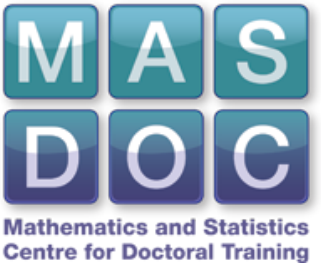


Statistical Inference for Diffusions in Genetics

by

Jaromir Sant

Thesis

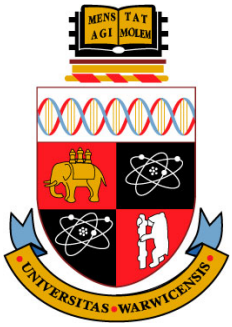
Submitted for the degree of

Doctor of Philosophy

Mathematics Institute

University of Warwick

December 2021



Contents

List of Figures

List of Tables

List of Algorithms

Acknowledgements

I would like to express my gratitude to my supervisors Dr. Paul Jenkins, Dr. Jere Koskela and Dr. Dario Spanò for their invaluable support and guidance over the past few years. I would like to thank Prof. Steven N. Evans and Prof. Adam Johansen for agreeing to be my examiners, Prof. Aleksandar Mijatovic for the helpful discussions and insights, as well as Ms. Deborah Walker and Ms. Alison Humphries for their prompt help and assistance in dealing with all my admin queries. To Hanson, Dom, Jacek (and Deus!!), Chris, Phil, Jack, Solly, Suzie, Ana, Alicia, Ian, Giorgos, Quirin, Osian and all the other postgrads and staff at both the Mathematics Institute and the Department of Statistics I've come across, thank you for making the whole experience all the more enjoyable. Thank you to James for sticking around despite the spatial discrepancy imposed by the past few years. And to Cinta, thank you for your unwavering support and unyielding patience with me. Last but not least, thank you to my parents for always being there for me no matter what.

Declarations

This thesis is submitted to the University of Warwick in support of my application for the degree of Doctor of Philosophy, and has not been previously submitted for any degree. The work presented here was conducted by the author in collaboration with Dr. Paul Jenkins, Dr. Jere Koskela and Dr. Dario Spanò, and is (to the best of my knowledge) original, unless otherwise stated.

Parts of Chapters 2 and 3 have been submitted for publication in [SJKS20].

Abstract

In this thesis we consider theoretical and practical aspects of conducting inference on data coming from the Wright–Fisher diffusion, which arises as the scaling limit of several discrete models used to describe the way in which allele frequencies change over time. This diffusion evolves on a bounded interval, and thus many standard results in diffusion theory assuming evolution on the entire real line do not apply.

Conditions ensuring the $\boldsymbol{\vartheta}$-uniform ergodicity of positively recurrent diffusions on bounded intervals with entrance or regular boundaries are established, and used to prove uniform in the selection and mutation parameters ergodicity for the Wright–Fisher case. The family of measures induced by the diffusion is further shown to be uniformly locally asymptotically normal, and these results are used to show the uniform (over compact sets in the parameter space) consistency, asymptotic normality, convergence of moments and asymptotic efficiency of the Maximum Likelihood and Bayesian estimators for the selection parameter in a continuous observation regime.

By appealing to a suitable state space augmentation and making use of the exact algorithm for the Wright–Fisher diffusion, we propose an exact Markov Chain Monte Carlo scheme which is able to directly target the joint posterior of the allele age and selection parameter. The method is subsequently tested on simulated data for a variety of prior distributions on both parameters.

Finally, a brief sketch of how $\boldsymbol{\vartheta}$-uniform ergodicity might be extended for the multidimensional Wright–Fisher diffusion is provided. The main techniques granting control over the rate of convergence in the ergodic theorem are developed, with a particular emphasis on why establishing such control is particularly challenging in the Wright–Fisher case.

Abbreviations

- aDNA - ancient DNA
- DNA - Deoxyribonucleic acid
- LAN - local asymptotic normality
- LHS - left hand side
- MCMC - Markov Chain Monte Carlo
- ML - Maximum Likelihood
- MLE - Maximum Likelihood estimator
- ODE - ordinary differential equation
- PIM - parent independent mutation
- RHS - right hand side
- SDE - stochastic differential equation

Chapter 1

Introduction

Over the past couple of decades, mathematical population genetics has been one of the main driving forces behind research in numerous areas of both mathematics and statistics. It primarily concerns itself with the study of how populations evolve over time, offering viable models to study how various biological phenomena such as selection and mutation interact and shape the genetic profile of the population they act upon. Many models have been proposed over the years to describe inheritance mechanisms between parents and offspring, but perhaps the most popular remains the Wright–Fisher model ([Wri31, Fis99]). In its simplest form, a haploid population of fixed size evolves in discrete generations, where the offspring inherit the type of the parent they choose uniformly at random from amongst the individuals present in the previous generation. Generalisations allowing for mutation, selection, migration, variable population size, overlapping generations, and various other genetically relevant phenomena are relatively easy to incorporate, making the process more realistic and thus appealing to practitioners.

Whilst formulating the Wright–Fisher model with mutation and selection is relatively straightforward, the granularity of the underlying process means that calculating any quantities of interest quickly becomes tedious and intractable. Thus it comes as no surprise that performing inference in this context is not feasible, and an alternative model approximating the original one needs to be sought. One way

to do this is to resort to a diffusion approximation where one rescales both space and time to recover a diffusive limit, commonly referred to as the Wright–Fisher diffusion ([Fel51, Kim64], and Chapter 10 in [EK86]), and which will be the main focus of this thesis. The Wright–Fisher diffusion is quite a robust process, in the sense that a broad class of Cannings models ([Can74]) converge to it when suitably scaled, thus making it an appealing process to study as it can be used to approximate any such underlying model. Besides taking values in the unit interval, it also has the favourable property that the only contribution to the diffusion coefficient comes from random mating whilst other mechanisms such as selection and mutation appear solely in the drift coefficient. This particular feature facilitates inference as it precludes any issues relating to identifying any parameters in the diffusion coefficient via the quadratic variation of the process. Thus one can concentrate solely on estimating the drift coefficient, treating the diffusion coefficient as a known expression.

In this thesis we aim to both establish theoretical guarantees for standard inferential techniques when applied to the Wright–Fisher diffusion, as well as provide a practicable inferential scheme with which to conduct exact inference. We start by considering the one-dimensional process, which is already quite well-understood ([Dur08, Chapters 7 and 8], [EK86, Chapter 10], [EM10, GJS18, Gri80, Gri79, Tav84, BEG00]) and widely used in the literature ([BYN, KPR21, SBS14, SES16, HDBY20b, MM13, Mat20]). However due to specific features of the process, standard results for scalar diffusions are not directly applicable. One specific instance of this is tackled in Chapter 3, where we show that the Maximum Likelihood (ML) and Bayesian estimators for the selection parameter have a set of desirable statistical properties in a continuous observation regime. These properties have been shown to hold for a broad class of diffusions defined on all of $\mathbb{R}$ in [Kut04], however the general theory there fails to hold for the Wright–Fisher diffusion. The main reason is the fact that the latter process has a diffusion coefficient which dies out at the boundaries (and is in fact equal to 0 at both 0 and

1), whilst in [Kut04] the diffusions considered are assumed to have a strictly positive diffusion coefficient everywhere. This assumption particularly highlights the discrepancy between diffusions defined on bounded intervals and those on all of $\mathbb{R}$, as the latter need to have some accompanying boundary behaviour ensuring that the process does not "escape" the bounded state space as it approaches the boundary. In the Wright–Fisher case, this is provided by the drift coefficient which becomes positive (or negative) as the process approaches 0 (or 1), provided the mutation parameters are strictly positive, whilst the diffusion coefficient vanishes sufficiently quickly to allow the process to be reflected back into the interior of the unit interval. This interplay between the drift and diffusion coefficient at the boundary has deeper implications on inferring the mutation parameters as will be highlighted in Chapter 3. Apart from the general case considered in [Kut04], we mention the work done by Watterson in [Wat79], where the maximum likelihood estimator (MLE) for selection in the absence of mutation (which in particular implies that the diffusion is absorbed at the boundary) is studied. By conditioning on absorption, the author derives the moment generating function, hypothesis tests and proves asymptotic normality, however these results do not extend to the case when the mutation rates are strictly positive as the boundaries now become reflecting.

