University of Turin
Dipartment of Mathematics

MSc Thesis in Stochastics and Data Science

Stochastic approximation of normalizing constants in genetic models with selection

Supervisor:
Prof. Matteo Ruggiero

Candidate:
Jacopo Tarantino

Academic Year 2023/2024

I hereby declare that this thesis is my own work and that no other sources have been used except those clearly indicated and referenced

Abstract

The analysis of population genetics and the evolution of allele frequencies over time can be framed as a filtering problem within the Hidden Markov Model (HMM) framework. To find a computable filter, we follow the work of Papaspiliopoulos and Ruggiero (2014), who established duality as a sufficient condition for filtering. We focus on the K-allele model and identify its dual, building on the work of Barbour et al. (2000). Through this duality, the core problem reduces to simulating a birth-and-death process and calculating its transition rates. When selection is introduced into the model, the tractability of these rates diminishes, as they depend on the ratio of multivariate density functions. These densities are the product of a normal distribution and a Dirichlet distribution, both defined over an n-dimensional simplex. We propose various methods to compute these ratios and compare their performances. First, we compute the normalizing constants for the numerator and denominator separately using Monte Carlo integration with importance sampling. We compare our approximations and their computational costs to the analytical method of nested integration proposed by Genz and Joyce (2000). To address the bias introduced by the first approach, we turn to direct approximations of the ratio using Annealed Importance Sampling (AIS) and Linked Importance Sampling (LIS), as described by Neal (2005). Finally, we evaluate all methods based on accuracy and computational time, ultimately defining the optimal approach for the K-allele model.

Contents

List of Tables

List of Figures

1 Introduction

The Wright-Fisher diffusion model is a cornerstone in population genetics, used to describe the evolution of allele frequencies over time in a given population. In its commonest form, the model assumes a population with a finite number of alleles, K , at a genetic locus, which can mutate, drift, or be subject to selection pressures. This gives rise to the K-allele diffusion model, where each allele frequency evolves according to a stochastic process driven by mutation rates and fitness parameters. The interplay between these forces—genetic drift, mutation, and selection—determines the long-term behavior of the population's genetic composition.

In the K-allele model, mutation rates define the likelihood of alleles changing, while fitness parameters influence reproductive success based on allele combinations. A key challenge in studying these dynamics is that the true allele frequencies are often unobservable, making the process effectively a hidden signal. This introduces the need for Hidden Markov Models (HMMs), where the allele frequencies represent the unobserved states, and the available data (e.g., sampled genetic data) form the observations. Our goal is to estimate the hidden distribution of allele frequencies over time, a task that falls under the domain of filtering problems in HMMs.

To solve this filtering problem in a computable and feasible way, we rely on certain structural assumptions outlined by Papaspiliopoulos and Ruggiero [10], which have their linchpin in duality. We explore the K-allele model and its dual counterpart focusing on the studies of Barbour et al [1]. Once the assumptions needed are satisfied, we leverage duality to transform the simulation of a complex diffusion process into a more tractable birth and death process which can be efficiently handled using established algorithms. This transformation provides a way to approximate the underlying dynamics of the allele frequencies by examining the rates of birth and death events. However, a central computational challenge remains since these rates contain the ratios of normalizing constants, which need to be computed for the process to be fully described. The challenge lies in the fact that these normalizing constants are often high-dimensional and not tractable.

The central goal of this thesis is to find efficient methods to approximate these normalizing constants and their ratios. Since direct computation of these constants is computationally prohibitive, we explore two different approaches. The first approach involves computing the numerator and denominator integrals separately. This can be done using two methods: nested integration outlined by Genz and Joyce [4], which breaks down the multi-dimensional integral into a series of one-dimensional integrals that can be solved analytically, and Monte Carlo integration with Importance Sampling, which employs stochastic sampling techniques to approximate the integrals. By carefully choosing the proposal distribution, this method focuses computational effort on the most critical regions of the integrand, trying to improve efficiency and to reduce variance in the estimates.

The second approach focuses on directly approximating the ratio of the normalizing constants, rather than calculating the numerator and denominator independently. This is done using advanced stochastic techniques described by Neal [9], such as Annealed Importance Sampling (AIS) and Linked Importance Sampling (LIS). AIS introduces a sequence of intermediate distributions that gradually transition between the numerator and denominator distributions, allowing for unbiased estimates even when the two distributions are far apart. This method is particularly effective in high-dimensional spaces, where traditional sampling methods may struggle. LIS, on the other hand, maintains the unbiasedness and combines elements of AIS and bridge sampling, linking samples between the distributions of interest to provide in some cases more stable and computationally efficient estimate of the ratio.

Finally, we compare the results obtained from different methodologies, focusing on a specific case of selective over dominance model where mutations are not considered to ensure comparability with

previous studies this. This study aims to develop and evaluate efficient strategies for approximating the normalizing constants in the K-allele diffusion model, providing insights into the computational trade-offs between analytical and stochastic methods in complex population genetic models.

2 Preliminaries

2.1 K-allele diffusion model

In this study, we begin our analysis by adopting the K-allele version of the Wright-Fisher diffusion model as the foundational framework. This model serves as the basis for our analysis and provides the essential structure upon which subsequent theoretical developments and methodologies are constructed. The K-allele model is a fundamental stochastic process in population genetics that describes how allele frequencies evolve over time in a population. This model is applicable when there are K distinct alleles at a genetic locus, where $K \geq 2$ and $K < \infty$.
The population is defined by the vector $x = (x_1, \ldots, x_k)$ and the process operates within the K-1 dimensional simplex defined as

$$\Delta_K = \left\{ \mathbf{x} = (x_1, \ldots, x_K) : x_i \geq 0 \ \forall i = 1, \ldots, K, \sum_{i=1}^{K} x_i = 1, \ K \in \mathbb{N}_{[2,\infty)} \right\} \tag{1}$$

It ensures that allele frequencies remain non-negative and always sum to one, reflecting the biological constraint that allele proportions must form a complete distribution.
The model captures the effects of genetic drift, mutation, and selection on allele frequencies, with these evolutionary forces embedded in the generator of the process, which can be defined as follows (formulation is due to [12]):

$$L = \frac{1}{2} \sum_{i,j=1}^{K} (x_i \delta_{ij} - x_j) \frac{\partial^2}{\partial x_i \partial x_j} + \sum_{i=1}^{K} \left(\sum_{j=1}^{K} \gamma_{ji} x_j + x_i \sum_{j=1}^{K} \sigma_{ij} x_j - \sum_{k,l=1}^{K} \sigma_{kl} x_k x_l \right) \frac{\partial}{\partial x_i} \tag{2}$$

where γ_{ji} for $j \neq i$ is the rate at which allele j mutates to allele i and $(\gamma_{ji})_{ji}$ is the $K \times K$ mutation matrix. This matrix satisfies the law of conservation of probability, constrainting its entries to be non negative for the off-diagonal elements, while the main diagonal $\gamma_{jj} \leq 0$ represents the rate at wich allele j leaves its current state. This ensures that the total probability across all alleles is conserved. On the other hand, $(\sigma_{ji})_{ji}$ is a $K \times K$ symmetric matrix, where σ_{ji} represents the effect of the interaction between allele i and allele j on their fitness. Specifically, $\sigma_{ji} > 0$ for $j \neq i$ represents a fitness advantage for heterozygote individuals in reproduction with both the alleles probably increasing their frequency and hence ensuring the maintenement of a certain genetic balance; the same logic can be applied to the homozygote case σ_{jj} which will instead lead to a genetic dominance. Finally if $\sigma_{ji} = 0 \ \forall i, j$ no allele combination would have a fitness edge, and the model with this specification goes under the name of neutral model. The generator domain is $\mathcal{D}(L) \coloneqq \{f|_{\Delta_k} : f \in C^2(\mathbb{R}^K)\}$ and we also introduce the notation $\partial f|_{\Delta_K}/\partial x_i \coloneqq \partial f/\partial x_i|_{\Delta_K}$.
Next, we turn our attention to the stationary distribution of the system, whose uniqueness and existence is guaranteed by different requirements. First condition outlined by Wright [16] has the following conditions:

$$\gamma_{ji} = \frac{1}{2}\theta_i > 0 \quad \text{with} \quad \theta_i = 4N_e u_i \qquad \forall i \neq j \ \wedge \ i, j \in \{1, \ldots, K\} \tag{3}$$

where θ_i is the scaled mutation rates and it quantifies the effect of mutations u_i (which measures the probability of a mutation occurring at the allele i during a single generation) in relation to the effective population size N_e; in particulat a large value of θ_i implies an high frequency of the allele in the population, thus, even committing an abuse of notation, from now on we willxx refer to θ_i as

alleles frequencies. If (3) holds, the diffusion has a unique stationary distribution Π_σ defined over $\mathcal{P}(\Delta_k)$ given by

$$\Pi_\sigma(dx) = c_\sigma(\theta)^{-1} x^{\theta_1 - 1} \cdots x^{\theta_K - 1} \exp\left(\sum_{i=1}^{K} \sum_{j=1}^{K} \sigma_{ji} x_i x_j \right) dx_1 \cdots dx_{K-1} \tag{4}$$

where $c_\sigma(\theta)$ is the normalizing constant of the distribution. It is significant to observe when considering a neutral model, i.e. deleting the σ_{ji} and the exponential factor, we revert to a Dirichlet distribution with parameters $(\theta_1, \ldots, \theta_K)$. Moreover, still assuming the neutrality of the model, we can define another condition for stationarity discovered by Shiga [13]: the diffusion has a unique stationary distribution $\Pi_\sigma \in \mathcal{P}(\Delta_K)$ if only the mutation matrix $(\gamma_{ji})_{ji}$ is irreducible and the diffusion is reversible with respect to Π_σ if and only if the previous condition (3) holds.
With the instruments and concepts defined thus far, we can proceed to define the transitional density $p(x, t|x_0)$ which measures the probability to go from the state x_0 at time 0 to the state x at time t and it is solution for the Fokker-Planck equation. It can be retrieved as function of the eigenfunctions $\psi(x)$ of the the linear second grade differential operator L described by our generator (for the eigenfunctions and L the following holds $L\psi(x) = \lambda\psi(x)$ where λ is a constant). A suitable system of eingefunctions was found by Shimakura [14] using a biortogonal combination of polynomials $(A_n)_n$ and its adjoint counterpart $(A_n^\star)_n$ with the Appel characterization defined as follows:

$$\frac{\partial A_n(x)}{\partial x} = n A_{n-1}(x) \quad \text{and} \quad A_0(x) \in \mathbb{R}_+ \tag{5}$$

Another expression for the transition function has been later found by Tavarè [15]. Let $\{N_t,\, t \geq 0\}$ be the pure death process in $\mathbb{Z}^+ \cup \{\infty\}$ starting at the entrance boundary ∞ where the states n denotes the number of ancestral lineages left as we go back in time and with death rates

$$q^\circ(\alpha, \alpha - 1) = \frac{1}{2}\alpha(\alpha - 1 + |\theta|) \quad \text{with} \quad |\theta| = \sum_{i=0}^{K} \theta_i \tag{6}$$

which focuses on the two states and the total amount of alleles, and define for $n \geq 0$, $t > 0$

$$d_n^\circ(t) = P(N_t = n) \tag{7}$$

$$= 1 + \sum_{i=1}^{\infty} \rho_i(t) \frac{(2k + |\theta| - 1)(-1)^i \theta_{i-1}}{i!} \quad \text{with} \quad \rho_i(t) = \exp\{-i(i + |\theta| - 1)t/2)\} \tag{8}$$

which is the probability of the process N_t being in the state n at time t. Finally to retrieve the transition function we state the following theorem:

Theorem 1 *If $2 \leq K < \infty$ and $\theta_1 > 0$, $\theta_K > 0$, then the neutral diffusion model in Δ_K with generator L as described above, in which the infinitesimal matrix γ_{ij} satisfies (1.3) and $\sigma_{ij} = 0$ for $i \neq j = 1, \ldots, K$, has transition function $P_t(x, dy)$ given for each $t > 0$ and $x \in \Delta_K$ by*

$$P_t(x, \cdot) = \sum_{n=0}^{\infty} d_n^\circ(t) \sum_{\alpha \in \mathbb{Z}^K : |\alpha| = n} \binom{n}{\alpha} x^\alpha \operatorname{Dir}(\alpha + \theta)(\cdot) \tag{9}$$

This expansion has several important consequences. In particular,

$$d_{TV}\left(P_t(x, \cdot), \operatorname{Dir}(\theta)(\cdot)\right) \leq 1 - d_0^\circ(t) \quad \text{with} \quad t > 0,\, x \in \Delta_K. \tag{10}$$

which combined with another result from Tavarè:

$$1 - d_0^\circ(t) \leq 1 + \theta e^{-\theta t/2} \quad \text{for every } t > 0 \tag{11}$$

it allows us to reach an explicit estimate on the rate of convergence to equilibrium. Therefore it is important to extend the expansion to encompass more general models, particularly those that incorporate selection. Before delving into these models, we would like to highlight an alternative formulation for (9). We consider a K-type pure death process $\{\alpha(t) \in \mathbb{R}^K : \alpha(t) = (\alpha_1, \dots, \alpha_K),\ t \geq 0\}$ starting at infinity, specifically characterized by $|\alpha(0)| = t$. The primary distinction from the model delineated in (6) are the states α: in this context α_i is the relative frequence of the allele i and we are not refering to the lineage anymore. The following equation describes the process deat rates

$$q(\alpha, \alpha - \varepsilon^i) = \frac{1}{2}\alpha_i \left(|\alpha| - 1 + \theta\right) \quad \text{with} \quad \varepsilon^i = \delta_i,\ \alpha \in \mathbb{Z}_+^K,\ 1 \leq i \leq K \tag{12}$$

which depends on the number of α_i people possessing allele i, the population amount $|\alpha|$ and the mutation rates θ. Given $x \in \Delta_K$, the process is uniquely determined if we assume an infinite population at the beginning and we let x describe their initial proportions:

$$|\alpha_t| \to \infty \quad \text{and} \quad \alpha_t/|\alpha_t| \to x \quad \text{as } t \to 0. \tag{13}$$

Then we define the one-to-one map $\rho : \mathbb{Z}_+^K \to \{0\} \cup \mathbb{N} \times \Delta_K$ by

$$\rho(0) = 0, \quad \rho(\alpha) = (|\alpha|, \alpha/|\alpha|) \quad \text{if } \alpha \neq 0, \tag{14}$$

The transition probabilities for the K-type pure death process are denoted with $P_{\alpha\beta}(t)$ and they account for the probability of moving from an initial configuration of alleles frequencies α to a new one β over a time interval t:

$$P_{\alpha\beta}(t) = P^\circ_{|\alpha|\,|\beta|}(t) \frac{\binom{\alpha_1}{\beta_1} \cdots \binom{\alpha_K}{\beta_K}}{\binom{|\alpha|}{|\beta|}}, \quad \alpha \geq \beta \tag{15}$$

They combine transition probabilities of one dimensional death process $P^\circ_{|\alpha||\beta|}(t)$ with rates described in (6) having as object of interest the population magnitude, i.e. $|\alpha|$ and $|\beta|$, with the ratio of binomials for the compositional transitions, resembling a multinomial distribution. To reach the convergence to a stationary distribution we can use transition probabilities to construct the following operator:

$$T_t f(\rho(\alpha)) = \sum_{\beta \in \mathbb{Z}_+^K} f(\rho(\beta)) P_{\alpha\beta}(t), \tag{16}$$

It is easy to show that the operator $\{T_t\}$ is a Feller Semigroup over $C(F)$ with $F = \{\rho(\alpha) : \alpha \in \mathbb{Z}_+^K\} \cup (\{\infty \times \Delta_K\})$ and this would ensure the convergence to a stationary distribution as the one described in (4). We can therefore define the probability to go from the initial state $|\alpha| = \infty$ and the composition x to the state β in time t:

$$d_\beta(t)(x) = \lim_{\rho(\alpha) \to (\infty, x)} P_{\alpha\beta}(t)(x) \overset{15}{=} d^\circ_{|\beta|}(t) \binom{|\beta|}{\beta} x_\beta \tag{17}$$

where again $d^\circ_{|\beta|}(t)$ refers to the death process at a population level. Consequently, we can rewrite the transition function (9) in the compact form

$$P(t, x, \cdot) = \sum_{\alpha \in \mathbb{Z}_+^K} d_\alpha(t, x)\ \mathrm{Dir}(\alpha + \theta)(\cdot) \tag{18}$$

Let us now include haploid selection in the model. We assume that there exists $\sigma = (\sigma_1, \dots, \sigma_K) \in \mathbb{R}^K$ such that

$$\sigma_{ij} = \sigma_i + \sigma_j, \quad i \neq j = 1, \dots, K \tag{19}$$

We then compute the Radon Nikodyn derivative of the stationary distribution $\Pi(\theta)$ with respect to the Dirichlet distribution $\mathrm{Dir}(\theta)$:

$$\frac{d\Pi_\sigma(\theta)}{d\mathrm{Dir}(\theta)}(x) = c_\sigma(\theta)^{-1}\, e^{2\sigma\cdot x} \tag{20}$$

The stationary distribution is absolutely continuous with respect to the Dirichlet Distribution and they differ by an exponential function and a normalizing constant $c_\sigma(\theta)$ which is a normalizing constant depending implicitly on σ, namely,

$$c_\sigma(\theta) = \int_{\Delta_K} e^{2\sigma\cdot y}\ \mathrm{Dir}(\theta)\, dy \in \left[e^{2\min\sigma_i}, e^{2\max\sigma_i}\right] \tag{21}$$

Reasoning in a similar way to what we have done in (18) we can expect a transition form like

$$P(t, x, \cdot) = \sum_{\alpha\in\mathbb{Z}_+^K} b_\alpha(t, x)\, \mathrm{Dir}(\alpha + \theta)(\cdot), \tag{22}$$

where the coefficients $b_\alpha(t, x)$ will be determined in the following section.

