself. As a result Eq. (3.34) is approximately:

$$\Delta_N x \approx \left[\frac{x(1-x)}{2N}\epsilon\right]^{(\frac{1}{2})} = \sqrt{\frac{x(1-x)}{2N}\epsilon} \tag{3.35}$$

We rearrange Eq. (3.35) imposing the time interval ϵ to calculate how much frequency relatively changes during this time period due to genetic drift and we have:

$$\frac{\Delta_N x}{x} = \sqrt{\frac{(1-x)}{2Nx}\epsilon} = \sqrt{\frac{\eta_1}{2N\delta x}} \tag{3.36}$$

Now, knowing that $\frac{\Delta_\delta x}{x} = \frac{\Delta_N x}{x} = \eta_1$ we solve for the frequency at which genetic drift and the negative shift are comparable and that frequency is:

$$x = \frac{1}{2N\delta\eta_1} \tag{3.37}$$

Specifically, we are interested in the case where $\eta_1 \sim 1$ representing a small but not infinitesimally small change. If η_1 is infinitesimally small then we are evaluating the influence of drift based on its initial growth rate. Such an approach would lead to a substantial overestimation because unlike the negative shift, the allele frequency changes due to drift change sign. Consequently, the total allele frequency change that accumulates over time is much less than a linear extrapolation of the initial displacement. In fact, the total displacement due to drift only grows as $\sqrt{t}$. Our analysis needs to capture this fact, so we want finite t and hence finite η_1.

Also, we need to make sure that we do not let t get so large that x changes significantly. The primary objective of this assumption is to analyze the influence of drift versus the negative shift on allele frequency dynamics rather than focusing on long-term outcomes, such as fixation. Our approach is to fix the total change in x of interest (i.e. defining η_1 and and selecting an appropriate value). The selected value of η_1 is not precise but rather an order-of-magnitude estimate. We avoid choosing $\eta_1 \ll 1$ to steer clear of infinitesimal

changes and likewise $\eta_1 \gg 1$ is avoided to prevent the analysis from focusing on long-term outcomes. Thus, we aim for $\eta_1 \sim 1$.

It is important to note that $\eta_1 \sim 1$ may not represent a strictly small change when x is large, as in the case where $x = \frac{1}{2}$. In such scenarios, $\eta_1 \sim 1$ would correspond to long-term outcomes. However, the focus of this research is on low frequencies, particularly the dynamics of a newly arising mutation. Such mutations typically start at very low frequencies and may remain at these levels for many generations if they do not extinct. The key question addressed here is what occurs during this time. We do not concern large frequencies because if one force only starts to matter near $x = \frac{1}{2}$, it is essentially irrelevant to the primary inquiry. Therefore, the equation (3.37) can be written as:

$$x = \frac{1}{2N\delta} \tag{3.38}$$

The equation (3.38) is similar to the well known result in constant selection models where δ could be the selection coefficient. Eg. 3.38 reveals that if $x < \frac{1}{2N\delta}$ genetic drift dominates and if $x > \frac{1}{2N\delta}$ the negative shift dominates over drift (Desai and Fisher, 2007a).

3.5.3 Selective Fluctuation vs. Negative Shift

We continue with the final comparison between selective fluctuation (v) and the negative shift in selection coefficient (δ). The partial differential equation (PDE) describing the dynamic between v and δ is:

$$\frac{\partial \phi(x,t)}{\partial t} = -\frac{\partial}{\partial x}[x(1-x)(\delta + 2v\left(\frac{1}{2} - x\right)\phi(x)] + \frac{\partial^2}{\partial x^2}[2vx^2(1-x)^2\phi(x)] \tag{3.39}$$

We assumed the population size is big enough that the genetic drift term $\frac{x(1-x)}{N}$ vanished from stochastic part of the equation, therefore, $a(x) = 2vx^2(1-x)^2$ and $b(x) = \delta + 2v\left(\frac{1}{2} - x\right)$.