To arrive to statements concerning the ML and Bayesian estimator however, we first show that the underlying process possesses certain properties which will be useful going forwards. To this end, in Chapter 2 we start by considering the more general setting of scalar diffusions defined on a bounded interval $[l, r]$, with $-\infty < l < r < \infty$ regular or entrance boundaries, and define the notion of $\boldsymbol{\vartheta}$-uniform ergodicity. This property extends the usual pointwise (in the parameter) ergodicity displayed by positively recurrent scalar diffusions to compact sets in the parameter space by finding a least rate at which the time averages of bounded measurable functions of the process converge to the state space average. Through the use of quantities that are very close to the ones used in classifying the boundary

behaviour of scalar diffusions, we provide verifiable criteria which allow us to deduce $\boldsymbol{\vartheta}$-uniform ergodicity for the above considered class of diffusions by making use of the regeneration arguments developed in [LLL11]. We are further able to extend this notion of ergodicity to a specific class of unbounded functions for the case of diffusions having solely entrance boundaries. These two general results admit the Wright–Fisher diffusion as a special case, as shown in Corollaries 2.5 and 2.6, and directly imply uniform local asymptotic normality for the family of measures induced by the solutions to the corresponding Wright–Fisher stochastic differential equations (SDEs), provided one considers a suitable class of initial laws.

These results will prove to be crucial in establishing certain statistical properties for the ML and Bayesian estimators for selection based on a continuously observed Wright–Fisher diffusion. As we are now in the regime where our data is a particular realisation from amongst the collection of all continuous functions mapping $[0, T] \mapsto [0, 1]$, the likelihood ratio function is given by the Radon–Nikodym derivative between the measures induced on this space of functions by the corresponding solutions to the SDE for different parameter configurations. This ratio can be analytically derived (provided the mutation parameters are sufficiently large) by the Girsanov transform, and shall be the main focus of Chapter 3. As stated above, a broad class of diffusions has already been tackled in [Kut04], establishing uniform (over compact sets in the parameter space) consistency, asymptotic normality and convergence of moments, together with asymptotic efficiency for a given set of loss functions. The arguments developed there hinge upon the Ibragimov-Has'minskii conditions ([IH81, Theorems I.5.1, I.5.2, I.10.1, I.10.2]), which consist of:

1. two bounds on the Hellinger distance of the likelihood ratio,
2. marginal convergence of the finite dimensional distributions of the likelihood ratio to those of a suitable limiting distribution
3. the uniqueness of the maximum of this limiting distribution to-

gether with that of the minimum of a particular function of this limiting quantity.

We show that these conditions directly imply the above stated properties for the ML and Bayesian estimators in Theorem 3.2, and prove that they hold true for the Wright–Fisher case in Propositions 3.5, 3.6, 3.7 and Corollary 3.4. The corollary follows directly from the previously proved uniform local asymptotic normality by considering the marginal convergence of the finite dimensional distributions of the associated likelihood ratio function, whilst Proposition 3.7 holds in general and is directly applicable for the Wright–Fisher diffusion. Propositions 3.5 and 3.6 require some more work. Extending this framework to include the mutation parameters is only possible when these are greater than or equal to 1 (to ensure that the Radon-Nikodỳm derivatives and resulting likelihood ratios are defined), and is rather delicate as illustrated in Section 3.4.

We point out here that the above described setup where we observe the whole path without error is clearly unrealistic, and thus cannot lead to a practicable inferential scheme. Even if one had access to the entire path, storing this infinite dimensional quantity on a machine is not possible. To this end we emphasise that the contributions developed in Chapter 3 are purely theoretical in nature and serve to establish a baseline from which one could hope to recover similar conclusions in the discrete observation setting, when an additional source of error is introduced as one does not know what happens to the path in between observations. As we show in Section 3.3, there is some empirical evidence which suggests that similar results might still hold in the case when observation times approach a densely sampled regime, however extending them formally for such a scenario is outside the scope of this thesis.

Having established that the selection parameter can be inferred from the data in this idealised setting, in Chapter 4 we proceed to develop an inferential technique for conducting inference based on noisy discrete observations coming from the scalar Wright–Fisher

diffusion. Recent developments in technologies relating to DNA sequencing and ancient DNA (aDNA) retrieval has led to a rapid increase in the amount of allele frequency times series datasets ([HSK+05, LPR+09, MM13, SCWGSL17, Mat20] to name a few, but we point out a growing repository of aDNA datasets at the Reich Lab website[1] in Harvard consisting of data used in published studies) which track the changes in allelic frequencies in a population across time. Traditional methods used by geneticists to quantify the degree of natural selection have relied solely on present day data, leading to estimates with a limited amount of statistical power since they depend only on a static snapshot of the population. In view of the fact that selection is a mechanism which is continuously operating on the population at hand, incorporating the temporal dimension into the picture by stringing together several observations through time (under the unifying assumption that all sampled individuals derive from the same population), should lead to more informative estimates. However, this approach does not come without its pitfalls. Whilst in the continuous observation case, the likelihood ratio is given analytically via the Girsanov theorem (with the nuisance there being the infinite dimensional random integrands involved), in the discrete observation case it is given by a product of Wright–Fisher transition densities. As with most diffusions, these quantities are analytically unavailable and cannot be computed exactly in finite time; in the neutral Wright–Fisher case the transition density admits an infinite series representation, whilst in the non-neutral case intractable terms make the analysis much harder. Most of the techniques developed thus far in the genetics time series literature have relied on some form of discretisation to deal with this intractability; be it solving the associated backward Kolmogorov equation via finite difference schemes ([BYN, HDBY20b]), working with truncated eigenfunction expansions of the transition densities ([SBS14]), or approximating Lebesgue integrals via Riemann sums ([SES16]). Such approximations allow for access to estimates of the selection coefficient as well as other genetically relevant quantities (such as allele age and effective

[1] https://reich.hms.harvard.edu/datasets

population size), however the discretisations and approximations employed introduce a bias into the inference which is impossible to quantify.

Notwithstanding these problems, it was proved in [JS17] that an exact algorithm targetting the neutral and non-neutral Wright–Fisher diffusion and bridge can be implemented in finite time, thereby paving the way to embedding this algorithm within a Markov Chain Monte Carlo (MCMC) framework to target the exact posterior of the allele age and selection parameter. By exact here we mean that the method suffers only from Monte Carlo (MC) and precision error as none of the intractable quantities are computed via discretised approximations. By relying upon a suitable state space augmentation, we gain access to a more tractable form of the likelihood function, which therefore enables us to construct a Gibbs sampler that alternates between updating the parameters of interest and the auxiliary random variables. The end result is a pseudo-marginal algorithm targetting the joint posterior of the allele age t_0, the selection coefficient σ, as well as the auxiliary variables which we can marginalise over to recover the posterior distributions of interest. The method has been tested extensively on simulated data (generated by means of the exact algorithm), and the output obtained suggests that the method produces reasonable results even when priors are somewhat mis-specified. The method developed in Chapter 4 is currently being extended in several directions, namely avoiding any pseudo-marginal updates, incorporating the mutation parameters into the inference, accounting for demographic history and allowing for selected alleles to arise from standing variation rather than necessarily stem from de novo mutations.