2.2 Duality

The process of projecting the operator in a dual space is centered around the following theorem

Theorem 2 *Let L be an operator with domain $\mathcal{D}(L) = C^2(\Delta_K)$ and range in $B(\Delta_K)$ be the generator for a Markov process in Δ_K with Feller transition function $P(t, x, dy)$ and stationary distribution Π with respect to which the process is reversible. Assume that ν charges nonempty open subsets of Δ_K. Define $f_\alpha \in \mathcal{D}(L)$ for each $\alpha \in \mathbb{Z}_+^K$ by $f_\alpha(x) = x^\alpha$ and assume that the matrix $r(\alpha, \beta)$ satisfies the following*

$$Lf_\alpha = \sum_{\beta\in\mathbb{Z}_+^K} r(\alpha, \beta) f_\beta \quad \textit{with} \quad \alpha \in \mathbb{Z}_+^K,\ r(\alpha, \beta) \geq 0\ \forall \alpha \neq \beta,\ r(\alpha, \alpha) \leq 0\ \forall \alpha \tag{23}$$

then define the following

$$m(\alpha) = \int_{\Delta_K} f_\alpha\ d\nu \to 0 = \int_{\Delta_K} Lf_\alpha\ d\nu = \sum_{\beta\in\mathbb{Z}_+^K} r(\alpha, \beta) m(\beta) \tag{24}$$

$$q(\alpha, \beta) = m(\alpha)^{-1} r(\alpha, \beta) m(\beta) \to \sum_{\beta\in\mathbb{Z}_+^K} q(\alpha, \beta) = 0 \tag{25}$$

$$d\nu_\beta = m(\beta)^{-1} f_\beta d\nu \tag{26}$$

Assume that $q(\alpha, \beta)$ is the infinitesimal matrix for a nonexplosive pure jump Markov process $\{\alpha_t\}_{t\geq 0}$ in $\mathbb{Z}_+^K$ with transition probabilities $P_{\alpha\beta}(t)$ and that, for any fixed $\beta \in \mathbb{Z}_+^K$, the function $H : \mathbb{Z}_+^K \to [0, \infty)$ defined as follows

$$H(\alpha) = \sum_{\beta\in\mathbb{Z}_+^K} \left(\delta_{\alpha\beta} + |q(\alpha, \beta)|\right) m(\beta)^{-1} \binom{|\beta|}{\beta} \tag{27}$$

satisfies

$$\{H(\alpha(t\wedge\tau_N)),\ t\in[0,t_0],\ N\geq 1\}\ \textit{is U.I with}\ \tau_N=\inf\{s\geq 0:|\alpha(s)|\geq N\} \tag{28}$$

for each initial state $\alpha\in\mathbb{Z}_+^K$ *and each* $t_0\geq 0$.
Assume that

$$b_\beta(t,y)=\liminf_{\rho(\alpha)\to(\infty,y)} P_{\alpha\beta}(t) \tag{29}$$

defines a probability distribution $b.(t,y)$ *on* $\mathbb{Z}_+^K$ *for each* $t>0$ *and* $y\in\Delta_K$, *and that, for each* $t>0$, *this probability distribution is weakly continuous in* $y\in\Delta_K$. *Then, for each* $t>0$ *and* $x\in\Delta_K$

$$P_t(x,\cdot)=\sum_{\alpha\in\mathbb{Z}_+^K} b_\alpha(t,x)\nu_\alpha(\cdot) \tag{30}$$

The following corollary is useful if the dual process, starting at infinity, absorbs at 0 with probability 1.

Corollary 2.1 *Under the hypothesis of theorem (2) the following holds:*

$$d_{TV}(P(t,x,\cdot),\Pi(\cdot))\leq 1-b_0(t,x) \tag{31}$$

following the theorem we define the Appel polynomials $\{f_\alpha(x)\}\in\mathcal{D}(\alpha)$ and we get

$$\frac{\partial}{\partial x_i}f_\alpha(x)=\alpha_i f_{\alpha-\varepsilon_i}\quad\text{and}\quad\frac{\partial^2}{\partial x_i\partial x_j}f_\alpha(x)=\alpha_i(\alpha_j-\delta_{ij})f_{\alpha-\varepsilon_i-\varepsilon_j} \tag{32}$$

then we substitue f_α into the K-allele diffusion model with haploid selection (19) outlined in (2) and assuming Wright conditions (3) hold, we obtain the following:

$$Lf_\alpha=\frac{1}{2}\sum_{i=1}^K\alpha_i(\alpha_i-1+\theta_i)f_{\alpha-\varepsilon_i}-\sum_{i=1}^K\sigma_i|\alpha|f_{\alpha+\varepsilon_i}-\left\{\frac{1}{2}|\alpha|(|\alpha|-1+|\theta|)-\sum_{i=1}^K\sigma_i\alpha_i\right\}f_\alpha \tag{33}$$

denoting $\sigma_i^-\coloneqq(\max\sigma_j)-\sigma_i$ and using the fact that $\sum_{i=1}^K f_{\alpha+\varepsilon_i}=f_\alpha$ for each $\alpha\in\mathbb{Z}_+^K$, it becomes

$$Lf_\alpha=\frac{1}{2}\sum_{i=1}^K\alpha_i(\alpha_i-1+\theta_i)f_{\alpha-\varepsilon_i}+\sum_{i=1}^K\sigma_i^-|\alpha|f_{\alpha+\varepsilon_i}-\left(\frac{1}{2}|\alpha|(|\alpha|-1+|\theta|)+\sum_{i=1}^K\sigma_i^-\alpha_i\right)f_\alpha \tag{34}$$

and defining

$$r(\alpha,\alpha-\varepsilon_i)=\frac{1}{2}\alpha_i(\alpha_i-1+\theta_i)\quad\text{and}\quad r(\alpha,\alpha+\varepsilon_i)=\sigma_i^-|\alpha| \tag{35}$$

The condition (23) is satisfied. Then we use the function $\gamma(\alpha,\theta)$ to represent the ratio between the normalizing constants of two Dirichlet distributions respectively with parameters α and $\alpha+\theta$:

$$\gamma(\alpha,\theta)=\underbrace{\frac{\Gamma(|\theta|)}{\Gamma(\theta_1),\ldots,\Gamma(\theta_K)}}_{c(\theta)}\underbrace{\left(\frac{\Gamma(|\alpha+\theta|)}{\Gamma(\alpha_1+\theta_1),\ldots,\Gamma(\alpha_K+\theta_1)}\right)^{-1}}_{c(\theta+\alpha)}\to\frac{d\,\mathrm{Dir}(\alpha+\theta)}{d\,\mathrm{Dir}(\theta)}=\gamma(\alpha,\theta)^{-1}f_\alpha \tag{36}$$

To satisfy (24) we revert to the equations (20) and (21) and characterize $m(\alpha)$ as

$$\begin{aligned} m(\alpha) = \int_{\Delta^K} f_\alpha d\Pi(\theta) &= c_\sigma(\theta)^{-1} \int_{\Delta^K} f_\alpha(x) e^{2\sigma\cdot x} \operatorname{Dir}(\theta)\ dx \\ &= \gamma(\alpha,\theta) c_\sigma(\theta)^{-1} \int_{\Delta^K} e^{2\sigma\cdot x} \operatorname{Dir}(\alpha+\theta)\ dx \\ &= \frac{\gamma(\alpha,\theta) c_\sigma(\alpha+\theta)}{c_\sigma(\theta)} \end{aligned} \tag{37}$$

Finally, following (25) we define the transition rates:

$$q(\alpha, \alpha-\epsilon_i) = \frac{1}{2}\alpha_i(|\alpha| - 1 + |\theta|)\frac{c_\sigma(\alpha-\varepsilon_i+\theta)}{c_\sigma(\alpha+\theta)} \tag{38}$$

$$q(\alpha, \alpha+\epsilon_i) = \frac{1}{2}\sigma_i^- |\alpha| \frac{\alpha_i+\theta_i}{|\alpha|+|\theta|}\frac{c_\sigma(\alpha+\varepsilon_i+\theta)}{c_\sigma(\alpha+\theta)} \tag{39}$$

These are the rates for a K-type birth-and-death process $\{X_t, t \geq 0\}$. We also note that deaths occur at a quadratic rate, while births at a linear rate. The generator $L^\#$ of the latter process has form:

$$L^\#\varphi(\alpha) = \sum_{i=1}^K q_\alpha(\alpha-\varepsilon_i)\left[\varphi(\alpha-\varepsilon_i) - \varphi(\alpha)\right] + \sum_{i=1}^K q_\alpha(\alpha+\varepsilon_i)\left[\varphi(\alpha+\varepsilon_i) - \varphi(\alpha)\right] \tag{40}$$

where

$$\varphi(\alpha) = \frac{d\Pi_\sigma(\alpha+\theta)}{d\Pi_\sigma(\theta)} = m(\alpha)^{-1} f_\alpha \in \mathcal{D}(L) \tag{41}$$

Verification of the hypothesis (27) and (29) are more complex and we refer to Barbour [1]. Moduled this, all the assumption and constraints of the theorem hold. Hence we can define the transition function of L and $L^\#$ as follows:

$$P(t,x,\cdot) = \sum_{\alpha\in\mathbb{Z}_+^K} b_\alpha(t,x)\,\Pi_\sigma(\alpha+\theta)(\cdot) \tag{42}$$

where Π is still outlined by (4) and has (21) as normalizing constants.

2.3 Filtering

While Wright-Fisher diffusion model underlying dynamics are well-studied, practical applications often require estimating hidden genetic states, such as allele frequencies, from discrete and noisy observational data. This task falls within the realm of Hidden Markov Models and sequential Bayesian inference.

Hidden Markov Models are widely used as statistical models for time series that have as main target an unobserved signal $\{X_t\}_{t>0}$ with $X_t \in \mathcal{X} \subset \mathbb{R}^K$ $\forall t$ and $K \geq 1$ which is a temporarly evolving parameter with an initial belief distribution ν. What can be detected instead is an observable process collected at discrete times $\{Y_{t_i}\}_{i=1,\dots,n}$ with $Y_{t_i} \in \mathcal{Y} \subset \mathbb{R}^D$ and $D \geq 1$; the observable process is driven by the hidden signal and is indipendent when conditioned to it, i.e given $X_{t_i} = x$ we have $Y_{t_i} \overset{\text{iid}}{\sim} f_x(\cdot)$ and f_x is called emission distribution. In this setting, the main goal is to estimate the trajectory of the signal given observations collected at discrete times $0 = t_0 < \cdots < t_n = T$, which amounts to performing sequential Bayesian inference by computing the so-called filtering distributions $\nu_{i|i-1} \coloneqq p(X_{t_i}|Y_{t_0} = y_{t_0}, \dots, Y_{t_{i-1}} = y_{t_{i-1}})$ from which comes the name of the section.

We can express the filtering distribution through the update operator $\phi_y(\nu)$ defined through the Bayes Theorem:

$$\nu_{i|i-1} = \phi_y(\nu)(x) = \frac{f_x(y)\nu(x)}{f(y)} \quad \text{with} \quad f(y) = \int_{\mathbb{R}^K} f_x(y)\nu(x) \tag{43}$$

Here, we assume all densities of interest exist with respect to an appropriate dominating measure and we define $f(y)$ as the marginal likelihood of a data point y when X_t has distribution ν. To better understand the next computations, we also define the propagation operator ψ_t which describes how the probability density of the signal process evolves over time. Specifically, it maps the probability density at time 0, denoted by ν, to the probability density at time t, based on the transition dynamics of the signal process P_t:

$$\psi_t(\nu)(x) = \int_{\mathbb{R}^K} \nu(x)P_t(x'|x) \tag{44}$$

Filters are usually not so feasible and easy to compute except some specific scenarios characterized by low dimensionality as for example the Kalman-Bucy filter, when both the signal and the observation process are formulated in a gaussian linear system. From these, the definition of finite-dimensional filters, i.e., a sequence of filtering distributions whose explicit identification is obtained through a parameter update based on the collected observations and on the time intervals between the collection times. Outside these classes, explicit solutions are difficult to obtain, and their derivation typically relies on ad hoc computations, but Chaleyat-Maurel and Genon-Catalo [3] comes with a weaker in terms of computational expenses but useful concept of computable filters extending the former class to a larger class of filters whose marginal distributions are finite mixtures of elementary kernels, rather than single kernels. Sufficient conditions for computable filters have been found by Papaspiliopoulos and Ruggiero [10] and can be summarized in the next described three assumptions. The first is reversibility, the signal X_t is reversible with respect to the probability measure $\pi(\cdot)$:

$$\pi(x)P_t(x'|x) = \pi(x')P_t(x|x') \tag{45}$$

Second assumption is conjugacy. Defining $m \in \mathcal{M} \subset \mathbb{R}^K$ a vector of multiplicies obtained by our data (resulting in $m = 0$ at $t = 0$) and $\theta \in \Theta \subset \mathbb{R}^l$ being the vector of hyperparameters; we outline $g(x, \theta, m)\pi(x)$ as the current prior of the signal. Given $f_x(y)$ as the likelihood, we say that $\pi(dx)$ is conjugate with respect to f_x if the posterior $g_y(x, \theta, m)$ belongs to the same distribution family of the prior; i.e. there exist functions $t : \mathcal{M} \times \mathcal{Y} \to \mathcal{M}$ and $T : \mathcal{Y} \times \Theta \to \Theta$ such that:

$$g_y(x, \theta, m)\pi(dx) = g(x, t(y, m), T(y, \theta))\pi(dx) \tag{46}$$

Third assumption is duality and concerns the existence of a certain type of dual process for the signal. We assume that $r : \Theta \to \Theta$ is such that the differential equation

$$\frac{d\Theta_t}{dt} = r(\Theta_t), \quad \Theta_0 = \theta_0 \tag{47}$$

has a unique solution for all θ_0. Let $\lambda : \mathbb{Z}^+ \to \mathbb{R}^+$ be an increasing function, $\rho : \Theta \to \mathbb{R}^+$ be a continuous function, and consider a two-component Markov process (M_t, Θ_t) with state-space $M \times \Theta$, where Θ_t evolves autonomously according to (47), and when at $(M_t, \Theta_t) = (m, \theta)$, the process jumps down to state $(m - e_j, \theta)$ with instantaneous rate

$$\lambda(|m|)\rho(\theta)m_j. \tag{48}$$

We assume (M_t, Θ_t) is dual to X_t with respect to the family of functions g defined in the previous assumption, in the sense that

$$\mathbb{E}_x[g(X_t, m, \theta)] = \mathbb{E}_{(m,\theta)}[g(x, M_t, \Theta_t)] \quad \forall x \in X,\, m \in M,\, \theta \in \Theta,\, t \geq 0. \tag{49}$$

Then we can state the following result (proposition 2.2 in [10]) which has a pivotal role in our approach:

Proposition 1 *Let the assumptions defined above hold and $\psi_t(x)$ be the propogation operator defined in (44)*

$$\psi_t(g_y(x, \theta, m)\pi(x)) = \sum_{0 \leq i \leq m} p_{m,m-i}(t, \theta) g(x, m - i, \Theta_t)\pi(dx) \tag{50}$$

where $p_{m,m-i}$ are the transition probabilities for the process M_t (Proposition 2.1 [10]) and Θ_t is the value in t of (47)

Moving to our specification, the signal process $\{X_t\}$ corresponds to the underlying allele frequencies $(\theta_i, \ldots, \theta_K)$ in the population that evolve over time according to the Wright-Fisher diffusion defined by the generator (2) and its stationary distribution $\pi(x)$ is described by $\Pi_\sigma(\theta)$ in (4) satisfying the first assumption. The observable process A_t refers to data collected from the population at discrete time intervals $(\alpha_{t_1}, \ldots, \alpha_{t_n})$ with $(\alpha_{t_i} = \alpha_{t_{i,1}}, \ldots \alpha_{t_{i,K}})$ and their distribution f_x representation a multinomial kernel as in (32). With these specifications for both the prior and the emission distribution we satisfy the second assumption since our posterior would be $\Pi_\sigma(\theta + \alpha)$ (to show this recall that α and θ does not affect the exponential term of Π_σ). Then we can define the $g(x, \alpha) \coloneqq \varphi(\alpha)$ (we do not need to specify Θ in our model) with φ expressed in (42) and generator (40) whose rates (39) are the one of B&D process $\{N_t\}$. From the generator definition we can state the following:

$$\mathbb{E}_x\left[g(X_t, \alpha)\right] = \mathbb{E}_\alpha\left[g(x, N_t)\right] \tag{51}$$

and with that we complete the assumptions that have to be satisfied.
Then we set the initial distribution $\nu_0 = \Pi_\sigma(\theta)$ and we update it with multiplicities α_0 observed at time 0 obtaining:

$$\nu_{0|0} = \phi_{\alpha_0}(\Pi_\sigma(\theta)) = \Pi_\sigma(\alpha_0 + \theta) \tag{52}$$

Next applying the proposition we can define the first predictive $\nu_{1|0}$ as:

$$\begin{aligned} \nu_{1|0}(x) &= \sum_{\beta \in \mathbb{Z}_+} p_{\alpha,\beta} g(x, \beta) \Pi_\sigma(\theta, x) && (53) \\ &= \sum_{\beta \in \mathbb{Z}_+} p_{\alpha,\beta} \varphi(\beta) \Pi_\sigma(\theta, x) \\ &\overset{39}{=} \sum_{\beta \in \mathbb{Z}_+} p_{\alpha,\beta} \Pi_\sigma(\theta + \beta, x) \end{aligned}$$

(54)

Here $p_{\alpha,\beta}$ are the transition probabilities of $\{N_t\}$ which are given by the rates $q(\alpha, \alpha - \epsilon_i)$ and $q(\alpha, \alpha + \epsilon_i)$ in (39). In this step we can finally see the arise of our computational problem of simulating the B&D process and compute the normalizing constants $c_\sigma(\alpha + \theta)$ and $c_\sigma(\theta)$ or directly

their ratio their as cardinal to solve filtering problems. Then, updating again with $\alpha_1 \in \mathbb{Z}_+^K$ observed at time 1 yields:

$$\nu_{1|0:1}(dx) = \sum_{\beta \in \mathbb{Z}_K^+} \frac{p_{\alpha,\beta} p_{\theta+\beta}(\alpha_1)}{\sum_k p_{\alpha,k} p_{\theta+k}(\alpha_1)} \Pi_\sigma(\theta + \beta + \alpha_1, x) \tag{55}$$

If we are able to simulate rates, the basic strategy suggested in [6] [7] would then be roughly the following: the first update with α_0 gives a singleton, then we simulate I instances of $N(\Delta)$ starting from $N_0 = \alpha_0$. Next, we use the empirical frequency $\hat{\pi}_\beta = \frac{1}{I} \sum_{i=1}^{I} \mathbf{1}(N^{(i)}(\Delta) = \beta)$ to approximate $p_{\alpha,\beta}$:

$$\nu_{1|0:1} \approx \sum_{\beta \in \mathbb{Z}_K^+} \hat{\pi}_\beta \Pi_\sigma(\theta + \beta) \tag{56}$$

and finally upadate α_1 to get (55)

2.4 Simulation Issues

To simulate the K-alleles genetic model (2) introduced at the beginning, we deploy its dual representation as a birth-and-death (B&D) process (40). This equivalence allows us to simulate the K-alleles model by effectively modeling the corresponding B&D process. Among the various techniques available for this purpose, the Gillespie algorithm (1) stands out as one of the most popular and effective methods. Specifically designed for exact stochastic simulation of continuous-time Markov processes, the Gillespie algorithm ensures that the dynamics of the system are accurately captured. To keep track of how fast the system is evolving, the algorithm computes the overall rates of birth and death and sums them together. It draws the time difference between two consecutive events from an exponential of the inverse of the sum of the rates, so that higher the rates smaller the time difference will be. Then it chooses death or birth based on their overall relative frequencies draws an allele using its rate distribution and increases or decreases by 1 its cardinality. A critical aspect of the simulation is the computation of the birth-to-death ratio, which is essential for accurately modeling the process. One key challenge in this computation arises from the form of the birth and death rates, as given in (39). Calculating these rates requires computing the normalizing constants $c_\sigma(\alpha + \theta)$ and $c_\sigma(\alpha + \theta \pm \varepsilon_i)$. Assigning a value to this integral is nontrivial. The main objective of the following sections is not only to compute it but to do so in a computationally efficient manner, as the continuous computation and updating of rates is necessary for applying the Gillespie algorithm or any other B&D process simulation technique.