Based on Allen (2003), the SDE corresponding to the equation (3.39) will be written as follows:

$$dx(t) = x(1-x)(\delta + 2v\left(\frac{1}{2} - x\right))dt + \sqrt{2vx^2(1-x)^2}dW(t) \tag{3.40}$$

First, we aim to determine the duration required for a small change in frequency due to directional first order derivative term including the negative shift in the selection coefficient, given by

$$\Delta_\delta x = \int_0^\epsilon dx(t)dt = \int_0^\epsilon x(1-x)(\delta + 2v\left(\frac{1}{2} - x\right))dt \tag{3.41}$$

Since we are integrating over a small time interval $[0, \epsilon]$, x does not have time dependence changes which means x is approximately constant with respect to t. As a result we approximate Eq. (3.41) to be

$$\Delta_\delta x = \int_0^\epsilon x(1-x)(\delta + 2v\left(\frac{1}{2} - x\right))dt \approx x(1-x)(\delta + 2v\left(\frac{1}{2} - x\right))\epsilon \tag{3.42}$$

Therefore, the relative change in frequency due to the negative shift is:

$$\frac{\Delta_\delta x}{x} = (1-x)(\delta + 2v\left(\frac{1}{2} - x\right))\epsilon \tag{3.43}$$

We call the small relative change in frequency due to the negative shift to be equal to η_2:

$$\frac{\Delta_\delta x}{x} = \eta_2 \tag{3.44}$$

With the same justification used for η_1, η_2 is also a small change but not infinitesimally

small. Imposing Eq. (3.44) into Eq. (3.43) and solving for ϵ, we have:

$$\epsilon = \frac{\eta_2}{(1-x)(\delta + 2v\left(\frac{1}{2} - x\right))} \tag{3.45}$$

which is the duration required for a small increase in frequency due to the negative shift in the selection coefficient. Next, we assess how much selective fluctuation changes during this time period. Given that we need to consider the directional term of the Eq. (3.40), we use Itô isometry property (3.33) leading to have:

$$\Delta_v x = E\left[\left(\int_0^{\epsilon} \sqrt{2vx^2(1-x)^2}dW(t)\right)^2\right]^{\frac{1}{2}} = \left[\left(\int_0^{\epsilon} E\left(2vx^2(1-x)^2\right)dt\right)\right]^{\frac{1}{2}} \tag{3.46}$$

We solve this integral and since $[0, \epsilon]$ is a small interval, the average change in frequency is approximately equal to the frequency itself. As a result Eq. (3.46) is approximately

$$\Delta_v x \approx \left[2v\epsilon x^2(1-x)^2\right]^{(\frac{1}{2})} = \sqrt{2v\epsilon}x(1-x) \tag{3.47}$$

So, we can define the relative change in frequency due to selective coefficient to be

$$\frac{\Delta_v x}{x} = \sqrt{\frac{2v\eta_2(1-x)}{\delta + 2v\left(\frac{1}{2} - x\right)}} \tag{3.48}$$

Since we know that $\frac{\Delta_\delta x}{x} = \frac{\Delta_v x}{x} = \eta_2$ and with the same justification used for η_1 in previous section, we are interested in the case where $\eta_2 \sim 1$. This yields

$$\frac{\Delta_v x}{x} = \sqrt{\frac{2v(1-x)}{\delta + 2v\left(\frac{1}{2} - x\right)}} \tag{3.49}$$

Solving this equation for x, reveals the relationship between v and δ where x disappears

during calculation:

$$v = \delta \tag{3.50}$$

This result is totally new and remarkable because, unlike the other two comparison results, it is not frequency-dependent. Instead, it establishes a linear relationship between v and δ. when $v < \delta$ the negative shift dominates, and when $v > \delta$ selective fluctuation dominates.

In order to see the effect of the negative shift, we zoomed in a smaller region of Fig. 3.1 where the red line is the Eq. 3.50:

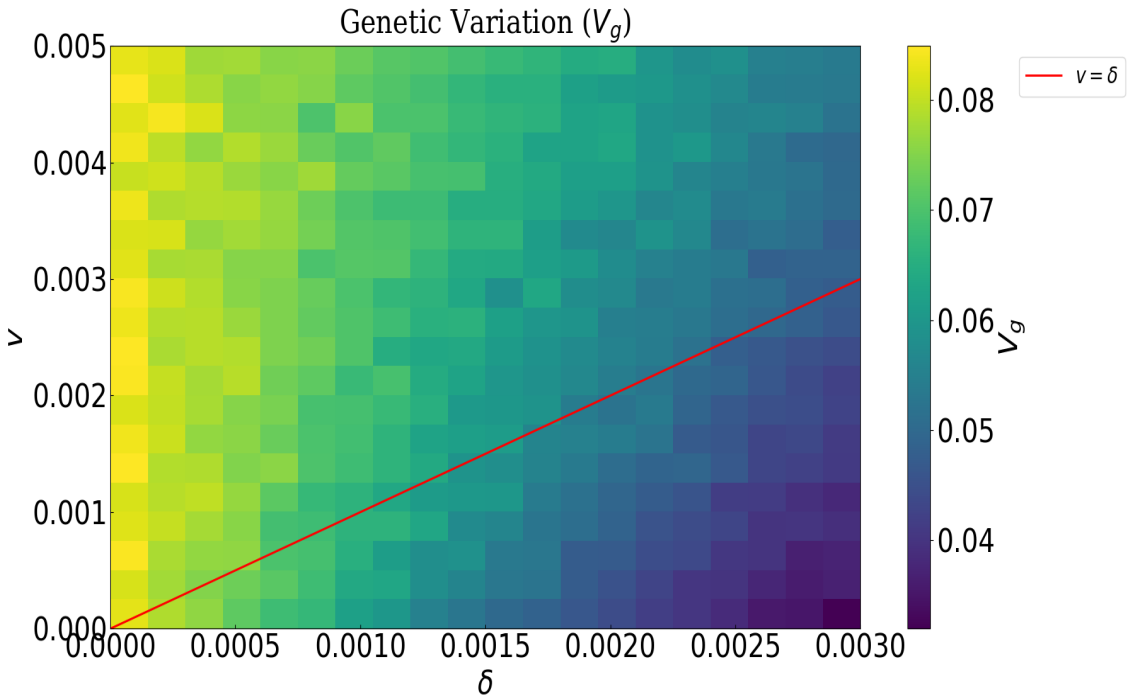

Figure 3.5: This plot zooms in on a specific region of Fig. 3.1. The red line derived in the section 3.5.3 separates regions with lower variation in darker shades from those with higher variation in lighter shades where selective fluctuation and the negative shift are comparable.

The red line with the equation $v = \delta$ split off the dark region from brighter one where the values of the parameters v and δ are equal. This line clearly shows the effect of the

negative shift on genetic variation since there the dark region is below the red line where δ is dominant and in each small interval of δ values (i.e. vertical columns), as we increase v we get more genetic variation.

Also, figure 3.2 shows that this result explains the overall dependency of V_g on δ and v. In other words, fluctuations are able to overcome pressure of negative bias keeping new mutations at low frequencies.

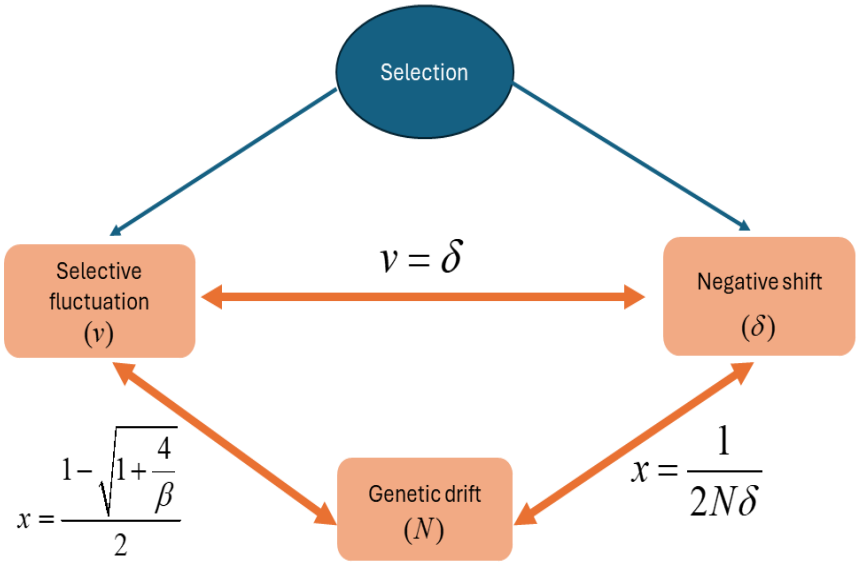

Figure 3.6: In this figure, orange boxes represent the three main forces in our model along with their parameters. The arrows between each pair show the equations describing the parameter values or frequencies at which the influence of those forces on allele frequency change are comparable. Only lower branch of Eq (3.25) included since our focus is on behavior at low frequencies where new mutations appear. We have $\beta = 4vN$ based on the Eq. (3.22).