Working in the one-dimensional case as above allows for a number of techniques and properties of one-dimensional processes to be invoked, leading to simpler calculations and arguments. These simplifications disappear when one moves to higher dimensions, as several concepts such as point recurrence and the notions of speed and scale do not extend. Thus establishing analogous results as in

the one dimensional case becomes much more intricate. In particular, in Chapter 5 we illustrate how the results from Chapter 2 can be extended for the K-dimensional Wright–Fisher diffusion with selection and mutation when $K \geq 2$, starting with $\boldsymbol{\vartheta}$-uniform ergodicity. Whilst extending the notion of $\boldsymbol{\vartheta}$-uniform ergodicity is quite straightforward, establishing it is rather more involved. The regenerative scheme coupled with the ordinary differential equation (ODE) approach used in Theorem 2.2 is no longer suitable, as the hitting times of sets (as opposed to points) do not offer the correct mathematical framework within which to entertain regenerative arguments in higher dimensions. Instead we show (following the approach in [LL13]) how a richer stochastic process constructed out of the original one allows for a suitable set of regeneration times to be defined, which in turn allows for regeneration arguments similar to those used in Chapter 2 to be employed. We point out here that the techniques used in [LL13] enable control over the rate of convergence in the ergodic theorem first by means of the moments of these regeneration times (via the arguments used in Theorem 5.2 therein) and subsequently via a suitable Lyapunov function (by making use of Theorem 4.1 therein). Whilst the results leading to the bounds involving the moments of the regeneration times apply verbatim to the Wright–Fisher case, showing how these can be controlled via Lyapunov functions remains an open problem, as the results used for this last step need not apply to the Wright–Fisher case. Nonetheless, we show how by considering the hitting times of a particular set (together with a specific choice of parameter configurations), standard multidimensional diffusion theory coupled with a suitable choice of Lyapunov function allow for control over the moments of said hitting time in terms of the Lyapunov functions. We emphasise here that this chapter does not contain original results, but rather provides a general overview of the problems involved in establishing $\boldsymbol{\vartheta}$-uniform ergodicity for the K-dimensional Wright–Fisher diffusion. Once this last hurdle is cleared, then one can start thinking about proving uniform local asymptotic normality for the corresponding family of measures, and eventually also looking into extending the results in Chapter 3 to the multidimensional setting.

The rest of this thesis is organised as follows: In Chapter 2 we tackle the problem of deriving conditions which guarantee $\boldsymbol{\vartheta}$-uniform ergodicity for bounded scalar diffusions with entrance or regular boundaries, and proceed to extend this notion for a specific class of unbounded functions in the case when both boundaries are entrance. We then use these results to deduce the uniform in the selection and mutation parameters ergodicity of the Wright–Fisher diffusion together with the uniform local asymptotic normality of the laws associated to the solutions of this SDE. Chapter 3 then looks at proving uniform (over compact sets in the parameter space) consistency, asymptotic normality and convergence of moments, together with asymptotic efficiency for the ML and Bayesian estimators for selection when the entire diffusion trajectory is observed without error in Theorem 3.2. The results rely on the Ibragimov–Has'minskii conditions which are translated to our setting in Theorem 3.3, and proved for the Wright–Fisher case in Section 3.2. Problems associated with extending the inferential framework to include the mutation parameters are elaborated on at the end of the chapter. In Chapter 4 we then turn towards devising a practicable way of performing inference on time-series data driven by an underlying Wright–Fisher diffusion. Here we develop an exact inferential scheme which allows us to directly target the posterior over the allele age and selection parameter via a suitable state space augmentation and access to the exact algorithm for the Wright–Fisher diffusion. We give a thorough explanation of the method, and proceed to illustrate its performance on simulated data before closing off with some final remarks on extensions to the current setup. We then move on to tackle the case of the K-dimensional Wright–Fisher in Chapter 5, where we illustrate the steps necessary to prove $\boldsymbol{\vartheta}$-uniform ergodicity in this setting, and give a brief sketch of how one might achieve this. In addition, we explain in greater detail the main problems relating to controlling the moments of regeneration times, and describe some preliminary attempts at solving this. We conclude with a brief discussion in Chapter 6, where we survey the results presented here and provide an outlook on related future work.

Chapter 2

$\boldsymbol{\vartheta}$-uniform ergodicity and local asymptotic normality

In this chapter we focus on the scalar Wright–Fisher diffusion which describes the allele frequency dynamics in a two-allele, haploid population undergoing both selection and mutation. We start by considering a general scalar diffusion taking values in an arbitrary bounded interval $[l, r]$, with $-\infty < l < r < \infty$ being either regular or entrance, and derive verifiable criteria to establish $\boldsymbol{\vartheta}$-uniform ergodicity (as defined in Definition 2.1). Subsequently, we introduce the Wright–Fisher diffusion, and by making use of the previously derived criteria, show that the diffusion is ergodic uniformly over both the selection and mutation parameters, and that the associated family of measures induced by the solutions to the corresponding stochastic differential equations (SDEs) is uniformly locally asymptotically normal (provided the mutation parameters are greater than or equal to 1). The importance of these two properties (particularly uniform local asymptotic normality) becomes more apparent in the next chapter when we use them to analyse the statistical properties of the Maximum Likelihood and Bayesian estimators for selection.

Using arguments developed in [LLL11], we bound the rate of convergence in the ergodic theorem for bounded positive recurrent scalar diffusions having either boundary being entrance or regular,

via moments of the hitting times of an arbitrary point in the interior of the state space. By making use of recursively defined ordinary differential equations (ODEs), these quantities can be bounded from above in terms of the underlying parameter $\boldsymbol{\vartheta}$. Thus, the pointwise (in the parameter) ergodicity is extended to arbitrary compact sets in the parameter space by finding the least rate at which this convergence occurs. We point out here that the general results we derive in Section 2.1 provide easily verifiable criteria with which to deduce whether an arbitrary diffusion on a bounded interval with entrance or regular boundaries displays $\boldsymbol{\vartheta}$-uniform ergodicity. In particular, as seen in the proof of Theorem 2.2, the criteria are closely linked to standard quantities used to classify the boundary behaviour of scalar diffusions. We further extend the notion of $\boldsymbol{\vartheta}$-uniform ergodicity for a specific class of unbounded functions when the diffusion has solely entrance boundaries.

Local asymptotic normality ([LC60]) is a particularly useful property for a family of statistical models to possess, as it allows for the log-likelihood ratio to be asymptotically viewed as a Gaussian random variable. This concept turns out to be crucial in proving several asymptotic properties of estimators such as consistency and asymptotic normality in the context of parametric models, and shall be one of the central properties upon which our results in Chapter 3 rely. Establishing uniform local asymptotic normality for a general class of diffusions taking values on the entire real line is rather straightforward given sufficient assumptions on the drift and diffusion coefficient (see Lemma 2.9 in [Kut04]), by applying a uniform (in the parameter) central limit theorem (Proposition 1.20 in [Kut04]) and a law of large numbers (by making use of Proposition 1.18 in [Kut04]). The latter proposition, however, holds only for SDEs for which the diffusion coefficient's inverse can be bounded from above by a polynomial, and thus immediately excludes the Wright–Fisher diffusion. Nonetheless, by resorting to $\boldsymbol{\vartheta}$-uniform ergodicity, we are able to establish uniform local asymptotic normality for the class of Wright–Fisher diffusions parametrised by the selection and mutation parameters (provided the latter are greater than or equal to 1) in Theorem 2.8.

The rest of this chapter is organised as follows: in Section 2.1 we derive the conditions which ensure that a scalar diffusion on a bounded interval with entrance or regular boundaries is $\boldsymbol{\vartheta}$-uniformly ergodic (a term we define precisely in Definition 2.1), and extend this property to a specific class of unbounded functions for diffusions with solely entrance boundaries. Section 2.2 then introduces the scalar Wright–Fisher diffusion, together with some well-known properties, before proving that the diffusion is uniformly in the selection and mutation parameters ergodic in Subsection 2.2.1, and that the family of measures (indexed by $\boldsymbol{\vartheta} \in \boldsymbol{\Theta} \subset \mathbb{R} \times [1, \infty)^2$ for $\boldsymbol{\Theta}$ open and bounded) induced by the solutions to the corresponding SDEs are uniformly locally asymptotically normal in Subsection 2.2.2.