Algorithm 1: Gillespie Algorithm for Birth-Death Process

Input: Initial population vector $\alpha^0 = [\alpha_1^0, \alpha_2^0, \ldots, \alpha_K^0]$
maximum simulation time $t_{\max}$

Output: $|\alpha|$

Initialize total population size pop_size $= |\alpha^0| = \sum \alpha_i$;
while $t < t_{max}$ ***and*** *pop_size* > 0 **do**
  $R_{\text{death}} \overset{39}{=} \sum_{i=1}^{K} q(\alpha, \alpha - \epsilon_i)$;
  $R_{\text{birth}} \overset{39}{=} \sum_{i=1}^{K} q(\alpha, \alpha + \epsilon_i)$;
  $R_{\text{total}} = R_{\text{death}} + R_{\text{birth}}$;
  $\Delta t \sim \text{Exp}(1/R_{\text{total}})$;
  $t = t + \Delta t$;
  $r = \text{Unif}(0, R_{\text{total}})$;
  if $r < R_{death}$ **then**
    draw $i : \mathbb{P}(i = j) = q(\alpha, \alpha - \epsilon_j)/R_{\text{death}}$;
    $\alpha_i = \alpha_i - 1$;
  else
    draw $i : \mathbb{P}(i = j) = q(\alpha, \alpha + \epsilon_j)/R_{\text{birth}}$;
    $\alpha_i = \alpha_i + 1$;
  end
  pop_size $= \sum \alpha_i$;
end

3 Normalizing Constant Approximation

In this section we provide two different techniques to estimate the normalizing constant described in (59). The first approach follows the strategy outlined by Genz and Joyce [4], which focuses on reducing the multidimensional integral into a series of one-dimensional integrals, each of which can then be computed using analytical methods. The second method is more stochastic in nature, utilizing Monte Carlo integration through importance sampling to approximate the value of the normalizing constant. Before delving deeper into the analysis, we introduce an additional assumption from Genz and Joyce's work, as we are going to compare our results with theirs. To ensure a genetic stability we consider the selective over dominance model in which heterozygotes have a selective advantage over homozygotes. The probability of choosing a homozygote at random from a population, called homozygosity, is given by

$$F = \sum_{i=1}^{K} x_i^2 \tag{57}$$

To match this concept with our notation expressed in (2) we transform the fitness matrix $(\sigma_{ji})_{ji}$ in a diagonal uniform matrix:

$$\sigma_{ii} = \sigma \in \mathbb{R} \ \forall i = 1, \dots, K \quad \text{and} \quad \sigma_{ij} = 0 \ \forall i \neq j \tag{58}$$

By plugging this result into the stationary distribution (4), the integral of interest, which we will focus on in the subsequent sections, becomes:

$$\begin{aligned} c_\sigma(\alpha) &= \int_{\Delta_K} e^{\sum_{j=1}^{K} \sigma_{ji} x_i x_j} \prod_{i=1}^{n} x_i^{\alpha_i - 1} dx_1, \dots, dx_{K-1} \\ &\overset{58}{=} \int_{\Delta_K} x_1^{\alpha_1} \dots x_{K-1}^{\alpha_{n-1}} \left(1 - \sum_{j=1}^{K-1} x_j\right)^{\alpha_n} e^{-\sigma\left(x_1^2 + \dots + x_{K-1}^2 + \left(1 - \sum_{j=1}^{K-1} x_j\right)^2\right)} dx_1, \dots, dx_{K-1} \end{aligned} \tag{59}$$

where the term $\exp\left(-\sigma(x_1^2 + \dots + x_n^2)\right)$ introduces a quadratic penalty on the allele frequencies x_i and $\sigma > 0$ represents the strength of selection. In the overdominance model, higher values of x_i , which correspond to greate homozygosity, are penalized. The sum $x_1^2 + \dots + x_n^2$ is closely related to the homozygosity (57). Thus, this term favors configurations with lower homozygosity (more heterozygotes) by reducing the stationary probability of states with large x_i^2.

3.1 Nested analytical integration

This strategy involves expressing the normalizing constant as a sequence of one-dimensional integrals and then, based on their order, we apply different analytical techniques to efficiently compute each of them. As first step we rewrite $c(\alpha)$ as follows:

$$\begin{aligned} c_\sigma(\alpha) = \int_0^1 x_1^{\alpha_1} \int_0^{1-x_1} x_2^{\alpha_2} \ \cdots \int_0^{1-\sum_{i=1}^{K-3} x_i} x_{K-2}^{\alpha_{K-2}} \int_0^{1-\sum_{i=1}^{K-2} x_i} x_{K-1}^{\alpha_{K-1}} \left(1 - \sum_{i=1}^{K-1} x_i\right)^{\alpha_K} \\ \exp\left(-\sigma\left(\sum_{i=1}^{K-1} x_i^2 + (1 - \sum_{i=1}^{K-1} x_i)^2\right)\right) \ dx_{K-1} \cdots dx_1 \end{aligned} \tag{60}$$

Next, we separate the innermost exponential term into $K-1$ factors, each pair with its corresponding x_i, and distribute them across the respective outer integrals:

$$c_\sigma(\alpha) = \int_0^1 x_1^{\alpha_1} e^{-\sigma x_1^2} \int_0^{1-x_1} x_2^{\alpha_2} e^{-\sigma x_2^2} \cdots \int_0^{1-\sum_{i=1}^{K-3} x_i} x_{K-2}^{\alpha_{K-2}} e^{-\sigma x_{K-2}^2} \int_0^{1-\sum_{i=1}^{K-2} x_i} x_{K-1}^{\alpha_{K-1}} (1 - \sum_{i=1}^{K-1} x_i)^{\alpha_K} e^{-\sigma\left(x_{K-1}^2 + (1\sum_{i=1}^{K-1} x_i)^2\right)} \, dx_{K-1} \cdots dx_1 \quad (61)$$

After this rearrangement, we shift the focus to the innermost integral. Defining $y_K = 1 - \sum_{i=1}^{K-1}$ and $t_K = x_K$ we can rewrite it as follows:

$$g\left(y_{K-1}\right) = \int_0^{y_{K-1}} \underbrace{t_{K-1}^{\alpha_{k-1}} (y_{K-1} - t_{K-1})^{\alpha_k} e^{-\sigma\left(t_{K-1}^2 + (y_{K-1} - t_{K-1})^2\right)}}_{f(t_{K-1})} \, dt_{K-1} \quad (62)$$

finally we recursively define the outer integrals as functions of the inner ones:

$$g(y_i) = \int_0^{y_i} \underbrace{t_i^{\alpha_i} e^{-\sigma t_i^2} g(i+1)}_{f(t_i)} \, dt_i \quad \text{with} \quad i = K-2, \ldots, 1 \quad (63)$$

and we can rewrite the normalizing constant as:

$$c_\sigma(\alpha) = g(1) \quad \text{with} \quad y_1 = 1 \quad (64)$$

We have now redefined our object of interest as $g(1)$, and we will approximate it by recursively approximating the sequence of one-dimensional integrals that constitute it.
We begin by approximating the innermost integral, $g(y_{K-1})$, using a 3-point Gauss-Legendre quadrature:

$$\int_{-1}^{1} f(x) \, dx \approx \sum_{i=1}^{n} w_i f(x_i) \quad (65)$$

Here $x_i \in [-1, 1]$ are the quadrature points where we evaluate the function, and w_i are the corresponding weights. For the 3-point Gauss-Legendre quadrature, the points and weights are:

$$(x_1, w_1) = (0, 8/9), \quad (x_2, w_2) = (\sqrt{3/5}, 5/9), \quad (x_3, w_3) = (-\sqrt{3/5}, 5/9) \quad (66)$$

To adapt the integration interval to $[0, y_{K-1}]$, we use the transformation:

$$\int_a^b f(z) \, dz = \int_{-1}^{1} f\left(\frac{b-a}{2} x + \frac{a+b}{2}\right) \frac{dz}{dx} dx \approx \frac{b-a}{2} \sum_{i=1}^{n} w_i f\left(\frac{b-a}{2} x_i + \frac{a+b}{2}\right) \quad (67)$$

Thus, the final result for $g(y_{K-1})$ becomes:

$$g(y_{K-1}) = \int_0^{y_{K-1}} f(t_{K-1}) \approx \frac{y_{K-1}}{2} \sum_{i=1}^{3} w_i f\left(\frac{y_{K-1}}{2} (x_i + 1)\right) \quad (68)$$

To compute the $g(y_i)$ with $i \neq K-1$ we employ the trapezoid rule. Letting $\{x_k\}_{k \in \mathbb{N}_{1,m}}$ a regular spaced partition of $[0, y_i]$ such that $0 = x_1 < \cdots < x_m = y_i$ with constant mesh Δx we approximate the outer integrals as:

$$g(y_i) = \int_0^{y_i} f(t_i) \, dt_i \approx \frac{\Delta x}{2} \sum_{k=1}^{m} f(x_{K-1}) + f(x_K) \quad (69)$$

It is clear that the accuracy of the nested integration is significantly influenced by the finesse m of the partition $\{x_i\}$. Increasing m improves the precision of our approximation but comes at a cost. Specifically, each increment in m leads to an exponential growth in computational complexity, as the number of hypercubes of the grid over which we are integrating expands across the $K-1$ dimensional space. Each evaluation of $g(y_i)$ has an asymptotic cost of $\mathcal{O}(m2^i)$, and since we will reach the maximum depth of $K-1$ the total computational complexity becomes $\mathcal{O}(m2^{K-1})$. Genz and Joyce provided different examples using different values of α and dimensions, but we focus only on the positive cases as they represent frequencies in our model. They picked 4 uniform vectors with $\alpha = 0.6$ with dimension 5 or 25 and paired with a σ of 10 or 100. Their approximation are summarized in the following tables:

m	$\sigma = 10$	$\sigma = 100$
16	7.20056e-06	6.15340e-15
32	7.23657e-06	6.12232e-15
64	7.24560e-06	6.12238e-15
128	7.24785e-06	6.12239e-15
256	7.24841e-06	6.12239e-15
512	7.24855e-06	6.12239e-15
1024	7.24858e-06	6.12239e-15

Table 1: Nested Integral Approximations for $\alpha = (0.6, \ldots, 0.6)$ and K = 5

m	$\sigma = 10$	$\sigma = 100$
16	11.46322e-49	10.41351e-51
32	1.86022e-48	6.16118e-51
64	1.51656e-48	6.29131e-51
128	1.53391e-48	6.52903e-51
256	1.55299e-48	6.62218e-51
512	1.56002e-48	6.65074e-51
1024	1.56213e-48	6.65872e-51

Table 2: Nested Integral Approximations for $\alpha = (0.6, \ldots, 0.6)$ K = 25

From these results, we observe that as the dimensionality K increases, convergence slows down, though some accuracy is still retained. The exponential increase in computational complexity translates directly into computational time. For example, in the second case discussed by the two authors, where $K - 1 = 24$, the estimated computational cost is $\mathcal{O}(m \times 2^{24}) \approx \mathcal{O}(m \times 10^8)$. This computation not only exceeds our available resources, requiring few hours for $m = 16$, but also fails to meet the time efficiency constraints necessary for simulating our model due to this approach requiring both the numerator and the denominator of the target ratio to be computed. Consequently, we shift our focus to the case where $K = 5$. We begin by analyzing the computational time as a function of the partition size, obtaining figure (1). There, we can easily notice how the exponential nature is correctly reproduced. Moreover, looking at the plot and searching for a feasible proposal for m in real calculations we would suggest $m = 128$.

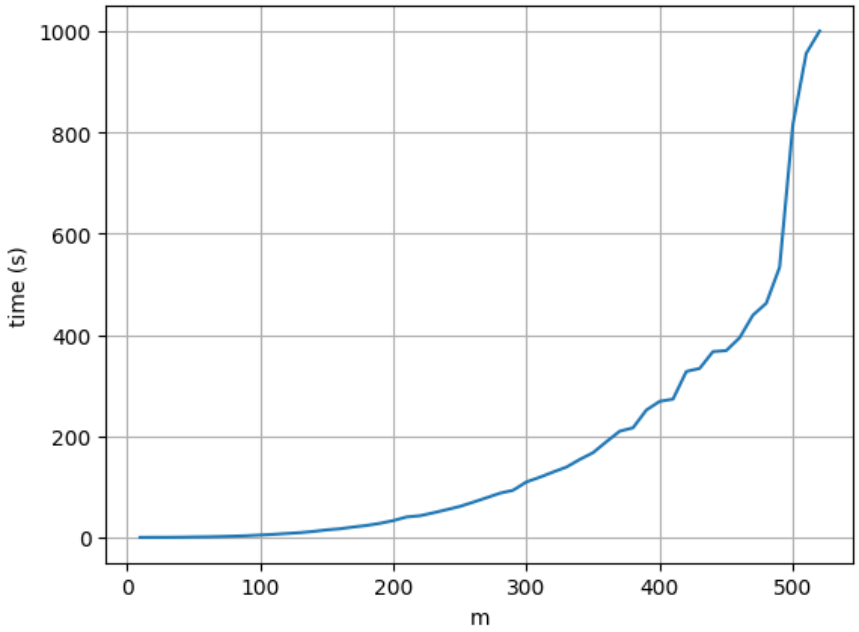

Figure 1: Evolution of computational time as function of partition cardinality for nested analytical integration with $K = 5$

However, we must still evaluate the quality of our approximation. To assess this, we look for the convergence between our computed integral approximations and the reference results in tables (1) and (2) plotting the difference as function of the partion of the grid.

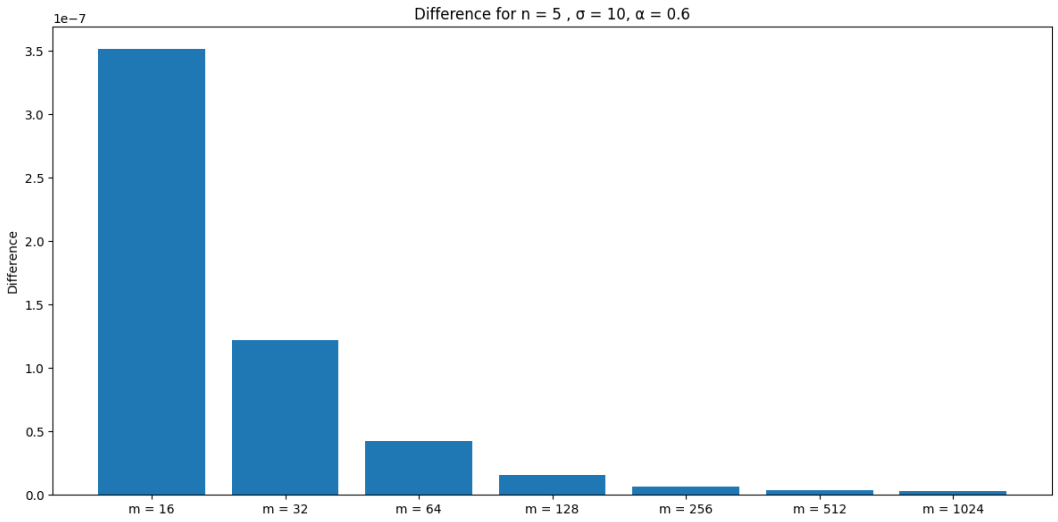

Figure 2: Convergence of the computed nested integral to Genz and joyce results for k = 5 and $\sigma = 10$

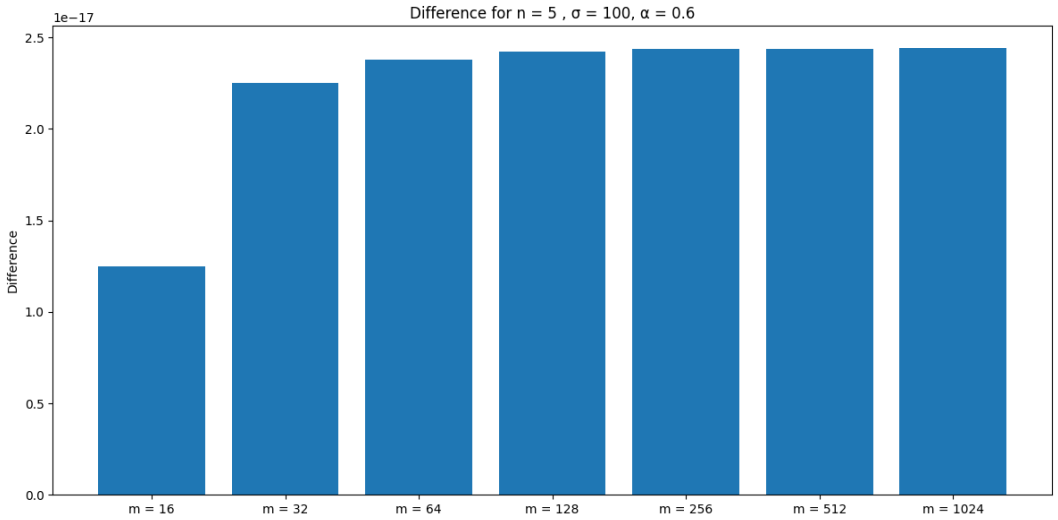

Figure 3: Convergence of the computed nested integral to Genz and joyce results for k = 5 and $\sigma = 100$

For $\sigma = 10$, we observe that our computed integral converges, with the difference gradually decreasing to 0. In contrast, for $\sigma = 100$, the difference between our approximation and the reference values increases as the grid refinement grows, eventually stabilizing. This behavior suggests a potential issue with the covergence value of table (1), which we will investigate further when applying Monte Carlo integration to approximate the integral. Finally, we can conclude stating that nested integration is a high valuable approach, which has in its stability in the convergence its main strenght, but it meets as major drawback in the computational time. Its level of computational cost makes it unsuited to simulate processes where time constraints are a core part for the simualtion effinciency.