Chapter 4

4 Conclusion

Previous studies such as Huerta-Sanchez et al. (2008) and Takahata and Kimura (1979) have shown that selective fluctuation as a source of stochasticity tends to reduce genetic variation. However, this study assumes a negative shift, stemming from the natural observation in real world scenarios. We know that most populations are presumably living in a regime where new mutations are on average deleterious (Eyre-Walker and Keightley, 2007). By incorporating the negative shift into the model of selective fluctuation, we find a significant increase in genetic variation, meaning that the influence of selective fluctuation is to increase diversity allowing new mutations to reach higher frequencies. This finding highlights the importance of the negative shift in selective fluctuation models, as it qualitatively changes the model's predictions and indicates that even in the absence of balancing selection, a considerable genetic diversity can arise in a fluctuating environment.

However, it is crucial to note that for this effect to be observable, selective fluctuations must be strong. This implies that the selection coefficient per generation is likely very high, requiring an intense and variable environment to manifest. The logic behind this is that δ has a scale of a typical selection coefficient per generation. On the other hand, selective fluctuation v has a scale of the variance of the selection coefficient per generation based on (2.9). Additionally, we derived the frequency independent equation of (3.50) representing the relationship between v and δ which implies if we want selective fluctuation to matter v has to be bigger than δ. In order to adjust the scale of comparison between v and δ, we use the standard deviation of selection coefficient that is $\sqrt{v}$ and will be much bigger than v given that $0 \leq v \ll 1$. (i.e. $\sqrt{v} \gg v \geq \delta$). This fundamentally explains why y-axis values v in Fig.3.1 has to go to such high values to see the full pattern of genetic variation. To our knowledge this analytical result about the relative importance of selective fluctuations vs the negative shift is completely new.

The requirement for unreasonably large values of v appears to undermine the relevance of fluctuations in our model for real populations. However, the necessity of choosing large values for v arises at least partially from the assumption of white noise. We know that the uncorrelated selective fluctuations over time is unrealistic. Other models that have incorporated autocorrelation in selective fluctuation have shown that the diffusion model derived in such studies has a very similar form. For instance, in the work of Huerta-Sanchez et al. (2008) we see the same diffusion equation that we used, with v replaced by V where the variable V in Eq. (19) is accumulated variance parameter that aims to get the effect of selective fluctuation throughout many generations. Unlike v, the parameter V accounts for the fact that the selective pressures are correlated over generations which gives fluctuation more time to accumulate allele frequency changes in a consistent direction. The result of considering this autocorrelation is that V could in principle be substantially larger than v if we were accounting for autocorrelation. This could adjust our model to reflect real-world scenarios better, thus refining our understanding of genetic variation under fluctuating selection. This is a promising avenue for future work.

Additionally, recent discussions such as Myhre et al. (2016) have highlighted an intriguing aspect of selective fluctuation when negative shift is absent. It has been suggested that selective fluctuation in such a model decreases the effective population size. The parameter representing genetic drift in our model is the population size. While we have incorporated genetic drift by fixing its parameter, future work will involve redefining this parameter in terms of effective population size instead. By doing so, we can explore how genetic diversity behaves under varying effective population sizes in the presence of a negative shift within our selective fluctuation model. This could potentially lead to a more comprehensive analysis, including a 3D heat map that incorporates effective population size as an additional dimension. The effective population size is crucial because it directly influences the strength of genetic drift, the smaller the effective population size, the stronger the genetic drift. By examining different effective population sizes, we can gain insights into how

the interplay between selective fluctuation, the negative shift and genetic drift affects genetic diversity. This approach may reveal new dynamics that are not captured when using a fixed population size, thereby offering a more nuanced understanding of how these factors interact over time.

References

Allen, L. J. S. (2003). *An Introduction to Stochastic Processes with Applications to Biology*. Pearson Education, Upper Saddle River, NJ.

Bertram, J. and Masel, J. (2019). Different mechanisms drive the maintenance of polymorphism at loci subject to strong versus weak fluctuating selection. *Evolution*, 73(5):883–896.