2.1 ϑ-uniform ergodicity for scalar diffusions on bounded intervals

We start by considering an arbitrary fixed interval $[l, r]$, with $-\infty < l < r < \infty$, on which we define the SDE

$$dY_t = \mu(\boldsymbol{\vartheta}, Y_t)dt + \alpha(Y_t)dW_t, \quad Y_0 \sim \nu, \quad \boldsymbol{\vartheta} \in \boldsymbol{\Theta} \subseteq \mathbb{R}^d, d \geq 1 \tag{2.1}$$

where ν is an arbitrary initial distribution on $[l, r]$, $(W_t)_{t \geq 0}$ a standard Wiener process defined on a given filtered probability space, μ and α are such that the SDE admits a unique strong solution which we denote by $Y := (Y_t)_{t \geq 0}$, $-\infty < l < r < \infty$ are both either entrance or regular boundaries for Y, and the observation interval is fixed to $[0, T]$. We denote by $\mathbb{P}_\nu^{(\boldsymbol{\vartheta})}$ the law induced on the space of continuous functions mapping $[0, T]$ into $[l, r]$ (endowed with the Borel σ-algebra, and henceforth denoted by $C_T([l, r])$) by the solution to (2.1) when the true diffusion parameter is set to $\boldsymbol{\vartheta}$, and $Y_0 \sim \nu$ (with dependence on T being implicit). Furthermore we denote taking expectation with respect to $\mathbb{P}_\nu^{(\boldsymbol{\vartheta})}$ by $\mathbb{E}_\nu^{(\boldsymbol{\vartheta})}$.

Assume further that Y is positive recurrent, then using stan-

dard one-dimensional diffusion theory (see Theorem 1.16 in [Kut04]), we get that the unique invariant density is given by

$$f^Y_{\boldsymbol{\vartheta}}(x) = \frac{1}{G^Y_{\boldsymbol{\vartheta}}} \frac{2}{\alpha^2(x)} e^{2\int^x \frac{\mu(\boldsymbol{\vartheta},z)}{\alpha^2(z)}dz}, \qquad x \in [l,r],$$
$$G^Y_{\boldsymbol{\vartheta}} := \int_l^r \frac{2}{\alpha^2(x)} e^{2\int^x \frac{\mu(\boldsymbol{\vartheta},z)}{\alpha^2(z)}dz} dx. \tag{2.2}$$

In what follows, we denote taking expectation with respect to $f^Y_{\boldsymbol{\vartheta}}$ by $\mathbb{E}^{(\boldsymbol{\vartheta})}$, where the omission of the subscript will indicate that we start from stationarity, and henceforth always assume that $\xi \sim f^Y_{\boldsymbol{\vartheta}}$.

In order to derive the results in Chapter 3, we will need a slightly stronger notion of ergodicity which we now define. The idea here is that we can extend pointwise ergodicity in the parameter $\boldsymbol{\vartheta}$ to any compact set $\mathcal{K} \subset \boldsymbol{\Theta}$ by finding the slowest rate of convergence which works within that compact set. More rigorously, we introduce the following definition.

Definition 2.1. A process Y is said to be *ergodic uniformly in the parameter* $\boldsymbol{\vartheta}$ (or $\boldsymbol{\vartheta}$-uniformly ergodic) if $\forall \varepsilon > 0$ we have that

$$\lim_{T\to\infty} \sup_{\boldsymbol{\vartheta}\in\mathcal{K}} \mathbb{P}^{(\boldsymbol{\vartheta})}_{\nu}\left[\left| \frac{1}{T}\int_0^T h(Y_t)dt - \mathbb{E}^{(\boldsymbol{\vartheta})}\left[h\left(\xi\right)\right] \right| > \varepsilon \right] = 0 \tag{2.3}$$

holds for any $\mathcal{K}$ compact subset of the parameter space, and for any function $h : [l,r] \to \mathbb{R}$ bounded and measurable, where $\xi \sim f^Y_{\boldsymbol{\vartheta}}$.

In the context of scalar diffusions defined on a bounded interval $[l,r]$ with $-\infty < l < r < \infty$, where both boundaries are either regular or entrance we have the following theorem:

Theorem 2.2. *Let Y be defined as above as the solution to* (2.1)*, with boundary points l and r either entrance or regular, and that the*

expressions

$$\kappa^l_{\boldsymbol{\vartheta}}(a,b) := \int_a^b e^{-\int^\xi \frac{2\mu(\boldsymbol{\vartheta},y)}{\alpha^2(y)}dy} \int_l^\xi \frac{2}{\alpha^2(\eta)} e^{\int^\eta \frac{2\mu(\boldsymbol{\vartheta},y)}{\alpha^2(y)}dy} d\eta d\xi,$$
$$\kappa^r_{\boldsymbol{\vartheta}}(a,b) := \int_a^b e^{-\int^\xi \frac{2\mu(\boldsymbol{\vartheta},y)}{\alpha^2(y)}dy} \int_\xi^r \frac{2}{\alpha^2(\eta)} e^{\int^\eta \frac{2\mu(\boldsymbol{\vartheta},y)}{\alpha^2(y)}dy} d\eta d\xi, \tag{2.4}$$

are bounded away from 0 on any compact set $\mathcal{K} \subset \boldsymbol{\Theta}$, and any $l < a < b < r$. Then Y is uniformly ergodic in the parameter $\boldsymbol{\vartheta}$ for any initial distribution ν.

Proof. We show $\boldsymbol{\vartheta}$-uniform ergodicity for scalar diffusions on the bounded interval $[l, r]$ having entrance or regular boundary points by making use of Theorem 3.2 in [LLL11], which allows us to bound the LHS of (2.3) in terms of the moments of the hitting times of the process. That result requires the diffusion coefficient to be positive everywhere, and the drift and diffusion coefficients to be locally Lipschitz and to satisfy a linear growth condition. These conditions however, are only used to guarantee the existence of a unique strong non-exploding solution to the SDE in Theorem 3.2, which we are guaranteeing explicitly in the statement of the theorem. None of these requirements on the drift and diffusion coefficients are used in the proof of Theorem 3.2 in [LLL11] when $p \in \{2, 3, \dots\}$, which allows us to employ this theorem for such p. All that remains to prove then is that these moments can be bounded in $\boldsymbol{\vartheta}$ over compact sets in the parameter space, for then (2.3) holds. To this end, we introduce some notation from [LLL11], namely let $a, b \in (l, r)$ be arbitrary fixed points such that $a < b$. Define $S_0 = 0$, $R_0 = 0$, and

$$S_{n+1} := \inf\{t \geq R_n : Y_t = b\}$$
$$R_{n+1} := \inf\{t \geq S_{n+1} : Y_t = a\}$$

for $n \in \mathbb{N}$, where we specify that here and throughout the rest of this thesis $\mathbb{N}$ includes 0. By the strong Markov property, $(R_k - R_{k-1})_{k \in \mathbb{N}\setminus\{0\}}$ is an i.i.d. sequence with law under $\mathbb{P}^{(\boldsymbol{\vartheta})}_\nu$ equal to the law of R_1 under $\mathbb{P}^{(\boldsymbol{\vartheta})}_a$, where $\mathbb{P}^{(\boldsymbol{\vartheta})}_a$ denotes the law of the process started from a. Related

to the process $(R_n)_{n\in\mathbb{N}}$ we have the process $(N_t)_{t\geq 0}$ which we define as

$$N_t := \sup\{n : R_n \leq t\}$$

and for which we observe that $\{N_t \geq n\} = \{R_n \leq t\}$. We also denote by

$$T_a := \inf\{t \geq 0 : Y_t = a\}$$

the hitting time of a, and note that $T_b = S_1$. Furthermore, let $\ell_{\boldsymbol{\vartheta}} := \mathbb{E}^{(\boldsymbol{\vartheta})}[N_1] = \mathbb{E}_a^{(\boldsymbol{\vartheta})}[R_1]^{-1}$ (see Lemma 2.7 in [LLL11]), and $\bar{\eta}_1 := -(R_2 - R_1 - \ell_{\boldsymbol{\vartheta}}^{-1})$. Then Theorem 3.2 in [LLL11] gives us that for $p \in \{2, 3, \dots\}$

$$\mathbb{P}_\nu^{(\boldsymbol{\vartheta})}\left[\left|\frac{1}{T}\int_0^T h(Y_t)dt - \mathbb{E}^{(\boldsymbol{\vartheta})}\left[h(\xi)\right]\right| > \varepsilon\right] \leq K(\boldsymbol{\vartheta}, Y, p)\varepsilon^{-p}\|h\|_\infty^p T^{-\frac{p}{2}},$$