3.2 Monte Carlo Integration through importance sampling

Unlike the analytical approach, Monte Carlo integration uses stochastic techniques to estimate the normalizing constant by directly addressing the full-dimensional integral. Instead of reducing the problem to a series of one-dimensional integrals, this method samples points from the probability distribution, whose normalizing constant we want to compute, and uses these samples to approximate the integral. Moreover, in this study, we employ a more sophisticated variant of MC integration, pairing it with importance sampling. This addition greatly enhances efficiency by concentrating computational effort on regions of the integrand that contribute the most to the integral's value. By doing so, we reduce the variance of the estimates and achieve greater accuracy with fewer samples.
In MC importance sampling, the objective is to compute a multidimensional integral of the form

$$I = \int_{\Omega} f(x)dx \tag{70}$$

where $\Omega \subset \mathbb{R}^k$ is the region of integration, and $f(x)$ is the function we wish to integrate. When the dimensionality of the integral is high, directly evaluating this expression can be computationally prohibitive because large portions of Ω contribute very little to the overall integral. To address this issue, we introduce a proposal distribution $q(x)$ which has to be nonnegative and to have a large enough support completely covering the one of the objective function. Its main task is to simplify

the sampling process by concentrating computational effort on regions where the integrand $f(x)$ is large, effectively "reweighting" the space. This is crucial because uniform sampling over the domain would result in wasted effort in regions that contribute little to the integral. Moreover, the proposal distribution $q(x)$ allows us to rewrite the original integral as:

$$I = \int_\Omega \underbrace{\frac{f(x)}{q(x)}}_{w(x)} q(x)dx \tag{71}$$

Here, we successfully transform our original problem into a new one where we can sample directly from $q(x)$. In this new form, the expression $f(x)/q(x)$ is the Radon-Nikodyn derivative and it behaves as correcting factor or "weight" for the change of draw function. This change in formulation let us to interpret the object integral I as the expectation of the derivative with rispect to the proposal distribution:

$$\mathbb{E}_{q(x)}\left[\frac{f(x)}{q(x)}\right] = \int_\Omega \frac{f(x)}{q(x)} q(x)dx = I \tag{72}$$

To estimate the expectation, and hence the integral, we draw n independent samples $x_1, \ldots, x_n$ from $q(x)$ and compute the sample mean of the weighted function $f(x_i)/q(x_i)$. This gives the Monte Carlo estimate with importance sampling:

$$\hat{I} = \frac{1}{n}\sum_{i=1}^{n} \frac{f(x_i)}{q(x_i)} \tag{73}$$

As the number of samples n increases, the law of large numbers guarantees that this estimate converges to the true value of the integral:

$$\lim_{n\to\infty} \hat{I} = I \tag{74}$$

Naturally, the more samples we take from $q(x)$, the more accurate our estimate becomes. The major streght of MC integration lies in its computational costs. Assuming that both drawing samples from the proposal distribution and evaluating the objective function have a fixed cost of c, the overall asymptotic cost for generating samples, computing densities, and averaging them scales as $\mathcal{O}(n \times K \times c)$, where n is the number of samples and K is the dimensionality of the problem. This formulation highlights a key advantage of MC integration compared to analytical methods: while the computational cost of analytical approaches typically grows exponentially with K, in MC integration, K only contributes linearly to the cost. Furthermore, the addition of importance sampling enhances the efficiency of the method. By focusing computations on the most relevant regions of the integrand, importance sampling allows us to achieve accurate results with fewer samples, effectively reducing n and the overall computational burden. Given these considerations, it is clear that our proposed Monte Carlo integration, particularly when combined with importance sampling, is well-suited for handling high-dimensional integrals efficiently.
To adapt this formulation to our problem, we assign $\Omega = \Delta_K$ and we pick as objective function the integrand of (59) so that $f : \Omega \to \mathbb{R}$, i.e.:

$$f(x) = x_1^{\alpha_1} \ldots x_{K-1}^{\alpha_{n-1}} \left(1 - \sum_{j=1}^{K-1} x_j\right)^{\alpha_n} e^{-\sigma\left(x_1^2 + \cdots + x_{K-1}^2 + \left(1 - \sum_{j=1}^{K-1} x_j\right)^2\right)} \tag{75}$$

To fully leverage the potential of importance sampling, it is essential to carefully select the proposal distribution $q(x)$ so to eliminate any oppurtunity leading to bias. A formula to find the optimal

proposal distribution exists $q^\star(x)$ and it was recently outlined in 2011 by Rubinstein [11]:

$$g^\star(x) = \frac{|x|f(x)}{\underbrace{\int |z|f(z)\,dz}_{c(\alpha)}} \tag{76}$$

However, the optimal proposal distribution depends on the normalizing constant of the objective function, $c(\alpha)$, which is precisely the quantity we aim to compute. This makes the optimal proposal infeasible. Therefore, to identify a practical and computationally feasible proposal, we need to experiment with various functional forms and parameter settings. Our approach in this research is based on two key criteria: the accuracy of our approximation and the stability of our results. For the accuracy, we compare our estimates with the best available approximations, i.e. the Genz and Joyce results (shown in Tables (1) and (2)) for $m = 1024$, and evaluate whether our method converges to these values. For stability, we rely on two performance measures:

$$\mathrm{Var}(\tilde{w}) = \frac{1}{N}\left(\sum_{i=1}^{N} \tilde{w}_i - \frac{1}{N}\right) \quad \text{with} \quad \tilde{w}(x_i) = \frac{w(x_i)}{\sum_{i=1}^{n} w(x_i)} \tag{77}$$

$$\mathrm{ESS} = \frac{1}{N}\frac{\left(\sum_{i=1}^{N} \tilde{w}_i\right)^2}{\sum_{i=1}^{N} {\tilde{w}_i}^2} \tag{78}$$

The first measure, the variance of the normalized weights $\tilde{w}(x_i)$, assesses how well the proposal distribution matches the target one. A high variance indicates a poor match and suggests that our choice of proposal distribution may introduce a significant bias. The second measure is the Effective Sample Size (ESS), which measures the proportion of the samples drawn that effectively contribute to the estimation.

We begin our decision process from the functional form of $q(x)$. Given that a component of the objective function (75) is the Dirichlet distribution, we opt for this family as our proposal distribution and from now on we are going to focus on the choice of the parameter vector α^q we use as input for the proposal. Before testing a range of different parameters we state three constraint which we impose onto α^q: it has to be nonnegative due the heuristic meaning of its components in our model, i.e. alleles frequencies; it has to mantein the same proportions of its objective counterpart among its components, i.e. in this setting $\alpha_i^q = \alpha_j^q \ \forall i, j$; and since the other exponential component of the objective function tends to push the distribution towards the center of the simple we would like to sample points from the inner regions too and so we assume $\alpha_i^q \geq 1 \ \forall i$

To have a better visualization of the reasons behind this constraints, without any loss of generality, we reduce dimensions to $K = 3$ and plot the first two dimensions of the objective and proposal distribution for different parameters vectors over the two dimensional simplex:

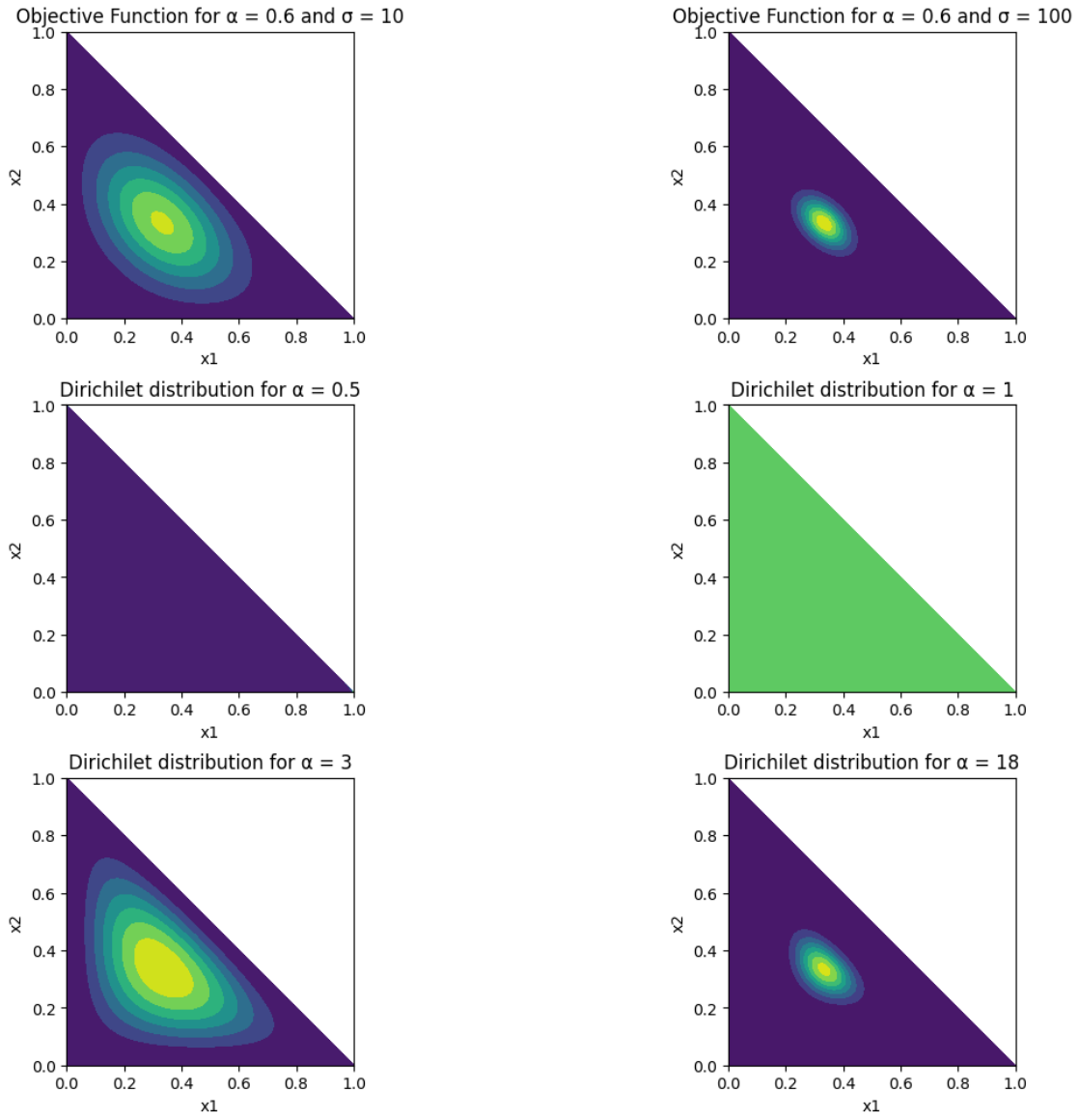

Figure 4: Comparison between objective function and Dirichlet distribution probability densities for different values of K and α

To select the optimal parameter vector, we systematically tune various candidate suggestions, which are functions of both the dimensionality K and the value of σ since the two factors affect the volume of the objective function, generally making it smaller as either increases. For each proposed parameter vector, we compute integral approximations using different sample sizes. Along with the latter, we record the three evaluation metrics: the deviation from the reference table values, the variance of the normalized weights, and the Effective Sample Size (ESS). Finally, for each combination of $\sigma = 10, 100$ and $K = 5, 25$, we plot these metrics as functions of the number of samples, obtaining the following results:

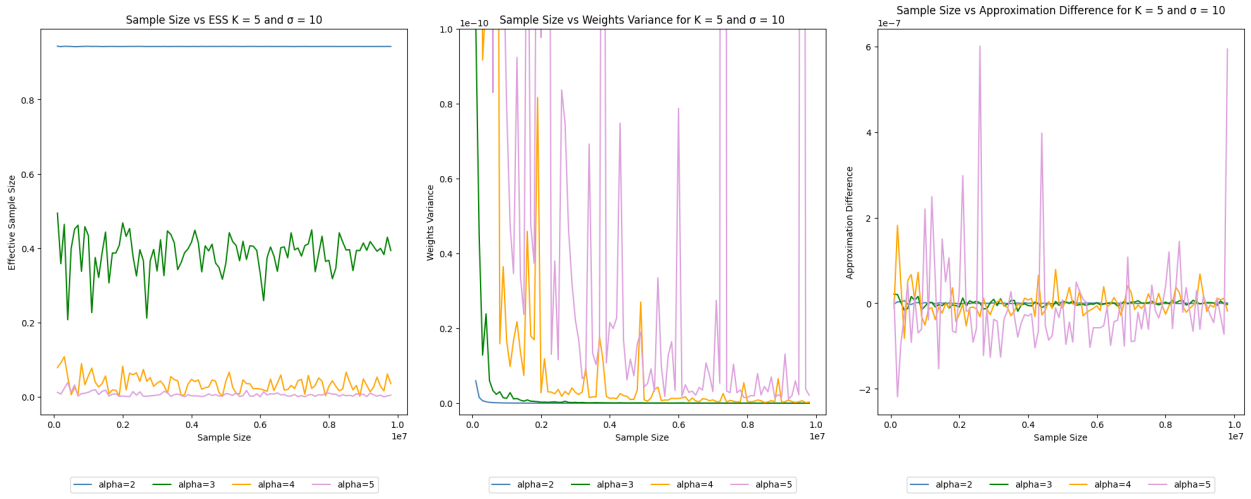

Figure 5: Tuning for proposal distribution with K = 5 and $\sigma = 10$

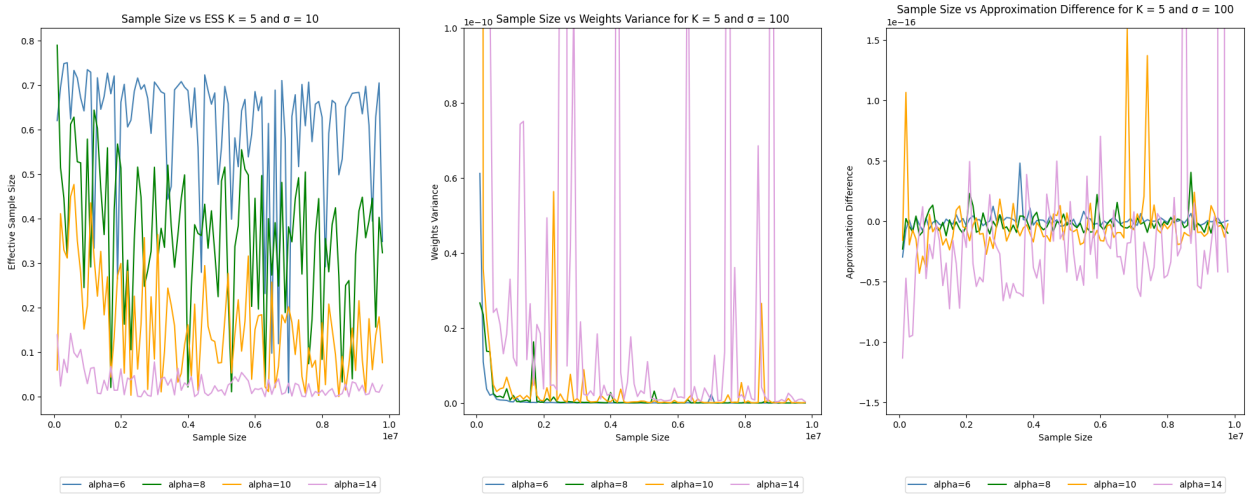

Figure 6: Tuning for proposal distribution with K = 5 and $\sigma = 100$

We begin from $K = 5$ and $\sigma = 10$, the best result is obtained for $\alpha^q = 2$ This choice results in a highly stable ESS, which closely resembles a horizontal line, indicating near-perfect stability. The variance quickly drops to zero, and the integral estimates exhibit an infinitesimal deviation from the reference values, remaining centered around zero with negligible fluctuations.
When we increase σ to 100, the optimal value shifts to $\alpha^q = 6$. Here, while the variance behaves similarly to the previous case, the ESS oscillates significantly, leading to slightly weaker stability. The difference between the estimated and reference integrals fluctuates within the range of $[-10^{-17}, 10^{-17}]$, reinforcing the concerns about the convergence of reference values raised in the previous subsection.

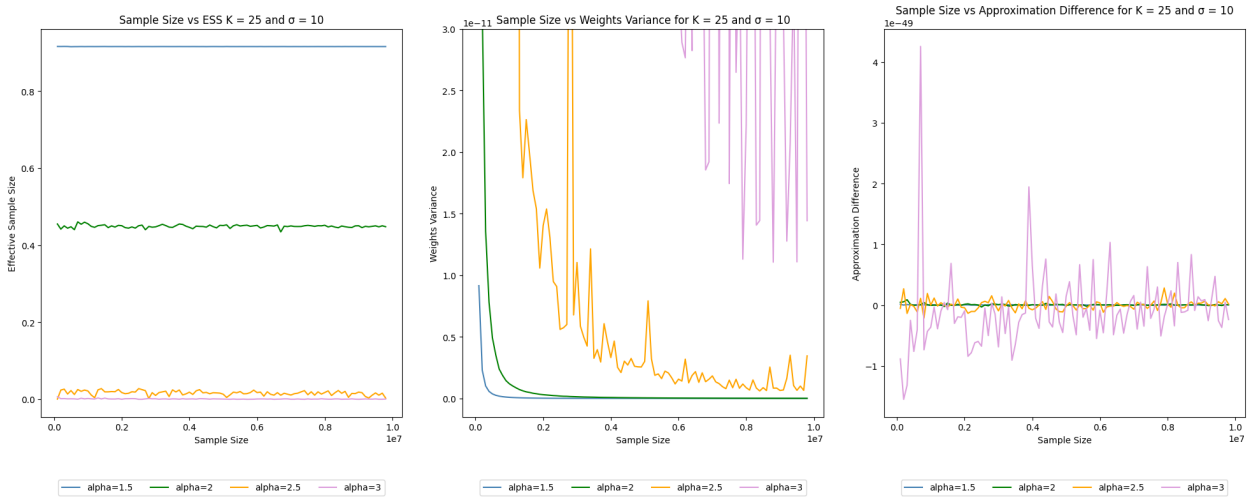

Figure 7: Tuning for proposal distribution with K = 25 and $\sigma = 10$

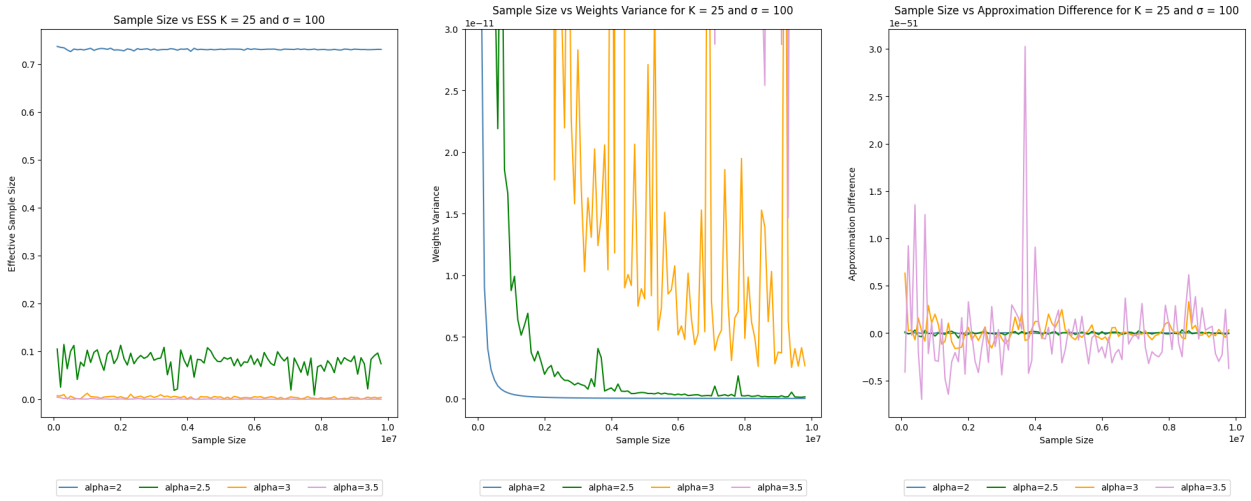

Figure 8: Tuning for proposal distribution with K = 25 and $\sigma = 100$

For $\sigma = 10$ we obtain $\alpha^q = 1.5$ which has a very similar behavior to the one for lower dimension, but scaled to a lower level due to the larger space and smaller volume.
For $\sigma = 100$ the optimum is at $\alpha^q = 2$, the ESS is still stable but its value is more far from 1 centered at 0.7, variance still fall very modest sample size same for which the output captures the reference value.
Next, we plot all the proposal distributions with their respective α^q values to provide a clearer visual representation.