Crow, J. and Kimura, M. (2009). *An Introduction to Population Genetics Theory*. Blackburn Press.

Desai, M. M. and Fisher, D. S. (2007a). Beneficial mutation selection balance and the effect of linkage on positive selection. *Genetics*, 176(3):1759–1798.

Desai, M. M. and Fisher, D. S. (2007b). Beneficial mutation–selection balance and the effect of linkage on positive selection. *Genetics*, 176(3):1759–1798.

Eyre-Walker, A. and Keightley, P. D. (2007). The distribution of fitness effects of new mutations. *Nature Reviews Genetics*, 8(8):610–618.

Gillespie, J. H. (1978). A general model to account for enzyme variation in natural populations. v. the sas-cff model. Theoretical population biology, 14(1):1–45.

Haldane, J. B. and Jayakar, S. D. (1963). Polymorphism due to selection of varying direction. *Journal of Genetics*, 58:237–242.

Hedrick, P. W. (1986). Genetic polymorphism in heterogeneous environments: a decade later. *Annual Review of Ecology and Systematics*, pages 535–566.

Hedrick, P. W. (2006). Genetic polymorphism in heterogeneous environments: the age of genomics. *Annual Review of Ecology and Systematics*, 37(1):67–93.

Huerta-Sanchez, E., Durrett, R., and Bustamante, C. (2008). Population genetics of polymorphism and divergence under fluctuating selection. *Genetics*, 178:325–37.

Karlin, S. and Levikson, B. (1974). Temporal fluctuations in selection intensities: Case of small population size. *Theoretical Population Biology*, 6(3):383–412.

Kimura, M. (1964). Diffusion models in population genetics. *Journal of Applied Probability*, 1(2):177–232.

Melsa, J. L. and Sage, A. P. (2013). *An Introduction to Probability and Stochastic Processes*. Courier Corporation.

Myhre, A. M., Engen, S., and Sæther, B.-E. (2016). Effective size of density-dependent populations in fluctuating environments. *Evolution*, 70(11):2431–2446.

Smỳkal, P. et al. (2014). Pea (Pisum sativum L.) in biology prior and after Mendel's discovery. *Czech Journal of Genetics and Plant Breeding*, 50(2):52–64.

Svardal, H., Rueffler, C., and Hermisson, J. (2015). A general condition for adaptive genetic polymorphism in temporally and spatially heterogeneous environments. *Theoretical Population Biology*, 99:76–97.

Takahata, N., Ishii, K., and Matsuda, H. (1975). Effect of temporal fluctuation of selection coefficient on gene frequency in a population. *Proceedings of the National Academy of Sciences*, 72(11):4541–4545.

Takahata, N. and Kimura, M. (1979). Genetic variability maintained in a finite population under mutation and autocorrelated random fluctuation of selection intensity. *Proceedings of the National Academy of Sciences*, 76(11):5813–5817.

Wittmann, M. J., Bergland, A. O., Feldman, M. W., Schmidt, P. S., and Petrov, D. A. (2017). Seasonally fluctuating selection can maintain polymorphism at many loci via

segregation lift. *Proceedings of the National Academy of Sciences*, 114(46):E9932–E9941.

Yi, X. and Dean, A. M. (2013). Bounded population sizes, fluctuating selection and the tempo and mode of coexistence. *Proceedings of the National Academy of Sciences*, 110(42):16945–16950.

Curriculum Vitae

Name:	Zahra Shafiei
Post-Secondary Education and Degrees:	Iran University of Science and Technology (IUST) Tehran, Iran 2018 - 2022 B.Sc in Mathematics and Applications University of Western Ontario (UWO) London, ON, Canada 2022 - 2024 M.Sc. in Applied Mathematics
Honours and Awards:	Western Graduate Research Scholarship (WGRS) 2022-2024 Selected as a Member of Iran University of Science and Technology Team National Student Scientific Olympiad Spring 2022 and Spring 2021 Ranked 2nd in Cumulative GPA Among Bachelor Students of Mathematics (Top 5 percent) at IUST 2018 – 2021
Related Work Experience:	Teaching Assistant The University of Western Ontario 2022-2024
Leadership roles	Flower Hour Organizer Fall 2023 Direct Reading Program Mentor Fall 2024 Math Challenge Program Assistant Fall 2024 Summer UWO Math Camp Assistant Summer 2023