where

$$\begin{aligned} K(\boldsymbol{\vartheta}, Y, p) := {} & 6^{\frac{p}{2}}\mathbb{E}_\nu^{(\boldsymbol{\vartheta})}\left[R_1^{\frac{p}{2}}\right] + 12^p C_p \ell_{\boldsymbol{\vartheta}}^{\frac{p}{2}}\mathbb{E}_\nu^{(\boldsymbol{\vartheta})}\left[|R_2 - R_1|^p\right] \\ & + 2(6^p)\ell_{\boldsymbol{\vartheta}}\mathbb{E}_a^{(\boldsymbol{\vartheta})}\left[R_1^p\right] + 2^{\frac{p}{2}}\mathbb{E}_\nu^{(\boldsymbol{\vartheta})}\left[\left|R_1 - \ell_{\boldsymbol{\vartheta}}{}^{-1}\right|^{\frac{p}{2}}\right] \\ & + 2^{\frac{3p}{2}} C_p \ell_{\boldsymbol{\vartheta}}^{\frac{p}{2}}\mathbb{E}_\nu^{(\boldsymbol{\vartheta})}\left[|\bar{\eta}_1|^p\right], \end{aligned}$$

and C_p is a constant depending only on p. We point out here that Theorem 3.2 in [LLL11] holds $\forall p \in (1, \infty)$ under additional assumptions, but for our case we need only $p \in \{2, 3, \dots\}$. Thus we are left with showing these moments can be bounded from above in $\boldsymbol{\vartheta}$ over compact sets, for then (2.3) follows. Now the only terms above that depend on $\boldsymbol{\vartheta}$ are

$$\begin{aligned} &\mathbb{E}_\nu^{(\boldsymbol{\vartheta})}\left[R_1^{\frac{p}{2}}\right], \qquad \ell_{\boldsymbol{\vartheta}}^{\frac{p}{2}}\mathbb{E}_\nu^{(\boldsymbol{\vartheta})}\left[|R_2 - R_1|^p\right], \qquad \ell_{\boldsymbol{\vartheta}}\mathbb{E}_a^{(\boldsymbol{\vartheta})}\left[R_1^p\right], \\ &\mathbb{E}_\nu^{(\boldsymbol{\vartheta})}\left[\left|R_1 - \ell_{\boldsymbol{\vartheta}}^{-1}\right|^{\frac{p}{2}}\right], \qquad \ell_{\boldsymbol{\vartheta}}^{\frac{p}{2}}\mathbb{E}_\nu^{(\boldsymbol{\vartheta})}\left[|\bar{\eta}_1|^p\right] \end{aligned}$$

and in light of the following inequalities

$$\begin{aligned}
\mathbb{E}_{\nu}^{(\boldsymbol{\vartheta})}\left[|\bar{\eta}_1|^p\right] &\leq 2^{p-1}\left(\mathbb{E}_{\nu}^{(\boldsymbol{\vartheta})}\left[|R_2 - R_1|^p\right] + \mathbb{E}_{\nu}^{(\boldsymbol{\vartheta})}\left[\ell_{\boldsymbol{\vartheta}}^{-p}\right]\right) \\
&= 2^{p-1}\left(\mathbb{E}_{a}^{(\boldsymbol{\vartheta})}\left[R_1^p\right] + \mathbb{E}_{a}^{(\boldsymbol{\vartheta})}\left[R_1\right]^p\right), \\
\mathbb{E}_{\nu}^{(\boldsymbol{\vartheta})}\left[\left|R_1 - \ell_{\boldsymbol{\vartheta}}^{-1}\right|^{\frac{p}{2}}\right] &\leq 2^{\frac{p}{2}-1}\left(\mathbb{E}_{\nu}^{(\boldsymbol{\vartheta})}\left[R_1^{\frac{p}{2}}\right] + \mathbb{E}_{\nu}^{(\boldsymbol{\vartheta})}\left[\ell_{\boldsymbol{\vartheta}}^{-\frac{p}{2}}\right]\right) \\
&= 2^{\frac{p}{2}-1}\left(\mathbb{E}_{\nu}^{(\boldsymbol{\vartheta})}\left[R_1^{\frac{p}{2}}\right] + \mathbb{E}_{a}^{(\boldsymbol{\vartheta})}\left[R_1\right]^{\frac{p}{2}}\right), \\
\mathbb{E}_{\nu}^{(\boldsymbol{\vartheta})}\left[|R_2 - R_1|^p\right] = \mathbb{E}_{a}^{(\boldsymbol{\vartheta})}\left[R_1^p\right] &\leq 2^{p-1}\left(\mathbb{E}_{a}^{(\boldsymbol{\vartheta})}\left[T_b^p\right] + \mathbb{E}_{b}^{(\boldsymbol{\vartheta})}\left[T_a^p\right]\right), \\
\mathbb{E}_{\nu}^{(\boldsymbol{\vartheta})}\left[R_1^{\frac{p}{2}}\right] &\leq 2^{\frac{p}{2}-1}\left(\mathbb{E}_{\nu}^{(\boldsymbol{\vartheta})}\left[T_b^{\frac{p}{2}}\right] + \mathbb{E}_{b}^{(\boldsymbol{\vartheta})}\left[T_a^{\frac{p}{2}}\right]\right), \\
\mathbb{E}_{a}^{(\boldsymbol{\vartheta})}\left[R_1\right] &= \mathbb{E}_{a}^{(\boldsymbol{\vartheta})}\left[T_b\right] + \mathbb{E}_{b}^{(\boldsymbol{\vartheta})}\left[T_a\right],
\end{aligned}$$

it suffices to consider only the terms $\ell_{\boldsymbol{\vartheta}}$ and $\mathbb{E}_{\nu}^{(\boldsymbol{\vartheta})}\left[T_b^p\right]$. Thus we are left with showing that these two terms can be bounded from above in $\boldsymbol{\vartheta}$ over any compact set $\mathcal{K} \subset \boldsymbol{\Theta}$. We further point out that we can reduce our considerations in the expressions above to integer moments, for if this is not the case then

$$\mathbb{E}_{\nu}^{(\boldsymbol{\vartheta})}\left[T_b^p\right] \leq \mathbb{E}_{\nu}^{(\boldsymbol{\vartheta})}\left[T_b^{\lceil p \rceil}\right] + \mathbb{E}_{\nu}^{(\boldsymbol{\vartheta})}\left[T_b^{\lfloor p \rfloor}\right]$$

where $\lceil \cdot \rceil$ and $\lfloor \cdot \rfloor$ denote the ceiling and floor functions respectively.

We make use of the backward equation for the quantity $U_{q,b}(x) := \mathbb{E}_x^{(\boldsymbol{\vartheta})}[T_b^q]$ for $q \in \{1, 2, \dots\}$, to derive the ODE (as can be found in [KT81] p. 203 and 210, and [WY08])

$$\frac{\alpha^2(x)}{2} U''_{q,b}(x) + \mu(\boldsymbol{\vartheta}, x) U'_{q,b}(x) + q U_{q-1,b}(x) = 0 \tag{2.5}$$

with boundary conditions $U_{q,b}(b) = 0$ and

$$\lim_{y \to l} S'(y)^{-1} \frac{\partial}{\partial y} U_{q,b}(y) = 0$$

when $x < b$, or

$$\lim_{y \to r} S'(y)^{-1} \frac{\partial}{\partial y} U_{q,b}(y) = 0$$

when $x > b$, where

$$S(x) := \int^x e^{-\int^y \frac{2\mu(z)}{\alpha^2(z)}dz} dy.$$

Solving (2.5) for $x < b$ leads to

$$\mathbb{E}_x^{(\boldsymbol{\vartheta})}[T_b^q] = \int_x^b e^{-\int^\xi \frac{2\mu(\boldsymbol{\vartheta},y)}{\alpha^2(y)}dy} \int_l^\xi \frac{2}{\alpha^2(\eta)} e^{\int^\eta \frac{2\mu(\boldsymbol{\vartheta},y)}{\alpha^2(y)}dy} q U_{q-1,b}(\eta) d\eta d\xi, \quad (2.6)$$

whilst for $x > b$ we have that

$$\mathbb{E}_x^{(\boldsymbol{\vartheta})}[T_b^q] = \int_b^x e^{-\int^\xi \frac{2\mu(\boldsymbol{\vartheta},y)}{\alpha^2(y)}dy} \int_\xi^r \frac{2}{\alpha^2(\eta)} e^{\int^\eta \frac{2\mu(\boldsymbol{\vartheta},y)}{\alpha^2(y)}dy} q U_{q-1,b}(\eta) d\eta d\xi. \quad (2.7)$$