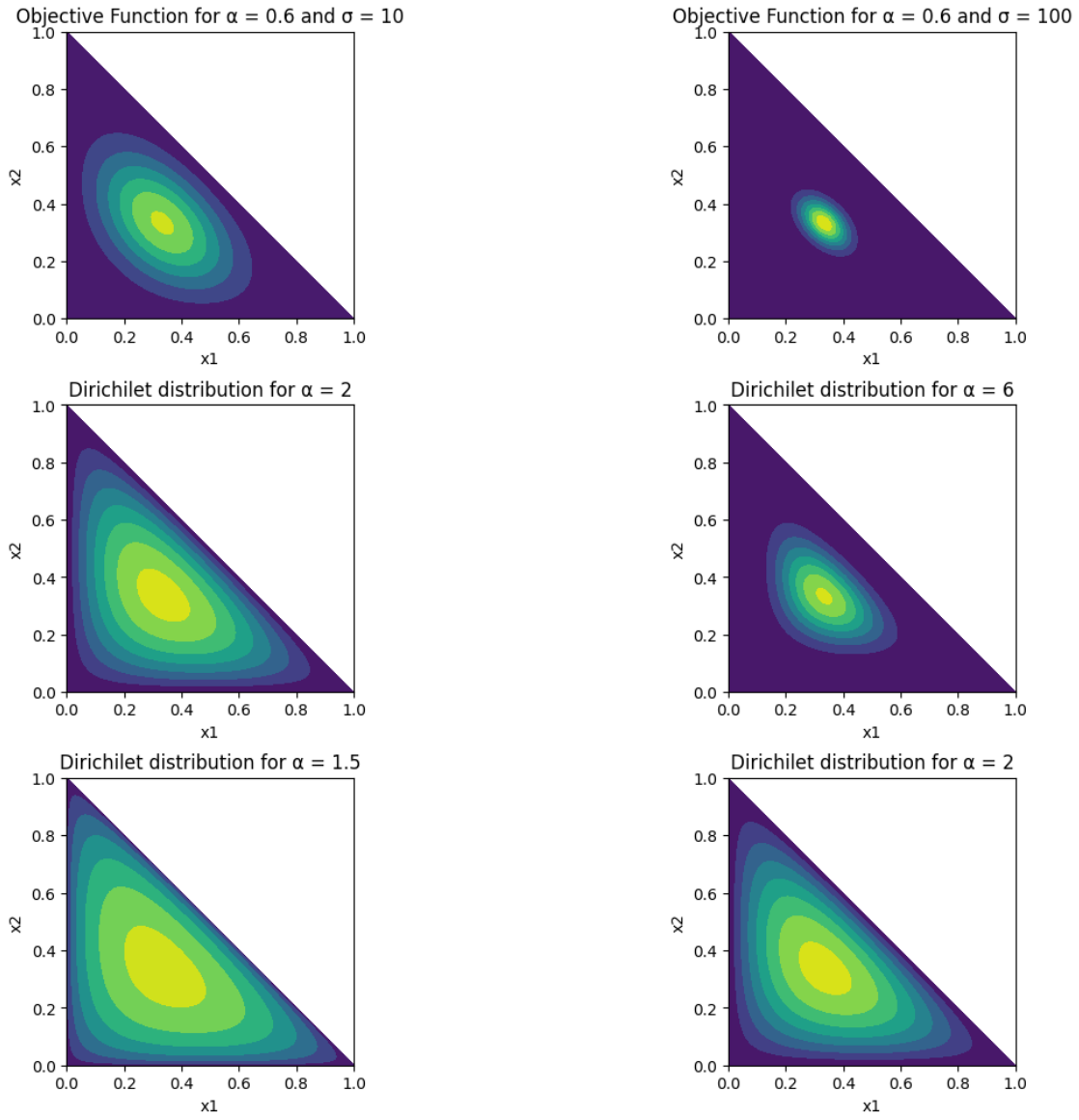

Figure 9: Comparison between objective function and the optimum proposal distributions for different values of K and α

Finally, we asses the computational burden implied by MC integration with importance sampling. We take the optimum proposals found for every level of K and σ and for each of them we plot the computational time require to execute the algorithm with respect to the sample size used as one of the inputs; expecting a linear relationship.

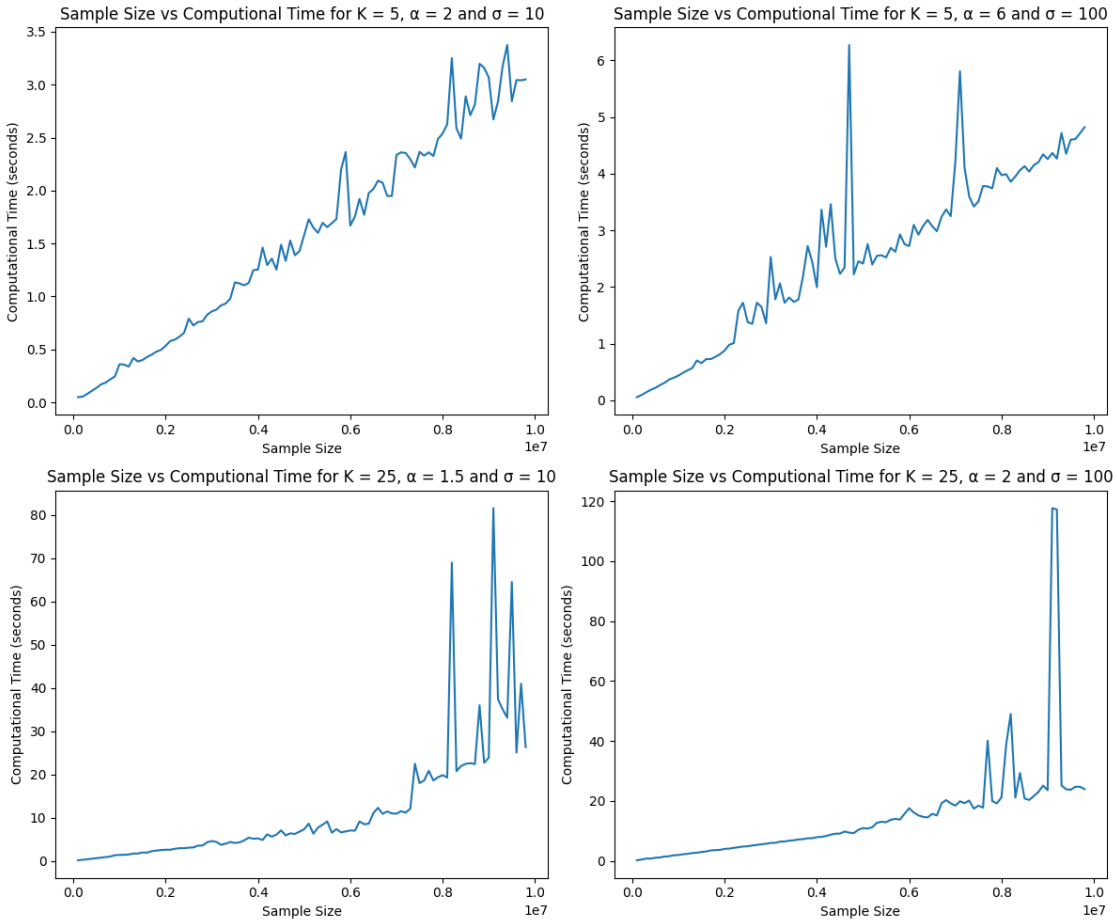

Figure 10: Computational Time vs Sample Size for optimal proposals

All the proposals have a linear computational cost with respect to the the sample size. Furthermore all the figures showed in the section clearly demonstrate the advantages of Monte Carlo integration with importance sampling over nested integration. Due to the enchantment provided by importance sampling, it achieves fast and very stable convergence for both small and large values of K and σ, which let us use a relative small number of sample to obtain an accurate value for our integral. It also drastically reduces computational time. While nested integration methods struggle with exponential complexity as the dimensionality increases, the Monte Carlo method scales linearly with respect to sample size, making it an efficient and practical choice for high-dimensional integrals. This is especially apparent for very large K-allele models, where our approach outperforms the analytical method by several orders of magnitude in terms of speed going from lfew hours to less than a minute. Its only limitation is the unfeasibility to compute the true optimal proposal distribution outlined in (76). This impossibility will lead to a small bias which can always have a non trivial effect considering the fact that we seek the ratio of normalizing constants (39). Moreover, the bias from both the numerator and denominator compounds, increasing the overall bias in the final estimate.

4 Normalizing Constants Ratio Approximation

Looking at the previous results, our goal now is to find an estimator that is both computationally feasible and stable, while also remaining unbiased. A potential solution, which we will describe in detail in this chapter, involves directly approximating the ratio of normalizing constants using the strategies introduced by Neal [1], specifically Annealed Importance Sampling (AIS) and Linked Importance Sampling (LIS). By approximating the ratio directly, we can reduce computational time, as it eliminates the need to estimate the numerator and denominator separately. In terms of stability, both AIS and LIS are unbiased methods, ensuring that this criterion is met. To clearly outline the setting and the strategies we will employ, we first introduce a general notation before adapting it to our specific case. Consider a probability space $(\Omega, \mathcal{H}, \mathbb{P})$, where a probability distribution is defined as:

$$\pi_i(x) = \frac{p_i(x)}{Z_i} \tag{79}$$

with Z_i being the normalizing constant and $p_i(x)$ the unnormalized distribution. We then identify with $i = 0$ and $i = 1$ respectively the distributions at the numerator and the denominator of the ratio we want to approximate; thus our study object can be written as:

$$\frac{Z_1}{Z_0} \quad \text{with} \quad Z_i = \int_\Omega p_i(x)\, dx \quad \forall i = 0, 1 \; \forall x \in \Omega \tag{80}$$

In Bayesian settings, it is common to interpret the numerator distribution π_0 as the prior and the denominator distribution π_1 as the posterior. Considering the normalizing constants of our original model (i.e., $c(\alpha+\theta)$ and $c(\alpha+\theta+\varepsilon_i)$ repsctively prior and posterior), the connection with Bayesian inference becomes clear. In this context, we can give a special meaning to the posterior normalizing constant Z_1, interpreting it as a measure of how well the model explains the observed data.
Various methods have been extensively explored in the literature to estimate ratios of normalizing constants from Monte Carlo data. One straightforward approach and link with the previous section is simple importance sampling (SIS), which relies on the following identity, derived under the assumption that no region with zero probability under π_0 has non-zero probability under π_1:

$$\frac{Z_1}{Z_0} = \mathbb{E}_{\pi_0}\left[\frac{p_1(X)}{p_0(X)}\right] \approx \frac{1}{N}\sum_{i=1}^{N} \frac{p_1(x^{(i)})}{p_0(x^{(i)})} = \frac{1}{N}\sum_{i=1}^{N} \hat{r}_{\text{SIS}}^{(i)} = \hat{r}_{\text{SIS}} \tag{81}$$

In this equation, $\mathbb{E}_{\pi_0}$ represents the expectation with respect to the distribution π_0, which is approximated by the Monte Carlo average over N samples $x^{(i)}, \ldots, x^{(N)}$ drawn from π_0 (either independently or using a Markov chain sampler). Main challenge of simple importance sampling estimate, $\hat{r}_{\text{SIS}}$, is that it can perform poorly if π_0 and π_1 are not sufficiently similar, particularly if regions with significant probability under π_1 have very low probability under π_0.

4.1 Annealed Importance Sampling (AIS)

To address this issue, we propose a more effective approach involving the introduction of intermediate distributions. By parameterizing these distributions using a variable η, we can construct a sequence of distributions $\pi_{\eta_0}, \ldots, \pi_{\eta_n}$, where $\eta_0 = 0$ and $\eta_n = 1$, such that the first and last distributions are π_0 and π_1, respectively, with the intermediate distributions interpolating between them. We can then redefine the ratio as:

$$\frac{Z_1}{Z_0} = \prod_{j=0}^{n-1} \frac{Z_{\eta_{j+1}}}{Z_{\eta_j}} \tag{82}$$

Provided that the distributions $\pi_{\eta_{j+1}}$ and π_{η_j} are sufficiently close, we can estimate each factor $Z_{\eta_{j+1}}/Z_{\eta_j}$ using simple importance sampling outlined by equation (81). From these estimates, we can compute an overall estimate for Z_1/Z_0. Regarding the choice of the functional form of intermediate distributions, different techniques can be employed to define them. Specifically, with $n = 2$, i.e. only one intermediate distribution we reach what is called Bridged sampling. To approximate the ratio, we first define a bridge distribution $\pi_*(x) = p_*(x)/Z_*$, where in this case $\pi_\star = \pi_{\eta_1}$, and then we proceed with the approximation through importance sampling of the ratios between bridge and both numerator and denominator, i.e. respectively Z_*/Z_0 and Z_*/Z_1. After this, we express our objective ratio Z_1/Z_0 as function of the latter subratios:

$$\frac{Z_1}{Z_0} = \frac{\overbrace{\mathbb{E}_{\pi_0}\left[\frac{p_*(X)}{p_0(X)}\right]}^{Z^*/Z_0}}{\underbrace{\mathbb{E}_{\pi_1}\left[\frac{p_*(X)}{p_1(X)}\right]}_{Z^*/Z_1}} \approx \frac{\frac{1}{N_0}\sum_{k=1}^{N_0}\frac{p_*(x_{0,k})}{p_0(x_{0,k})}}{\frac{1}{N_1}\sum_{k=1}^{N_1}\frac{p_*(x_{1,k})}{p_1(x_{1,k})}} = \hat{r}_{\text{bridge}} \tag{83}$$

where $x_{0,1}, \ldots, x_{0,N_0}$ are drawn from π_0 and $x_{1,1}, \ldots, x_{1,N_1}$ are drawn from π_1. A simple choice for the bridge distribution is the geometric bridge:

$$p_{\text{geo}*}(x) = \sqrt{p_0(x)p_1(x)} \tag{84}$$

while, as found out by Bennet [2] and Meng & Wong [8], asymptotically optimal bridge is:

$$p_{\text{opt}*}(x) = \frac{p_0(x)p_1(x)}{r(N_0/N_1)p_0(x) + p_1(x)} \tag{85}$$

where $r = Z_1/Z_0$. Of course, we cannot use this bridge distribution in practice, since we do not know r. We can, however, use a preliminary guess at r to define an initial bridge distribution, which will provide a bridge sampling estimate for Z_1/Z_0. Using this estimate as the new value of r, we can refine our bridge distribution, iterating this process as many times as desired.

Although both simple importance sampling and bridge sampling have been successfully used in many applications, they have some deficiencies. A first issue is that the bridge sampling estimate of equation (83) is biased, and the same would appear to be the case for an estimate using intermediate distributions. This criticality will, at the end, preclude averaging independent replications of the bridge sampling estimate to obtain a better estimate, since the bias would prevent convergence to the correct value as the number of replications increases. The second challenge, which actually requires a major tuning effort, is that, except sometimes for π_0, sampling from the distributions π_η must usually be done by Markov chain methods which approach the desired distribution only asymptotically. To speed convergence, the Markov chain for sampling π_{η_j} is often started from the last state sampled for $\pi_{\eta_{j-1}}$, but it is unclear how many iterations should then be discarded before an adequate approximation to the correct distribution is reached.

To correct the bias we resort to another sampling strategy: Annealed Importance Sampling. The theoretical foundation of this strategy is derived by Jarzinsky [5] and Neal [9]. They show that an estimate for Z_1/Z_0 using n intermediate distributions as in equation (82) will be exactly unbiased if each of the ratios $Z\eta_{j+1}/Z_{\eta_j}$ for $j = 0, \ldots, n-1$; is estimated using simple importance sampling with only one point sampled x^j (we are using (81) with $N = 1$). Moreover every x^j is drawn from its respective distribution π_{η_j} through Markov chain updates starting from x^{j-1}, i.e. $\text{Trans}(x^{j-1}, \pi_{\eta_j})$.

Averaging the estimates obtained from M independent replications of this process produces the following estimate:

$$\frac{Z_1}{Z_0} \approx \frac{1}{M}\sum_{i=1}^{M}\prod_{j=0}^{n-1}\frac{p_{\eta_{j+1}}(x_j^i)}{p_{\eta_j}(x_j^i)} = \frac{1}{M}\sum_{i=1}^{M}\hat{r}_{AIS}^i = \hat{r}_{AIS} \tag{86}$$

In this case, we can even increase the transitions of the Markov chain, but it is not strictly necessary since, even if only one step is insufficient to reach equilibrium, our estimate $\hat{r}_{AIS}$ is nevertheless exactly unbiased and will converge to the true value as M increases. Finally, another strenght of AIS is that the condition for this is slightly weaker than for importance sampling, as we only need to ensure that no region having zero probability under η_j has nonzero probability under η_{j+1}. To be more precise, what we have described so far is typically referred to as Forward AIS, but there is a counterpart called Backward AIS (or Reverse AIS), which estimates Z_0/Z_1. The strategy is the same as for the forward case, but it begins with a sample x_n drawn from π_{η_n} and progresses by approximating the ratios $Z_{\eta_j}/Z_{\eta_{j-1}}$ for $j = n-1, \ldots, 1$. We denote the backward approximation as $\hat{r}^*_{AIS}$, which can be derived as:

$$\frac{Z_0}{Z_1} \approx \frac{1}{M}\sum_{i=1}^{M}\prod_{j=n}^{1}\frac{p_{\eta_j}(x_j^i)}{p_{\eta_j-1}(x_j^i)} = \frac{1}{M}\sum_{i=1}^{M}\hat{r}_{AIS}^{i\star} = \hat{r}^{\star}_{AIS} \tag{87}$$

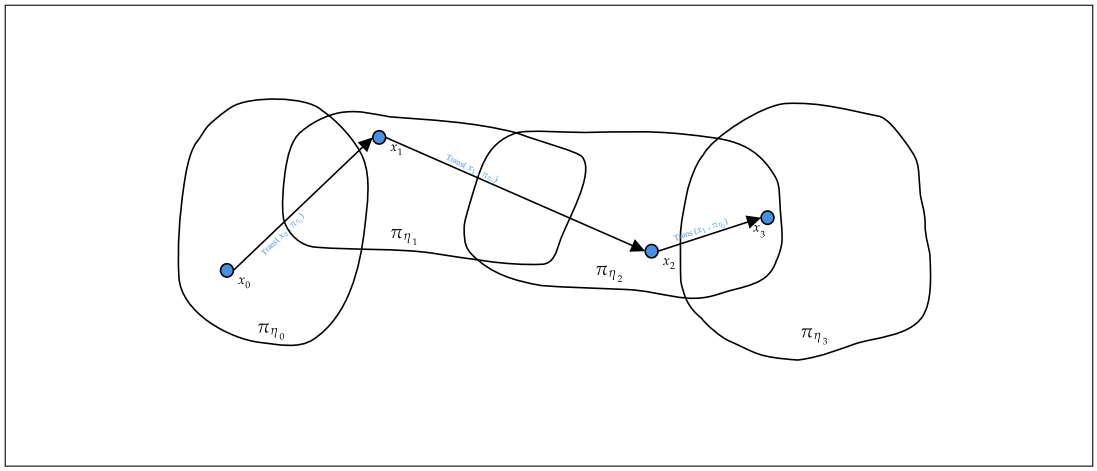

Figure 11: AIS Forward example

We conclude this strategy presentation with a brief example of AIS implementation, accompanied by the pseudo-code used in our approach which for computational reasons will treat the logarithm of the ratio. The example is illustrated in Figure (11), depicting the process of transitioning between successive distributions π_{η_j} and $\pi_{\eta_{j+1}}$. Each distribution π_{η_j} is shown as an oval representing regions of high probability, and the arrows indicate the Markov chain updates that move the sample from one distribution to the next. By reversing the direction of the arrows, we obtain the reverse AIS procedure. For the pseudo-code, we present it in terms of the log-ratio, $\log(Z_1/Z_0)$, as logarithms provide better numerical stability during computation. The pseudo-code is as follows:

Algorithm 2: Annealed Importance Sampling (AIS) Algorithm for Log Ratios

Input: Log probability function $lpr(x, \eta)$;
Sequence of parameters $\boldsymbol{\eta} = [\eta_0, \eta_1, \ldots, \eta_n]$;
Transition function $trans(x, lpr, \eta, stepsize)$;
Burn-in period *burn_in*;
Number of samples *num_samples*;
Thinning interval *thinning*;
Forward flag *forward*;
Iteration number **m**;

Output: Log ratio estimates

for $j < m$ **do**
  if *forward* **then**
    Draw $\mathbf{x_0^j} \sim \pi_0$;
    Set $order = [1, 2, \ldots, n]$;
    Set $k = 1$;
  end
  else
    Draw $\mathbf{x_n^j} \sim \pi_n$;
    Set $order = [n-1, n-2, \ldots, 1]$;
    Set $k = -1$;
  end
  for $i \in order$ **do**
    Compute : $\log(r)_i = lpr(x_{i-k}, \eta_i) - lpr(x_{i-k}, \eta_{i-k})$;
    if $(i < n \ \wedge \ forward) \ \vee \ (i > 1 \ \wedge \ \neg forward)$ **then**
      Perform MCMC transition:
      $\mathbf{x}_{i+k}^j \leftarrow trans(\mathbf{x}_i, lpr, \eta_i, \text{burn\_in}, \text{num\_samples}, \text{thinning})$;
    end
  end
  Compute $\hat{r}_{AIS}^j = \sum_i \log(r)_i$
end
return *Compute* $\hat{r}_{AIS} = \frac{1}{m} \sum_j \hat{r}_{AIS}^j$;

Here, the objective function f is substitued by its logarithms $lpr = \log(F)$ and *forward* is a flag such that if it is true, we are performing a Forward AIS; otherwise, we execute a Backward AIS. Moreover, the implementation and draw of only one sample with only one update are translated into using, as parameters of trans, the tuple $(x_{j-1}, \log(\pi_{\eta_j}), 1, 1)$.