We claim that for any $x < b$ and any $q \in \{1, 2, \dots\}$,

$$\begin{aligned}\mathbb{E}_x^{(\boldsymbol{\vartheta})}[T_b^q] &\leq q! \left(\int_l^b e^{-\int^\xi \frac{2\mu(\boldsymbol{\vartheta},y)}{\alpha^2(y)}dy} \int_l^\xi \frac{2}{\alpha^2(\eta)} e^{\int^\eta \frac{2\mu(\boldsymbol{\vartheta},y)}{\alpha^2(y)}dy} d\eta d\xi \right)^q \\ &= q! \kappa_{\boldsymbol{\vartheta}}^l(l,b)^q < \infty. \end{aligned} \quad (2.8)$$

To see this, observe that

$$\begin{aligned}\mathbb{E}_x^{(\boldsymbol{\vartheta})}[T_b] &= \int_x^b e^{-\int^\xi \frac{2\mu(\boldsymbol{\vartheta},y)}{\alpha^2(y)}dy} \int_l^\xi \frac{2}{\alpha^2(\eta)} e^{\int^\eta \frac{2\mu(\boldsymbol{\vartheta},y)}{\alpha^2(y)}dy} d\eta d\xi \\ &\leq \int_l^b e^{-\int^\xi \frac{2\mu(\boldsymbol{\vartheta},y)}{\alpha^2(y)}dy} \int_l^\xi \frac{2}{\alpha^2(\eta)} e^{\int^\eta \frac{2\mu(\boldsymbol{\vartheta},y)}{\alpha^2(y)}dy} d\eta d\xi \\ &= \kappa_{\boldsymbol{\vartheta}}^l(l,b), \end{aligned} \quad (2.9)$$

and we observe that $\kappa_{\boldsymbol{\vartheta}}^l(l,b)$ is finite for all $\boldsymbol{\vartheta} \in \boldsymbol{\Theta}$ in virtue of l being either an entrance or regular boundary (see Table 6.2 in [KT81, Chapter 15, Section 6] p. 234, and note that $\kappa_{\boldsymbol{\vartheta}}^l(l,b)$ here corresponds to $N(l)$ as defined in (6.19) there). Observe that the RHS of (2.9) is independent of x, so we can use the recursion in (2.7) to conclude by induction that (2.8) holds for $q \in \{1, 2, \dots\}$ as required. Similar arguments to those presented above coupled with the requirement that the boundary point at r is either entrance or regular, allows us to conclude that for $x > b$

and $q \in \{1, 2, \dots\}$,

$$\mathbb{E}_x^{(\boldsymbol{\vartheta})}[T_b^q] \leq q! \left(\int_b^r e^{-\int^\xi \frac{2\mu(\boldsymbol{\vartheta},y)}{\alpha^2(y)} dy} \int_\xi^r \frac{2}{\alpha^2(\eta)} e^{\int^\eta \frac{2\mu(\boldsymbol{\vartheta},y)}{\alpha^2(y)} dy} d\eta d\xi \right)^q$$
$$= q! \kappa_{\boldsymbol{\vartheta}}^r(b,r)^q < \infty. \tag{2.10}$$

Both RHS of (2.8) and (2.10) are independent of x, so trivially

$$\mathbb{E}_\nu^{(\boldsymbol{\vartheta})}[T_b^q] \leq q! \left(\kappa_{\boldsymbol{\vartheta}}^l(l,b)^q + \kappa_{\boldsymbol{\vartheta}}^r(b,r)^q \right). \tag{2.11}$$

All the terms on the RHS of (2.8), (2.10) and (2.11) are finite for any $\boldsymbol{\vartheta} \in \boldsymbol{\Theta}$, so we have our required bound when taking the supremum over a compact set $\mathcal{K} \subset \boldsymbol{\Theta}$ for $\mathbb{E}_\nu^{(\boldsymbol{\vartheta})}[T_b^q]$. It remains to show that we can bound $\ell_{\boldsymbol{\vartheta}}$ from above. Observe that by definition

$$\ell_{\boldsymbol{\vartheta}} = \mathbb{E}_a^{(\boldsymbol{\vartheta})}[R_1]^{-1} = \left(\mathbb{E}_a^{(\boldsymbol{\vartheta})}[T_b] + \mathbb{E}_b^{(\boldsymbol{\vartheta})}[T_a] \right)^{-1},$$

and recall that we will take the supremum in $\boldsymbol{\vartheta}$ over a given compact set $\mathcal{K}$. Using (2.6) and (2.7) with $q = 1$, coupled with (2.4), we deduce that $\mathbb{E}_a^{(\boldsymbol{\vartheta})}[T_b]$ and $\mathbb{E}_b^{(\boldsymbol{\vartheta})}[T_a]$ are bounded away from 0 for any compact $\mathcal{K} \subset \boldsymbol{\Theta}$, and thus we have the required upper bound on $\ell_{\boldsymbol{\vartheta}}$. □

Note that the definition of $\boldsymbol{\vartheta}$-uniform ergodicity given above involves only bounded functions h, however the result above can be extended to a specific class of unbounded functions if one restricts their attention to diffusions on $[l, r]$ where $-\infty < l < r < \infty$ are both entrance boundaries.

Theorem 2.3. *Let Y be as in Theorem 2.2, and suppose that all the conditions stated there hold, but that both l and r are now entrance boundaries. Assume further that the function h is integrable with respect to the invariant density $f_{\boldsymbol{\vartheta}}^Y$ but possibly unbounded, that for any $l < a < b < r$, $\sup_{y \in [a,b]} h(y) < \infty$, and that for any $x < b$ the following*

hold

$$\sup_{\boldsymbol{\vartheta}\in\mathcal{K}}\int_x^b e^{-\int^\xi \frac{2\mu(\boldsymbol{\vartheta},y)}{\alpha^2(y)}dy}\int_l^\xi \frac{2h(\eta)}{\alpha^2(\eta)}e^{\int^\eta \frac{2\mu(\boldsymbol{\vartheta},y)}{\alpha^2(y)}dy}d\eta d\xi<\infty, \tag{2.12}$$

$$\sup_{\boldsymbol{\vartheta}\in\mathcal{K}}\int_x^b e^{-\int^\xi \frac{2\mu(\boldsymbol{\vartheta},y)}{\alpha^2(y)}dy}\int_l^\xi \frac{2h(\eta)}{\alpha^2(\eta)}e^{\int^\eta \frac{2\mu(\boldsymbol{\vartheta},y)}{\alpha^2(y)}dy} \times \mathbb{E}_\eta^{(\boldsymbol{\vartheta})}\left[\int_0^{T_b} h(Y_t)dt\right]d\eta d\xi<\infty, \tag{2.13}$$

$$\sup_{\boldsymbol{\vartheta}\in\mathcal{K}}\int_l^r \mathbb{E}_x^{(\boldsymbol{\vartheta})}\left[\int_0^{T_b} h(Y_t)dt\right]\nu(\boldsymbol{\vartheta},dx)<\infty, \tag{2.14}$$

for any compact set $\mathcal{K}\subset\boldsymbol{\Theta}$*, and* $T_b:=\inf\{t\geq 0: Y_t=b\}$*. Then* (2.3) *holds for the function* h*.*

Remark 2.4. Note that the above conditions imply that h is only unbounded at the end points (because the supremum between a and b of h is finite for any $l<a<b<r$), which in particular ensures that all integrals of the form $\int_0^T h(Y_t)dt$ above are well-defined as both boundary points are unattainable (in view of them being entrance).

Proof. Recall the notation introduced in Theorem 2.2, namely the regeneration times $\{S_n,R_n\}_{n\in\mathbb{N}}$ and the number of upcrossings up to time t, $\{N_t\}_{t\geq 0}$. We want to prove that

$$\lim_{T\to\infty}\sup_{\boldsymbol{\vartheta}\in\mathcal{K}}\mathbb{P}_\nu^{(\boldsymbol{\vartheta})}\left[\left|\frac{1}{T}\int_0^T h(Y_t)dt-\mathbb{E}^{(\boldsymbol{\vartheta})}\left[h(\xi)\right]\right|>\varepsilon\right]=0 \tag{2.15}$$

holds for any compact set $\mathcal{K}\subset\boldsymbol{\Theta}$, with h as defined in the statement of the theorem. The strategy here will be to decompose the sample path of the diffusion into i.i.d. blocks of excursions as done in Theorem 3.5 in [LLL11]. However, we will deal with the resulting expectations in a different way, namely by applying the ODE approach used in Theorem 2.2 to bound these quantities from above in $\boldsymbol{\vartheta}$ over a compact set $\mathcal{K}$. To this end, fix $\varepsilon\in(0,\mathbb{E}^{(\boldsymbol{\vartheta})}[h(\xi)])$ and choose $\delta\in(0,1)$ such that $\varepsilon=\delta\mathbb{E}^{(\boldsymbol{\vartheta})}[h(\xi)]$, and set $\Omega_T:=\{|N_TT^{-1}-\ell_{\boldsymbol{\vartheta}}|\leq\ell_{\boldsymbol{\vartheta}}\delta/4\}$ for $\ell_{\boldsymbol{\vartheta}}=\mathbb{E}_a^{(\boldsymbol{\vartheta})}[R_1]^{-1}$. Then as in the proof of Theorem 3.5 in [LLL11], we

get the following decomposition

$$\begin{aligned}
&\mathbb{P}_\nu^{(\boldsymbol{\vartheta})}\left[\left|\frac{1}{T}\int_0^T h(Y_t)dt - \mathbb{E}^{(\boldsymbol{\vartheta})}\left[h(\xi)\right]\right| > \varepsilon\right] \\
&\quad\leq \mathbb{P}_\nu^{(\boldsymbol{\vartheta})}\left[\left|\int_0^{R_1} h(Y_t)dt\right| > \frac{T\varepsilon}{4}\right] \\
&\qquad + \mathbb{P}_\nu^{(\boldsymbol{\vartheta})}\left[\left|\int_{R_1}^{R_{N_T}+1} h(Y_t)dt - N_T\mathbb{E}^{(\boldsymbol{\vartheta})}\left[h(\xi)\right]\mathbb{E}_a^{(\boldsymbol{\vartheta})}\left[R_1\right]\right| > \frac{T\varepsilon}{4};\Omega_T\right] \\
&\qquad + \mathbb{P}_\nu^{(\boldsymbol{\vartheta})}\left[\left|N_T\mathbb{E}^{(\boldsymbol{\vartheta})}\left[h(\xi)\right]\mathbb{E}_a^{(\boldsymbol{\vartheta})}\left[R_1\right] - T\mathbb{E}^{(\boldsymbol{\vartheta})}\left[h(\xi)\right]\right| > \frac{T\varepsilon}{4};\Omega_T\right] \\
&\qquad + \mathbb{P}_\nu^{(\boldsymbol{\vartheta})}\left[\left|\int_T^{R_{N_T}+1} h(Y_t)dt\right| > \frac{T\varepsilon}{4};\Omega_T\right] + \mathbb{P}_\nu^{(\boldsymbol{\vartheta})}\left[\Omega_T^c\right] \\
&\qquad =: A + B + E + C + D
\end{aligned}$$

Dealing with E and D can be achieved as in equations (3.10) and (3.14) in [LLL11], to deduce that $E = 0$ and

$$D \leq \frac{1}{T\varepsilon^2}\mathbb{E}^{(\boldsymbol{\vartheta})}\left[h(\xi)\right]^2\left(2\mathbb{E}_\nu^{(\boldsymbol{\vartheta})}\left[\left|R_1 - \ell_{\boldsymbol{\vartheta}}^{-1}\right|\right] + 2^3C_1^2\mathbb{E}_\nu^{(\boldsymbol{\vartheta})}\left[\left|\bar{\eta}_1\right|^2\right]\ell_{\boldsymbol{\vartheta}}\right),$$

for C_1 the constant from the Burkholder-Davis-Gundy inequality. All the above expressions are either constant or have been shown to be bounded in $\boldsymbol{\vartheta}$ over compact sets in the parameter space in Theorem 2.2, so it remains to deal with terms A, B and C above.

Applying Markov's inequality to A gives

$$A \leq \frac{4}{T\varepsilon}\mathbb{E}_\nu^{(\boldsymbol{\vartheta})}\left[\int_0^{R_1} h(Y_t)dt\right]$$

and we can decompose the above integral

$$\begin{aligned}
\mathbb{E}_\nu^{(\boldsymbol{\vartheta})}\left[\int_0^{R_1} h(Y_t)dt\right] &= \mathbb{E}_\nu^{(\boldsymbol{\vartheta})}\left[\int_0^{S_1} h(Y_t)dt\right] + \mathbb{E}_\nu^{(\boldsymbol{\vartheta})}\left[\int_{S_1}^{R_1} h(Y_t)dt\right] \\
&\leq \mathbb{E}_\nu^{(\boldsymbol{\vartheta})}\left[\int_0^{T_b} h(Y_t)dt\right] + \sup_{y\in[a,b]} h(y)\mathbb{E}_\nu^{(\boldsymbol{\vartheta})}\left[R_1\right]. \quad (2.16)
\end{aligned}$$

So it remains to prove that the first term on the RHS can be bounded

from above in $\boldsymbol{\vartheta}$. It turns out that B and C can be bounded by similar quantities, so we do this first and subsequently show that the resulting quantities can be bounded in $\boldsymbol{\vartheta}$ too.

Indeed, set $\xi_k := \int_{R_k}^{R_{k+1}} h(Y_t)dt$, $M_0 = 0$, and

$$M_n := \sum_{k=1}^{n} \left(\xi_k - \mathbb{E}_\nu^{(\boldsymbol{\vartheta})}\left[\xi_k\right]\right).$$

Then

$$\begin{aligned} B = \mathbb{P}_\nu^{(\boldsymbol{\vartheta})}\left[|M_{N_T}| > \frac{T\varepsilon}{4}; \Omega_T\right] &\leq \mathbb{P}_\nu^{(\boldsymbol{\vartheta})}\left[\sup_{n \leq \lfloor T\ell_{\boldsymbol{\vartheta}}(1+\delta/4)\rfloor} |M_n| > \frac{T\varepsilon}{4}\right] \\ &\leq \left(\frac{4}{T\varepsilon}\right)^2 \mathbb{V}_\nu^{(\boldsymbol{\vartheta})}\left[M_{\lfloor T\ell_{\boldsymbol{\vartheta}}(1+\delta/4)\rfloor}\right] \end{aligned}$$

by the Kolmogorov inequality where $\mathbb{V}_\nu^{(\boldsymbol{\vartheta})}$ denotes the variance with respect to the measure $\mathbb{P}_\nu^{(\boldsymbol{\vartheta})}$. Now observe that

$$\begin{aligned} \mathbb{V}_\nu^{(\boldsymbol{\vartheta})}\left[M_{\lfloor T\ell_{\boldsymbol{\vartheta}}(1+\delta/4)\rfloor}\right] &= \sum_{k=1}^{\lfloor T\ell_{\boldsymbol{\vartheta}}(1+\delta/4)\rfloor} \mathbb{V}_\nu^{\boldsymbol{\vartheta}}\left[\left(\xi_k - \mathbb{E}_\nu^{(\boldsymbol{\vartheta})}\left[\xi_k\right]\right)^2\right] \\ &= \lfloor T\ell_{\boldsymbol{\vartheta}}(1+\delta/4)\rfloor \mathbb{E}_\nu^{(\boldsymbol{\vartheta})}\left[\left(\xi_1 - \mathbb{E}_\nu^{(\boldsymbol{\vartheta})}\left[\xi_1\right]\right)^2\right] \\ &\leq \lfloor T\ell_{\boldsymbol{\vartheta}}(1+\delta/4)\rfloor 2\left(\mathbb{E}_a^{(\boldsymbol{\vartheta})}\left[\xi_0^2\right] + \mathbb{E}_a^{(\boldsymbol{\vartheta})}\left[\xi_0\right]^2\right). \end{aligned}$$

because the $\{\xi_k\}_{k=1}^\infty$ are i.i.d., and moreover we have that under $\mathbb{P}_\nu^{(\boldsymbol{\vartheta})}$ they are equal in distribution to ξ_0 under $\mathbb{P}_a^{(\boldsymbol{\vartheta})}$. So

$$B \leq \frac{4^2\lfloor \ell_{\boldsymbol{\vartheta}}(1+\delta/4)\rfloor}{T\varepsilon^2} 2\left(\mathbb{E}_a^{(\boldsymbol{\vartheta})}\left[\xi_0^2\right] + \mathbb{E}_a^{(\boldsymbol{\vartheta})}\left[\xi_0\right]^2\right). \tag{2.17}$$