4.2 Linked Importance Sampling (LIS)

Another sampling technique to estimate the ratio of normalizing constants, also outlined by Neal [9], is called Linked Importance Sampling (LIS). LIS is a hybrid method that combines elements of Annealed Importance Sampling (AIS) and Bridge Sampling, and depending on the context, it can significantly enhance performance compared to either technique when applied individually. From Bridge Sampling, LIS borrows the concept of using auxiliary sub-ratios Z^*/Z_{j+1} and Z_j/Z^* , which are computed through bridge distributions p^*. The two samples required for these sub-ratios are "linked" through a single shared state, used in both, giving rise to the name "Linked Importance

Sampling". From AIS, LIS incorporates the use of multiple auxiliary distributions $\{\pi_{\eta_j}\}_j$ and retains the key advantage of being exactly unbiased. This property holds even when intermediate distributions are used or when sampling is performed with Markov chain transitions that have not yet reached their equilibrium distributions.
In this procedure we denote again with the values $\eta_0 = 0$ and $\eta_n = 1$ the two distributions π_{η_0} and π_{η_1} we are interested in, for which the normalizing constants are Z_0 and Z_1. Moreover, we assume to have for each distribution, π_η a pair of Markov chain transition probability (or density) functions, denoted by $T_\eta(x, x')$ and $T'_\eta(x, x')$, satisfying $\int T_\eta(x, x')dx' = 1$ and $\int T'_\eta(x, x')dx' = 1$, for which the following mutual reversibility relationship holds:

$$\pi_\eta(x)T_\eta(x, x') = \pi_\eta(x')T'_\eta(x', x) \qquad \forall x, x' \in \Omega \tag{88}$$

From this relationship, one can easily show that both T_η and T'_η leave π_η invariant, i.e.

$$\begin{aligned}\int \pi_\eta(x)T_\eta(x, x')dx &\overset{88}{=} \int \pi_\eta(x')T'_\eta(x', x)dx \\ &= \pi_\eta(x') \underbrace{\int T'_\eta(x', x)dx}_{=1} \\ &= \pi_\eta(x')\end{aligned}$$

and the same holds for T'_η. These Markov chain transitions are used to obtain samples that are approximately drawn from each of the $n+1$ distributions, $\pi_{\eta_0}, \ldots, \pi_{\eta_n}$. Furtermore if T_η is reversible (i.e., satisfies 'detailed balance') then T_η will be the same as T'_η. Non-reversible transitions often arise when components of the state are updated in some predetermined order (think to classical Gibbs sampling), in which case the reverse transition simply updates components in the opposite order. As a special case, T_η might draw the next state from π_η independently of the current state. Such independent sampling may often be possible for T_0.
To perform LIS sampling we set our starting point from π_{η_0}. We sample from the latter distribution N_0 points $x_{0i} \sim \pi_{\eta_0}\ \forall i = 1, \ldots, N_0$, and we randomly pick an integer ν_0 from the uniform discrete distribution over N_0, i.e. $\nu_0 \sim \text{Unif}(\{0, N_0\})$. Then for $j = 0, \ldots, n$ we iterate the following: first, if $j > 0$ we again pick a random integer ν_j from the uniform distribution over N_j and set $x_{j\nu_j} = x_{j*j+1}$

$$x_{j\nu_j} = x_{j*j+1} \quad \text{with} \quad \nu_j \sim \text{Unif}(\{0, N_j\}) \tag{89}$$

Then if $\nu_j < N_j$ we use the forward Markov chain transition probability T_{η_j} to draw all the succesive x_{jk} until $k = N_j$.

$$x_{jk} \sim T_{\eta_j}(x_{jk-1}, x_{jk})\ \forall k \in \{\nu_{j+1}, \ldots, N_j\} \tag{90}$$

if instead $\nu_j > 0$ we do the reverse of what we have done until now, employing the backward markov transition function $T^*_{\eta_j}$ to draw the remaining elements of the sample:

$$x_{jk} \sim T'\eta_j(x_{jk+1}, x_{jk})\ \forall k \in \{\nu_{j-1}, \ldots, 0\} \tag{91}$$

Finally, excluded for the last iteration we set $x_{j*j+1} = x_{j\mu_j}$ where μ_j takes value in $\mathbb{N}_{0,N_j}$ according to the following probability:

$$x_{j*j+1} = x_{j\mu_j} \quad \text{with} \quad \mu_j \sim \Pi_0(\mu_j \mid x_j) = \frac{p_{j^*j+1}(x_{j\mu_j})}{p_{\eta_j}(x_{j\mu_j})} \Big/ \sum_{k=0}^{N_j} \frac{p_{j^*j+1}(x_{jk})}{p_{\eta_j}(x_{jk})} \tag{92}$$

where p_{j*j+1} is a bridge distribution build using formulas (84) and (85) and intermediate distributions π_{η_j} and $\pi_{\eta_{j+1}}$. Then we can compute the LIS approximation for one run as follows:

$$\hat{r}^i_{\text{LIS}} = \prod_{j=0}^{n-1} \left[\frac{1}{N_j+1} \sum_{k=0}^{N_j} \frac{p_{j^*j+1}(x_{jk})}{p_{\eta_j}(x_{jk})} \Big/ \frac{1}{N_{j+1}+1} \sum_{k=0}^{N_{j+1}} \frac{p_{j^*j+1}(x_{j+1,k})}{p_{\eta_{j+1}}(x_{j+1,k})} \right] \tag{93}$$

As in the AIS case we can even execute the a backward version of LIS $\hat{r}^\star_{\text{LIS}}$ estimating the ratio Z_0/Z_1. The procedure is the same of the forward version but with the inverted order. We begin sampling N_n samples form π_{η_n} and through link samples we move from one intermediate distribution to the other in the exact same way, with only differences are bridge distribtuions which cahnge direction. Hence the fromula to choose link state becomes:

$$x_{j-1*j} = x_{j\mu^\star_j} \quad \text{with} \quad \mu^\star_j \sim \Pi_1(\mu^\star_j \mid x_j) = \frac{p_{j-1^*j}(x_{j\mu^\star_j})}{p_{\eta_j}(x_{j\mu^\star_j})} \Big/ \sum_{k=0}^{N_j} \frac{p_{j-1^*j}(x_{jk})}{p_{\eta_j}(x_{jk})} \tag{94}$$

and our one-run estimation is computed as:

$$\hat{r}^{\star i}_{\text{LIS}} = \prod_{j=1}^{n} \left[\frac{1}{K_j+1} \sum_{k=0}^{K_j} \frac{p_{j-1^*j}(x_{jk})}{p_{\eta_j}(x_{jk})} \Big/ \frac{1}{K_{j-1}+1} \sum_{k=0}^{K_{j-1}} \frac{p_{j-1^*j}(x_{j-1,k})}{p_{\eta_{j-1}}(x_{j-1,k})} \right] \tag{95}$$

Finally to get the final estimate for both directions we average the outputs of M runs:

$$\hat{r}_{\text{LIS}} = \frac{1}{M} \sum_{i=1}^{M} \hat{r}^i_{\text{LIS}} \quad \text{and} \quad \hat{r}^\star_{\text{LIS}} = \frac{1}{M} \sum_{i=1}^{M} \hat{r}^{\star i}_{\text{LIS}} \tag{96}$$

Furthermore, since the LIS estimate can be viewed as a simple importance sampling estimate on an extended space, we can propose a third form called Bridged LIS. In this proposal we first perform both backward and forward LIS. Then we define top level unnormalized probability densities $P_1(x,\mu)$ and $P_0(x,\mu)$ defined functions of Π_1 and Π_0 which are distributions over all the quantities genereated through all the forward and backward steps and such that:

$$\Pi_0(x_j \mid x_j, \mu_j) = \frac{\pi_{\eta_j}(x_j, 0)}{\pi_{\eta_j}(x_j, \mu_j)} \prod_{k=1}^{n} T_{\eta_j}(x_j, k-1, x_j, k) \quad \text{and} \quad \Pi_0(\mu_j) = \text{Unif}(\{0, N_j\}) \tag{97}$$

$$\Pi_0(x_j \mid x_j, \mu^\star_j) = \frac{\pi_{\eta_j}(x_j, 0)}{\pi_{\eta_j}(x_j, \mu^\star_j)} \prod_{k=1}^{n} T_{\eta_j}(x_j, k-1, x_j, k) \quad \text{and} \quad \Pi_1(\mu^\star_j) = \text{Unif}(\{0, N_j\}) \tag{98}$$

then the formula for the top level distributions provided by Neal [9]:

$$P_0(x,\mu) = Z_0 \Pi_0(\mu_n) \pi_{\eta_n}(x_n, \mu_n) \prod_{j=0}^{n-1} \Pi_0(\mu_j) \prod_{j=0}^{n} \Pi_z(x_j \mid x_j, \mu_j) \prod_{j=1}^{n} \Pi_0(\mu_j \mid x_j) \Pi_0(\mu_0) \tag{99}$$

$$P_1(x,\mu^\star) = Z_1 \Pi_1(\mu^\star_n) \pi_{\eta_n}(x_n, \mu^\star_n) \prod_{j=0}^{n-1} \Pi_1(\mu^\star_j) \prod_{j=0}^{n} \Pi_1(x_j \mid x_j, \mu^\star_j) \prod_{j=1}^{n} \Pi_1(\mu^\star_j \mid x_j) \Pi_1(\mu^\star_0) \tag{100}$$

and it can be proved that the ratio for the normalizing constant of the two is equal to our object ratio Z_1/Z_0

$$\frac{\int_\Omega P_1(x,\mu^\star)}{\int_\Omega P_0(x,\mu)} = \frac{Z_1}{Z_0} \tag{101}$$

After this, we also need a suitable bridge distribution, P_*, for which we must be able to evaluate the ratios P_*/P_0 and P_*/P_1. Assuming the forward procedure is performed M times and the reverse procedure $M^\star$ times we define the optimal bridge as follows:

$$\frac{P_{\text{opt}^*}(x,\mu,\mu^\star)}{P_0(x,\mu)} = \left[r(M/M^\star) \left(\frac{P_1(x,\mu^\star)}{P_0(x,\mu)} \right)^{-1} + 1 \right]^{-1} \tag{102}$$

$$\frac{P_1(x,\mu^\star)}{P_{\text{opt}^*}(x,\mu,\mu^\star)} = \left[r(M/M^\star) + \left(\frac{P_1(x,\mu^\star)}{P_0(x,\mu)} \right)^{-1} \right]^{-1} \tag{103}$$

while we outline the geometric one as:

$$\frac{P_{\text{geo}^*}(x,\mu,\mu^\star)}{P_0(x,\mu)} = \sqrt{\frac{P_1(x,\mu^\star)}{P_0(x,\mu)}} \tag{104}$$

$$\frac{P_{\text{geo}^*}(x,\mu,\mu^\star)}{P_1(x,\mu^\star)} = \sqrt{\frac{P_0(x,\mu)}{P_1(x,\mu^\star)}} \tag{105}$$

These expressions allow us to express bridged LIS estimates in terms of the simple LIS estimate forward $\hat{r}$ and backward $\hat{r}^\star$. For the optimal bridge, we get

$$\hat{r}_{\text{opt}}^{\text{LIS-bridged}} = \frac{1}{M} \sum_{i=1}^{M} \frac{1}{r(M)/\hat{r}_{\text{LIS}}^i + 1} \Big/ \frac{1}{M^\star} \sum_{i=1}^{M^\star} \frac{1}{r(M^\star) + 1/\hat{r}_{\text{LIS}}^{i\star}} \tag{106}$$

Similarly, for the geometric bridge, we get

$$\hat{r}_{\text{geo}}^{\text{LIS-bridged}} = \frac{1}{M} \sum_{i=1}^{M} \sqrt{\hat{r}_{\text{LIS}}^i} \Big/ \frac{1}{M^\star} \sum_{i=1}^{M^\star} \sqrt{\hat{r}_{\text{LIS}}^{i\star}} \tag{107}$$

Following the structure we used for AIS, we now illustrate a simple but straightforward example of LIS and how K states are drawn and linked for each π_η distribution. Assuming $n = 3$, $K = 7$ and as given the Markov transition functions T and T' so that we can define $\pi_{\eta 0}$ and $\pi_{\eta 3}$ as our target distribution at the numerator and denominator respectively.

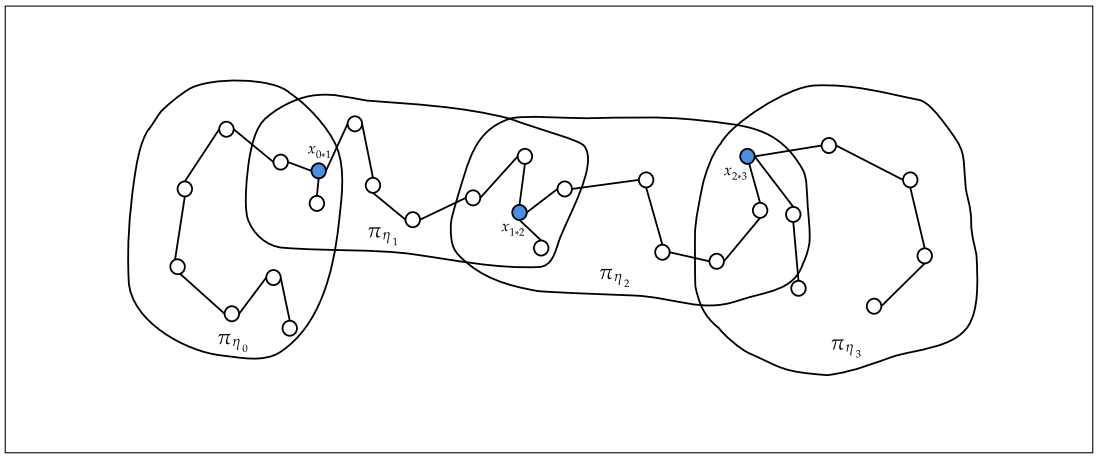

Figure 12: LIS forward example

Here again, each distribution π_η is represented by an oval enclosing the regions of high probability. Six Markov chain transitions are performed at each stage and the link states x_{i*i+1} are shown as blue scatters.

Finally we provide with (3) a pesudo code of our implementation for the estimation of the logarithms of the ratio. As for the prevoius one, *forward* is again a flag such thta if it is true we are performing the forward LIS and the backward if it is false. Furtermore to differentiate between μ and $\mu^\star$ we implement a list *order* which switches from $0 \dots n$ to $n \dots 0$ based on the flag assignement

Algorithm 3: Linked Importance Sampling (LIS) Algorithm for Log Ratios

Input: Log probability function $lpr(x, \eta)$;
Sequence of parameters $\boldsymbol{\eta} = [\eta_0, \eta_1, \ldots, \eta_n]$;
Step size vector **stepsize**;
Markov transition functions T, T';
Iteration number **m**;
Samples drawn number **K**;

Output: Log ratio estimates

for $j < m$ **do**
  if *forward* **then**
    Draw $\nu_0 \sim \text{Unif}(\{0, K\})$;
    Set $x^j_{0,\nu_0} = x \sim \pi_0$;
    Set $order = [0, 1, \ldots, n]$;
    Set $z = 1$;
  end
  else
    Draw $\nu_n \sim \text{Unif}(\{0, K\})$;
    Set $x^j_{n,\nu_n} = x \sim \pi_n$;
    Set $order = [n, n-1, \ldots, 0]$;
    Set $z = -1$;
  end
  for $i \in order$ **do**
    if $(i < n \ \wedge \ forward) \ \vee \ (i > 1 \ \wedge \ \neg forward)$ **then**
      Draw $\nu_i \sim \text{Unif}(\{0, K\})$;
      Set $x^j_{i,\nu_i} = x_{i*i+k}$
    end
    for $k \in \{\nu_i + 1, k\}$ **do**
      Draw $x^j_{i,k} \leftarrow T_{\eta_j}(x_{j,k-1}, x_{j,k})$
    end
    for $k \in \{\nu_i - 1, 0\}$ **do**
      Draw $x^j_{i,k} \leftarrow T'_{\eta_j}(x_{j,k+1}, x_{j,k})$
    end
    if $(i > 0 \ \wedge \ forward) \ \vee \ (i < n \ \wedge \ \neg forward)$ **then**
      Draw $\mu_i \sim \Pi(\mu_i | x_i s)$;
      Set $x^j_{i*i+v} = x_{j,\mu_j}$
    end
  end
  Compute $\hat{r}^j_{\text{LIS}}$
end
return *Compute* $\hat{r}_{LIS} = \frac{1}{m} \sum_j \hat{r}^j_{LIS}$;

4.3 Implementation

In this section, we will evaluate the performance of the proposed strategies, including AIS (Annealed Importance Sampling) in both forward and backward modes, as well as LIS (Linked Importance Sampling) in backward, forward, and bridged modes, using both the optimal and the geometric bridge. Since we will later compare these methods to those used for approximating a single normalizing constant (naturally for te latter we will estimate both numerator and denominator), we adopt the selective overdominance model used by Genz and Joyce, as described in equation (58). We begin by defining the integrals Z_0 and Z_1 as follows:

$$Z_0 = c_\sigma(\alpha) \quad \text{and} \quad Z_1 = c_\sigma(\alpha + \varepsilon) \tag{108}$$

Here, $c_\sigma(\alpha)$ represents the normalizing constant for a parameter vector $\alpha \in \mathbb{R}^K$, given a fitness matrix $\sigma \in \mathbb{R}^{K\times K}$, as outlined in equation (59). The vector $\varepsilon \in \mathbb{R}^K$ is a unit vector where its only non-zero component ε_i with $i \in \mathbb{N}_{0,K}$ is either 1 or -1 , depending on whether we are considering a birth or death event in the birth-and-death process for the corresponding i^{th} allele. Then we generate the sequence of intermediate distributions $\pi_{\eta_0}, \pi_{\eta_1}, \ldots, \pi_{\eta_n}$, with $\eta_0 = 0$, $\eta_n = 1$, and $\eta_i > \eta_j$ for all $i > j$, we define the unnormalized density function $p_{\eta_k}(x)$ for π_{η_k} as follows:

$$p_{\eta_k}(x) = x_1^{\alpha_1+\eta_k\varepsilon_1} \ldots x_{K-1}^{\alpha_{n-1}+\eta_k\varepsilon_{n-1}} \left(1 - \sum_{j=1}^{K-1} x_j\right)^{\alpha_n+\eta_k\varepsilon_n} e^{-\sigma\left(x_1^2+\cdots+x_{K-1}^2+\left(1-\sum_{j=1}^{K-1} x_j\right)^2\right)} \tag{109}$$

From this, setting $k = 0$ and $k = 1$, we obtain the unnormalized distributions generating Z_0 and Z_1 , respectively:

$$p_{\eta_0}(x) = x_1^{\alpha_1} \ldots x_{K-1}^{\alpha_{n-1}+} \left(1 - \sum_{j=1}^{K-1} x_j\right)^{\alpha_n} e^{-\sigma\left(x_1^2+\cdots+x_{K-1}^2+\left(1-\sum_{j=1}^{K-1} x_j\right)^2\right)} \tag{110}$$

$$p_{\eta_k}(x) = x_1^{\alpha_1+\varepsilon_1} \ldots x_{K-1}^{\alpha_{n-1}+\varepsilon_{n-1}} \left(1 - \sum_{j=1}^{K-1} x_j\right)^{\alpha_n+\varepsilon_n} e^{-\sigma\left(x_1^2+\cdots+x_{K-1}^2+\left(1-\sum_{j=1}^{K-1} x_j\right)^2\right)} \tag{111}$$

Finally, we define the forward and backward transition functions $T(x, x\prime)$ and $T\prime(x, x\prime)$. We implement a Metropolis algorithm using a Dirichlet proposal distribution $\mathrm{Dir}(\alpha^q)$, where α^q is a parameter vector that we will later tune to reduce the number of samples required while still maintaining accurate results. For the other parameters, we follow the previous section, setting the dimensionality $K = 5$. Additionally, to leverage one of the core features of AIS and LIS—computing the ratio between two probability distributions with significantly distant high-density regions—we let σ increase to 500. This creates an extremely concentrated probability mass, where even small changes result in a mismatch between the regions of high density. We choose α to be a uniform vector with $\alpha_i = 1\ \forall i \in \mathbb{N}_{0,K}$. This choice is arbitrary since our focus is on the magnitude of the shift in the probability mass after adding ε . To better visualize the regions being approximated, we reduce the dimensionality to $K = 3$ and plot the first two dimensions on the simplex.

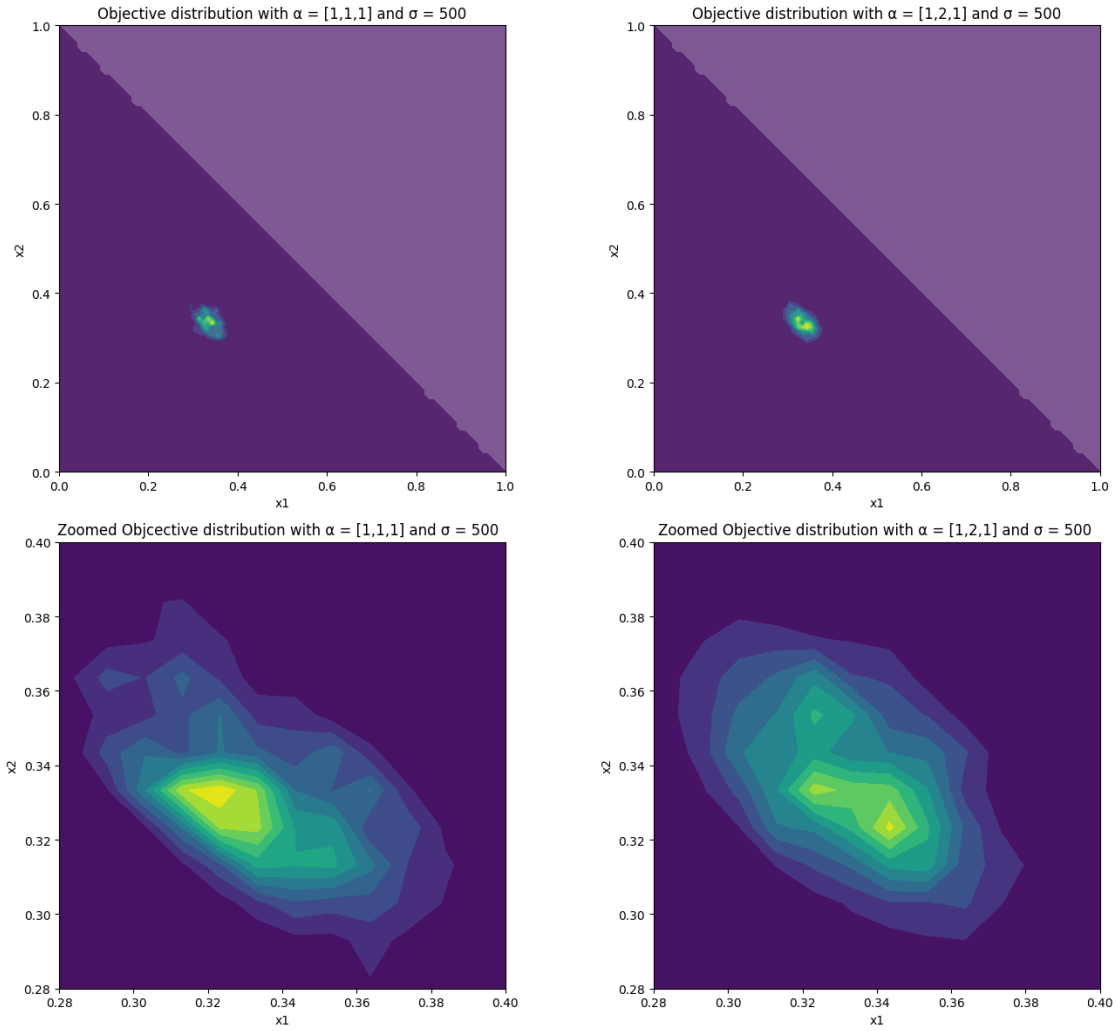

Figure 13: Probability mass of p_{η_0} and p_{η_1}

In the plot, we zoom in on the x_1 and x_2 axes due to the small size of the probability distributions. Next, we employ Monte Carlo (MC) integration with importance sampling, using an extremely large number of samples to retrieve a highly accurate benchmark for the ratio of interest. The results are as follows:

$$\frac{Z_1}{Z_0} = 1.99997 \times 10^{-1} \quad \text{and} \quad \frac{Z_0}{Z_1} = 5.000075 \tag{112}$$

We begin our tests with AIS, structuring our approach as follows: first, we tune the parameter vector α^q for the Metropolis updates, and then we move the optimization onto the other two parameters: M, the total number of runs over which we average our results, and n, the number of intermediate distributions used. As a criterion for the quality of our estimation, we use the difference between our results and those described in (112).

For the forward AIS, we first obtain the chart in Figure 14 for the optimal proposal vector: we keep the partition of the interval [0,1] fixed at 1000, which will be an assumption from now on, and we plot the difference between our ratio integrals as a function of the number of iterations to observe convergence. From this, we clearly see that the final choice is $\alpha^q = 3$, as both lower and higher values of α introduce oscillations. Next, we move to optimizing M and n. Keeping α^q fixed, we plot the

difference between our outputs and the reference values, varying n, while observing the results as a function of M. This is shown in Figure 15. We observe that increasing the number of intermediate distributions improves accuracy, with near-perfect convergence for $n = 1000$. Moreover, regarding the final choice of M, we see that for $M > 2000$, the difference curve becomes almost horizontal, allowing us to select $M = 2000$ as a sufficient number of iterations.

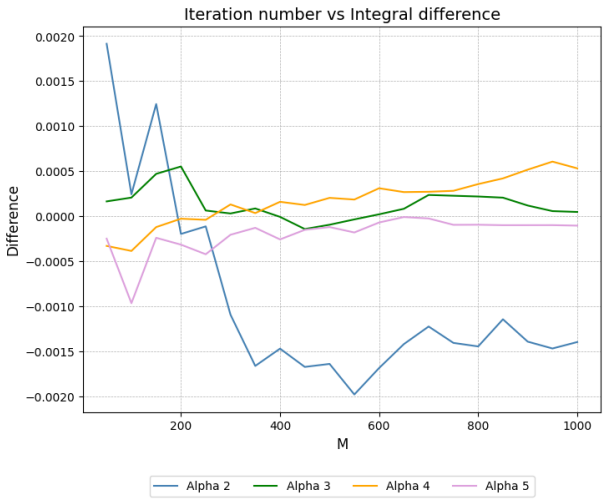

Figure 14: Forward AIS convergence for different α^q proposals

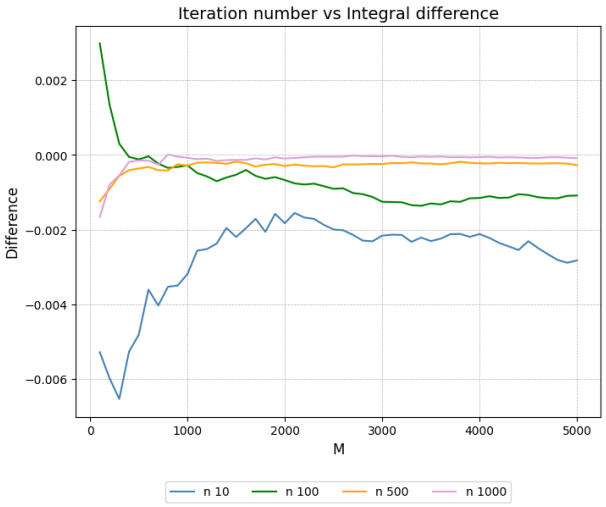

Figure 15: Forward AIS convergence for different n given α^q

Moving on to the backward AIS, we again compute the plot for the optimal α^q, settling on $\alpha^q = 4$, as lower values result in excessive oscillations. For n, we can choose 500, since the convergence at $n = 500$ is very close, if not better than, that for $n = 1000$. As before, we determine that $M = 2000$ is a sufficient number of iterations.

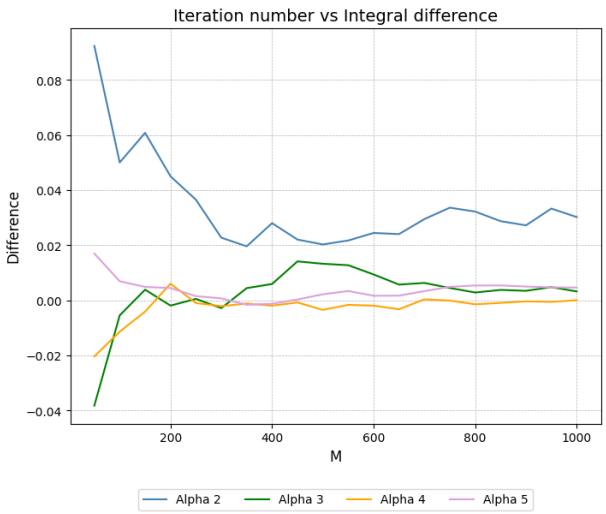

Figure 16: Backward AIS convergence for different α^q proposals

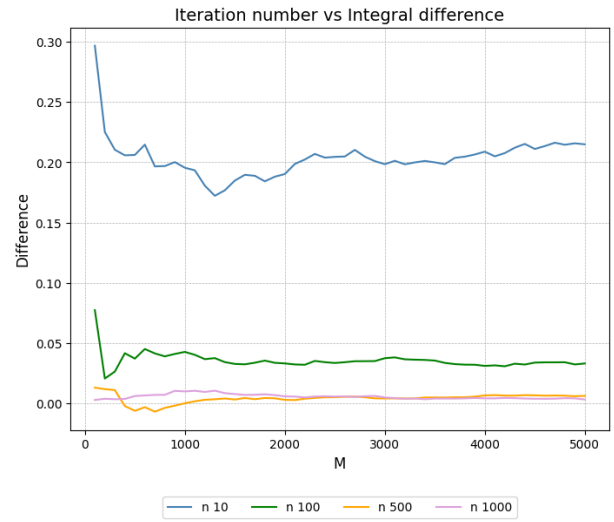

Figure 17: Backward AIS convergence for different n given α^q

Next, we move to LIS, starting with both the backward and forward methods using the geometric

bridge. After that, we explore the backward and forward methods with the optimal bridge. The bridged version will be computed last, as it is a function of the previous two. For LIS, we do not need the Metropolis-Hastings update to reach the stationary distribution. Therefore, we set a burn-in of 100 and a thinning factor of 1, though other implementations are possible. To maintain computational costs similar to AIS, we reduce the number of iterations, using a maximum of $M = 10$. For the geometric forward LIS we recover the next plots:

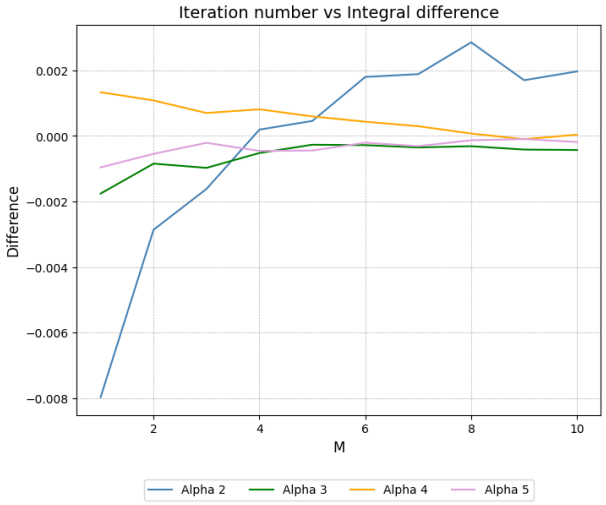

Figure 18: Geometric Forward LIS convergence for different α^q proposals

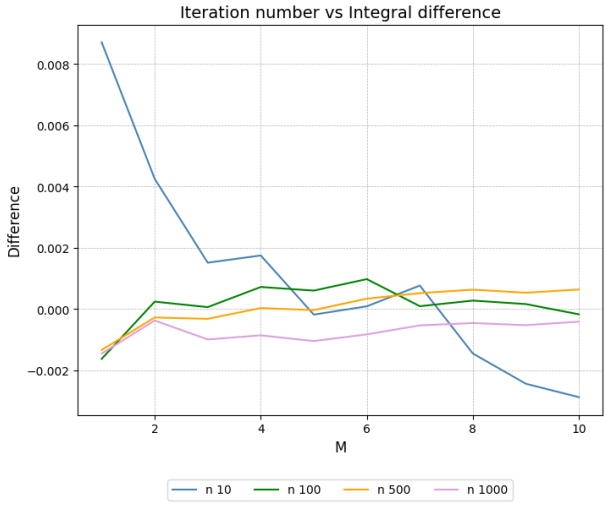

Figure 19: Geometric Forward LIS convergence for different n given α^q

For the proposal parameter, we can select either $\alpha^q = 3$ or $\alpha^q = 4$ with minimal distinction, but we ultimately opt for $\alpha^q = 4$. Regarding n and M, $n = 500$ works well as the number of intermediate distributions, while oscillations become negligible for $M > 6$. Its geometric counterpart is represented in the figures below:

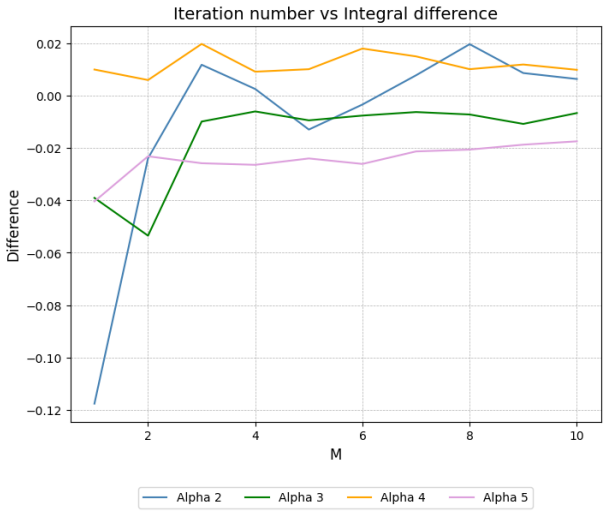

Figure 20: Geometric Backward LIS convergence for different α^q proposals

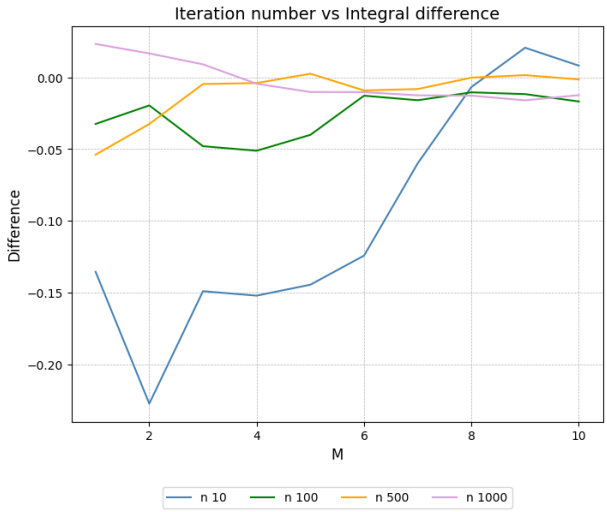

Figure 21: Geometric Backward LIS convergence for different n given α^q

Here, while it is worth note tht we have a weaker convergence in figure (20), we ultimately pick the same set of parameters as the geometric forward. We plot now the LIS with optimal bridge distribution.

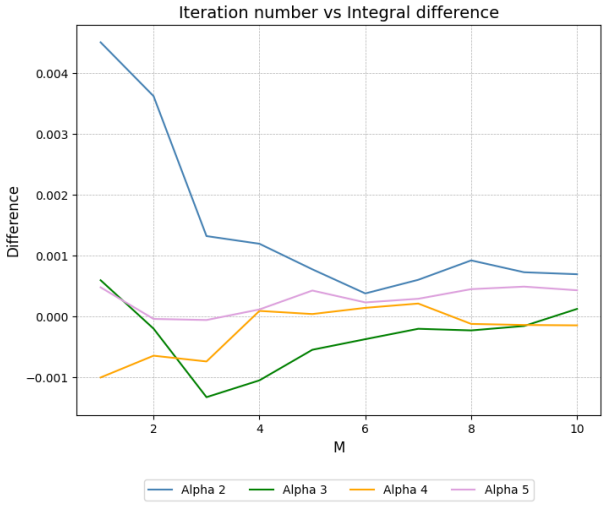

Figure 22: Optimal Forward LIS convergence for different α^q proposals

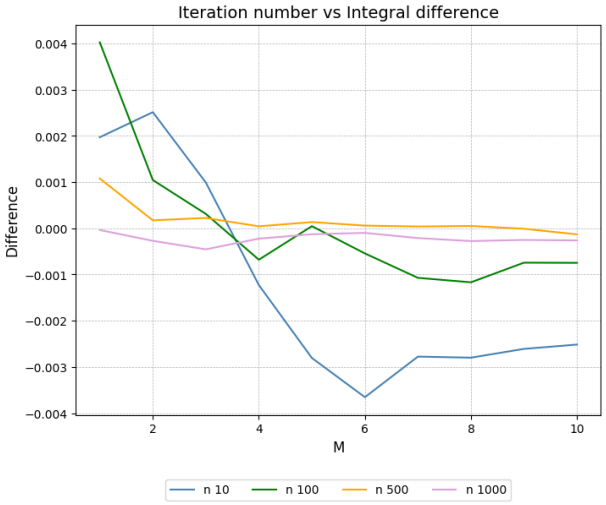

Figure 23: Optimal Forward LIS convergence for different n given α^q

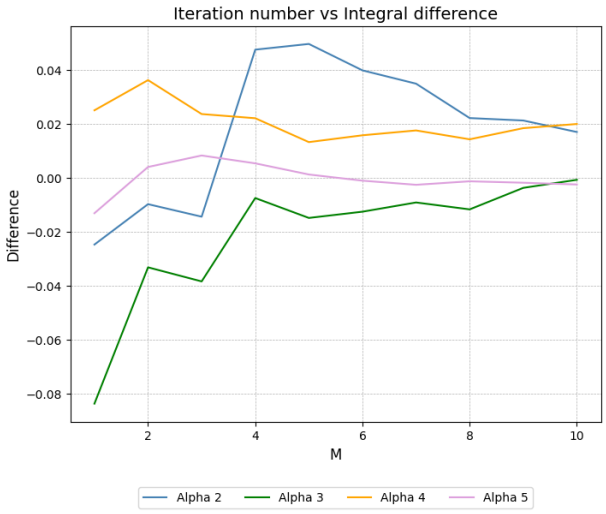

Figure 24: Optimal Backward LIS convergence for different α^q proposals

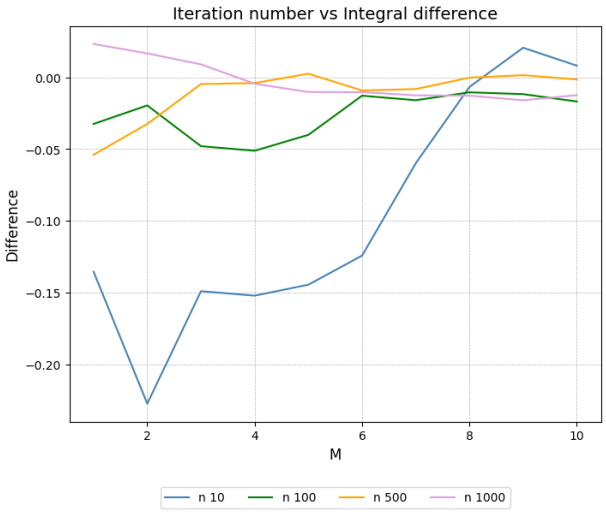

Figure 25: Optimal Backward LIS convergence for different n given α^q

We note that optimal bridge does affect LIS in a similar way to the geometric one, since the likeness in behaviour repsented in the quartets of plots. Moreover, we choose the triplet (4, 500, 6) and for (5, 500, 6) describing in order α^q, n and M respectively.
Finally, to determine the most effective strategy, we plot the difference between each approach using the optimum parameters and examine its computational cost as a function of M. The results are shown in the following figures:

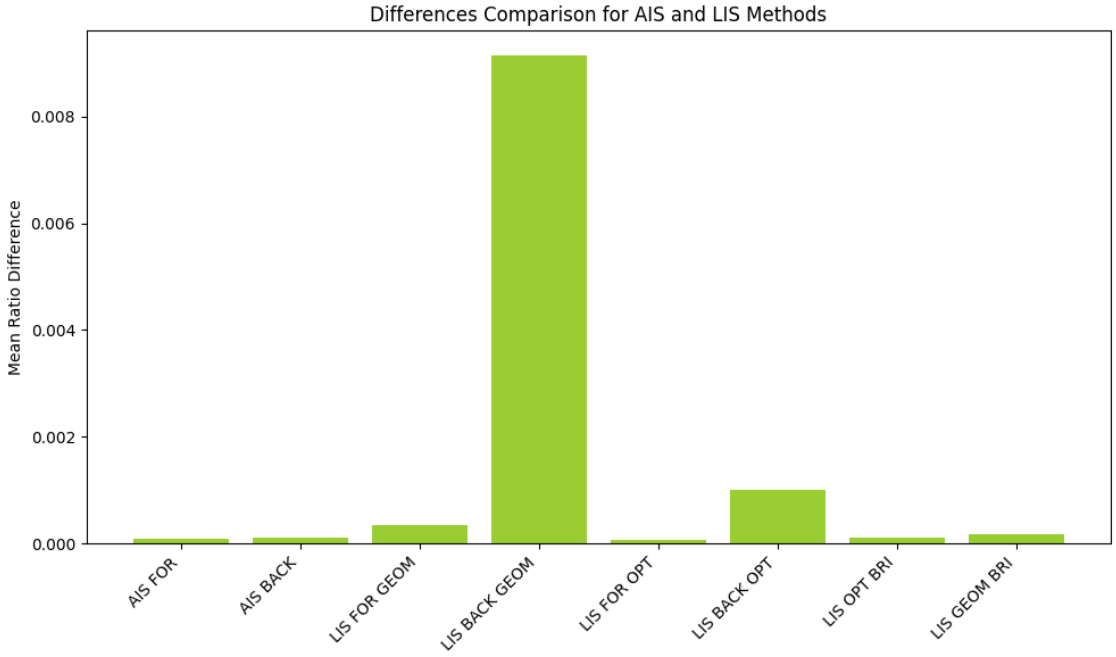

Figure 26: Strategies Comparisons

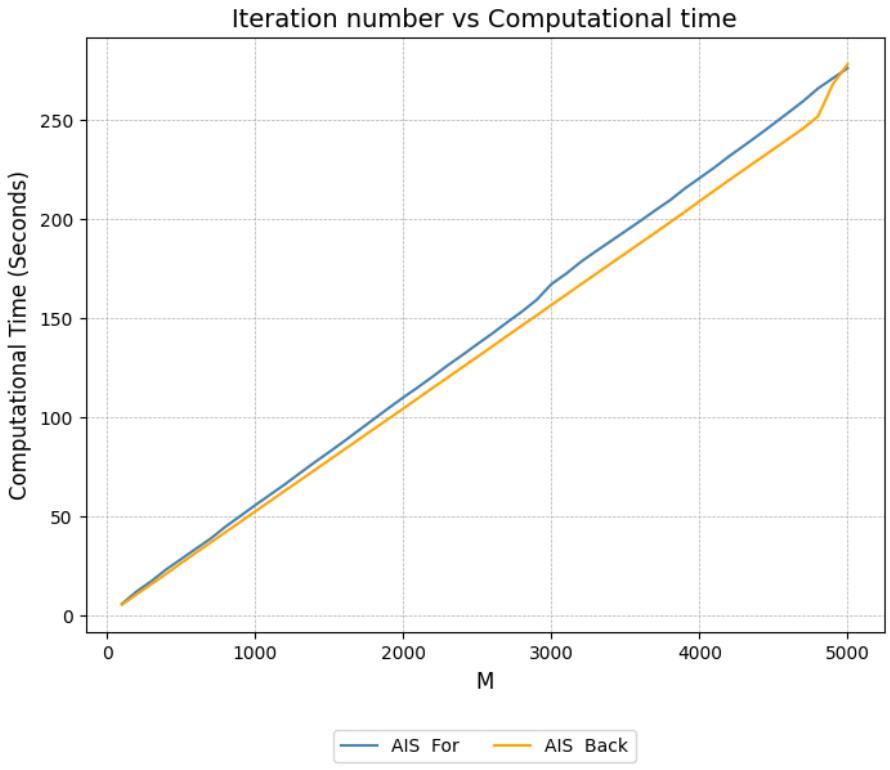

Figure 27: AIS Computational Costs

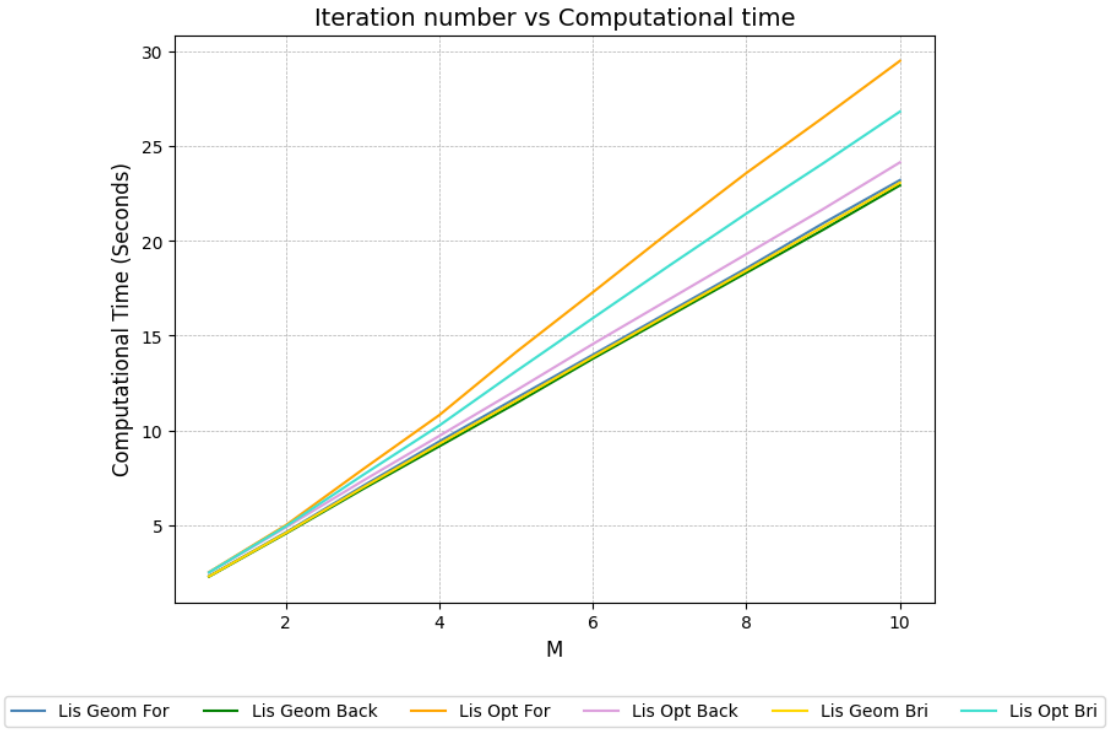

Figure 28: LIS Computational Costs

As we can observe, the best results are achieved using both the forward and backward AIS methods, as well as the LIS method with the optimal forward bridge approach. However, given the computational cost, which scales linearly in both cases, we opt for the Geometric Bridged LIS method in our case study. To further assess whether the newly chosen method offers additional benefits beyond unbiasedness, we compare it with a naïve Monte Carlo (MC) integration using importance sampling. We evaluate both methods in terms of computational time (in seconds, plotted on the X-axis) and the deviation of their outputs from the reference values (plotted on the Y-axis). The results are shown in figure (29). The plot clearly illustrates that Monte Carlo integration exhibits slower and less stable convergence to the reference values. By contrast, the Geometric Bridged LIS method demonstrates more accurate and consistent results for the same computational cost, which ultimately leads us to select it as the optimal method for our case study.

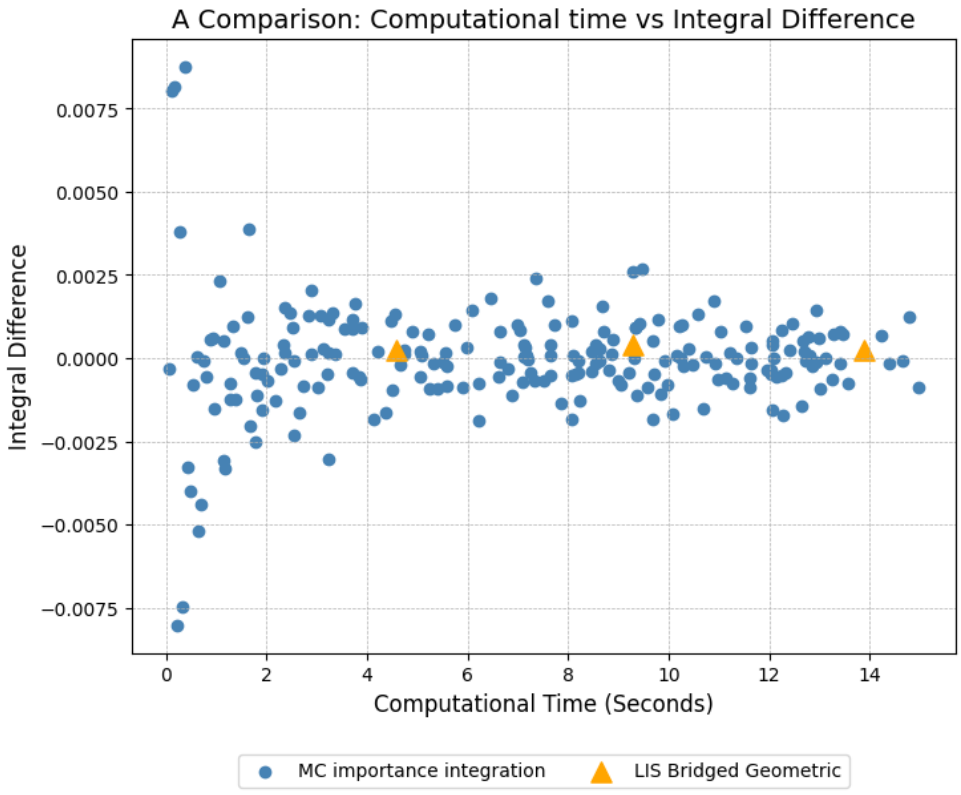

Figure 29: LIS Computational Costs

5 Conclusion

This thesis has explored the computational challenges associated with simulating the K-allele diffusion model in population genetics, with a particular focus on the approximation of normalizing constants using various integration techniques. Through the development and evaluation of both analytical and stochastic methods, this work aims to propose some contributions to the field by offering more efficient strategies for simulating allele frequency dynamics and computing key probabilistic quantities.

The main findings of this thesis highlight the strengths and weaknesses of different approaches. Nested analytical integration, while stable and theoretically sound, is computationally expensive, particularly as the dimensionality of the problem increases. The computational cost grows exponentially with the number of alleles, making this approach impractical for large populations or for models with many alleles. In contrast, Monte Carlo integration with Importance Sampling offers a scalable and computationally efficient alternative. By carefully selecting a proposal distribution, this method is able to focus computational effort on regions of the parameter space that contribute the most to the integral, significantly reducing the variance of the estimates and improving convergence. However, even Importance Sampling is not completely unbiased, as finding the optimal proposal distribution is for our problem not feasible. The inability to select this optimal distribution introduces bias into the estimates, requiring careful tuning of the proposal distribution, which may not always be affordable for have a time efficient simulation.

To overcome these limitations, we moved from methods that approximate a single normalizing constant to those designed to directly approximate the ratio of normalizing constants, which requires less tuning. Annealed Importance Sampling (AIS) introduces intermediate distributions to bridge the gap between the prior and posterior distributions, ensuring unbiased estimates even when the distributions are far apart. This method is particularly effective in high-dimensional settings where traditional importance sampling struggles. Similarly, Linked Importance Sampling (LIS), while maintaining the unbiasedness, provides a more stable and computationally efficient way to approximate these ratios, combining elements of AIS and bridge sampling. Both AIS and LIS require less tuning, making them practical tools for high-dimensional problems.

While the methods developed in this thesis provide significant improvements in terms of computational efficiency, there are limitations that must be acknowledged. While they do not show an heavy proposal tuning, even sophisticated methods like AIS introduce additional complexity, such as fine-tuning the number of intermediate distributions and transitions between them. Future research could focus on adaptive methods that automatically optimize the number of intermediate steps or proposal distributions, reducing the need for manual tuning.

It is important to note that this study applied these methods to a simplified version of the K-allele model, where mutations were not considered, in order to ensure comparability with previous studies. While this simplification allowed us to evaluate the computational trade-offs between different methods, future research could extend these findings by applying these techniques to more complex and complete versions of the model, such as those incorporating mutation dynamics. This would provide a deeper understanding of how evolutionary forces like mutation influence computational efficiency and accuracy, especially when approximating the normalizing constants.

In conclusion, this thesis tries to make some contributions to the field of population genetics by testing computational techniques for simulating genetic models and approximating normalizing constants. The methods developed here are not only applicable to the K-allele diffusion model but also have broader implications for Bayesian inference and Markov chain models in other areas of evolutionary biology and genetics. Future research could build upon these findings by exploring alternative sampling methods, such as Sequential Monte Carlo or adaptive importance sampling,

and by applying these techniques to more complex genetic models that incorporate additional evolutionary forces, such as recombination or migration.

References

[1] A. D. Barbour, S. N. Ethier, and R. C. Griffiths. "A transition function expansion for a diffusion model with selection". In: *The Annals of Applied Probability* 10.1 (2000), pp. 123–162.

[2] C. H. Bennett. "Efficient estimation of free energy differences from Monte Carlo data". In: *Journal of Computational Physics* 22 (1976), pp. 245–268.

[3] M. Chaleyat-Maurel and V. Genon-Catalot. "Computable infinite-dimensional filters with applications to discretized diffusion processes". In: *Stochastic Processes and their Applications* 116 (2006), pp. 1447–1467.

[4] Alan Genz and Paul Joyce. "Computation of the Normalization Constant for Exponentially Weighted Dirichlet Distribution Integrals". In: (2003).

[5] C. Jarzynski. "Nonequilibrium equality for free energy differences". In: *Physical Review Letters* 78 (1997), pp. 2690–2693.

[6] G. Kon Kam King et al. "Approximate filtering via discrete dual processes". In: *Stochastic Processes and their Applications* 168 (Feb. 2024). ISSN: 0304-4149. DOI: 10.1016/j.spa.2023.104268.

[7] Guillaume Kon Kam King, Omiros Papaspiliopoulos, and Matteo Ruggiero. "Exact inference for a class of hidden Markov models on general state spaces". In: *Electronic Journal of Statistics* 15 (Feb. 2021), pp. 2832–2875. ISSN: 0304-4149. DOI: https://doi.org/10.1214/21-EJS1841.

[8] X.-L. Meng and H. W. Wong. "Simulating ratios of normalizing constants via a simple identity: A theoretical exploration". In: *Statistica Sinica* 6 (1996), pp. 831–860.

[9] R. M. Neal. "Estimating Ratios of Normalizing Constants Using Linked Importance Sampling". In: (2005).

[10] O. Papaspiliopoulos and M. Ruggiero. "Optimal filtering and the dual process". In: *Bernoulli* 20.4 (2014), pp. 1999–2019.

[11] R. Y. Rubinstein and D. P. Kroese. "Simulation and the Monte Carlo Method". In: *John Wiley & Sons* 707 (2011).

[12] K. Sato. "Diffusion operators in population genetics and convergence of Markov chains". In: *Measure Theory Applications to Stochastic Analysis* 695 (1978), pp. 127–137.

[13] T. Shiga. "Diffusion processes in population genetics". In: *J. Math. Kyoto Univ.* 21 (1981), pp. 133–151.

[14] N. Shimakura. "Equations differentielles provenant de la genetique des populations". In: *Tohoku Math. J.* 29 (1977), pp. 287–318.

[15] S. Tavare. "Line-of-descent and genealogical processes, and their applications in population genetics models". In: *Theor. Popul. Biol.* 26 (1984), pp. 119–164.

[16] S. Wright. "Adaptation and selection". In: *Genetics, Paleontology, and Evolution* (1949), pp. 365–389.