The second term of (2.17) can be bounded in the same way as in (2.16),

whilst for the first term we can use a similar decomposition to get

$$\mathbb{E}_a^{(\boldsymbol{\vartheta})}\left[\xi_0^2\right] \leq 2\Bigg(\mathbb{E}_a^{(\boldsymbol{\vartheta})}\left[\left(\int_0^{T_b} h(Y_t)dt\right)^2\right] + \sup_{y\in[a,b]} h(y)^2\mathbb{E}_a^{(\boldsymbol{\vartheta})}\left[R_1^2\right]\Bigg). \tag{2.18}$$

Finally, for C we use the same arguments as in [LLL11] (just before equation (3.13)) to get that

$$\begin{aligned} C &\leq \sum_{k=1}^{\lfloor T\ell_{\boldsymbol{\vartheta}}(1+\delta/4)\rfloor} \mathbb{P}_\nu^{(\boldsymbol{\vartheta})}\left[\int_{R_k}^{R_k+1} h(Y_t)dt > \frac{T\varepsilon}{4}\right] \\ &\leq \frac{4^2\lfloor T\ell_{\boldsymbol{\vartheta}}(1+\delta/4)\rfloor}{T^2\varepsilon^2}\mathbb{E}_\nu^{(\boldsymbol{\vartheta})}\left[\left(\int_{R_1}^{R_2} h(Y_t)dt\right)^2\right] \\ &\leq \frac{4^2\ell_{\boldsymbol{\vartheta}}(1+\delta/4)}{T\varepsilon^2}\mathbb{E}_a^{(\boldsymbol{\vartheta})}\left[\left(\int_0^{R_1} h(Y_t)dt\right)^2\right], \end{aligned}$$

and we can apply the same reasoning as in (2.18). It remains to show that the terms

$$\mathbb{E}_a^{(\boldsymbol{\vartheta})}\left[\int_0^{T_b} h(Y_t)dt\right],\quad \mathbb{E}_\nu^{(\boldsymbol{\vartheta})}\left[\int_0^{T_b} h(Y_t)dt\right],\quad \mathbb{E}_a^{(\boldsymbol{\vartheta})}\left[\left(\int_0^{T_b} h(Y_t)dt\right)^2\right]$$

can be bounded from above in $\boldsymbol{\vartheta}$. The same arguments used to derive the ODEs in Theorem 2.2 can be used here to derive an ODE for $U_n(x) := \mathbb{E}_x^{(\boldsymbol{\vartheta})}[(\int_0^{T_b} h(Y_t)dt)^n]$ for the cases when $x < b$ and $x > b$ with the same boundary conditions as in Theorem 2.2. Thus for $n \in \mathbb{N}\setminus\{0\}$, the following recursion holds for $U_n(x)$ when $x < b$

$$U_n(x) = n\int_x^b e^{-\int^\xi \frac{2\mu(\boldsymbol{\vartheta},y)}{\alpha^2(y)}dy}\int_l^\xi \frac{2h(\eta)}{\alpha^2(\eta)}e^{\int^\eta \frac{2\mu(\boldsymbol{\vartheta},y)}{\alpha^2(y)}dy}U_{n-1}(\eta)d\eta d\xi, \tag{2.19}$$

and for $x > b$ we have

$$U_n(x) = n\int_b^x e^{-\int^\xi \frac{2\mu(\boldsymbol{\vartheta},y)}{\alpha^2(y)}dy}\int_\xi^r \frac{2h(\eta)}{\alpha^2(\eta)}e^{\int^\eta \frac{2\mu(\boldsymbol{\vartheta},y)}{\alpha^2(y)}dy}U_{n-1}(\eta)d\eta d\xi. \tag{2.20}$$

Now for $n = 1$, we get that for $x < b$,

$$\mathbb{E}_x^{(\boldsymbol{\vartheta})}\left[\int_0^{T_b} h(Y_t)dt\right] = \int_x^b e^{-\int^\xi \frac{2\mu(\boldsymbol{\vartheta},y)}{\alpha^2(y)}dy} \int_l^\xi \frac{2h(\eta)}{\alpha^2(\eta)} e^{\int^\eta \frac{2\mu(\boldsymbol{\vartheta},y)}{\alpha^2(y)}dy} d\eta d\xi$$

which is bounded over any compact set $\mathcal{K} \subset \boldsymbol{\Theta}$ by (2.12). In view of condition (2.14), we get that $\mathbb{E}_\nu^{(\boldsymbol{\vartheta})}[\int_0^{T_b} h(Y_t)dt]$ is also bounded from above in $\boldsymbol{\vartheta}$ over compact sets $\mathcal{K} \subset \boldsymbol{\Theta}$, and finally, using the recursions in (2.19), we get that for $x < b$,

$$\begin{aligned}
&\mathbb{E}_x^{(\boldsymbol{\vartheta})}\left[\left(\int_0^{T_b} h(Y_t)dt\right)^2\right] \\
&\quad = 2\int_x^b e^{-\int^\xi \frac{2\mu(\boldsymbol{\vartheta},y)}{\alpha^2(y)}dy} \int_l^\xi \frac{2h(\eta)}{\alpha^2(\eta)} e^{\int^\eta \frac{2\mu(\boldsymbol{\vartheta},y)}{\alpha^2(y)}dy} \mathbb{E}_\eta^{(\boldsymbol{\vartheta})}\left[\int_0^{T_b} h(Y_t)dt\right] d\eta d\xi
\end{aligned}$$

which is bounded from above in $\boldsymbol{\vartheta}$ over a given compact set $\mathcal{K} \subset \boldsymbol{\Theta}$ by (2.13), giving the required bounds for the quantities A, B, and C. Combining these with the bounds for D and E we conclude that (2.15) holds. □

2.2 The scalar Wright–Fisher diffusion

We now give a brief overview of the Wright–Fisher diffusion before showing that the diffusion is ergodic uniformly in the selection and mutation parameters, and subsequently use this to prove the uniform local asymptotic normality (LAN) of the family of measures associated to the solution of the SDE.

Consider an infinite haploid population undergoing selection and mutation, where we are interested in two alleles A_1 and A_2. Suppose that $\boldsymbol{\vartheta} = (\sigma, \theta_1, \theta_2) \in \boldsymbol{\Theta} = \mathbb{R} \times (0, \infty)^2$ are the selection and mutation parameters respectively, where σ describes the extent to which allele A_2 is favoured over A_1, alleles of type A_1 mutate to A_2 at rate $\theta_1/2$, and those of type A_2 mutate to A_1 at rate $\theta_2/2$. Let X_t denote the frequency of A_2 in the population at time t. Then the dynamics of X_t can be described by a diffusion process on $[0, 1]$, which, after

expressing the parameters on an appropriate timescale, satisfies the SDE

$$\begin{aligned} dX_t &= \mu_{\mathrm{WF}}(\boldsymbol{\vartheta}, X_t)dt + \alpha_{\mathrm{WF}}(X_t)dW_t \\ &:= \frac{1}{2}\left(\sigma X_t(1-X_t) - \theta_2 X_t + \theta_1(1-X_t)\right)dt \\ &\quad + \sqrt{X_t(1-X_t)}dW_t, \end{aligned} \tag{2.21}$$

with $X_0 \sim \nu$ for some initial distribution ν, and $[0, T]$ the observation interval. We point out that (2.21) with $\sigma = 0$ is commonly referred to as the *neutral* Wright–Fisher diffusion, whilst $\sigma \neq 0$ is known as the *non-neutral* case. A strong solution to (2.21) exists by the Yamada–Watanabe condition (see Theorem 3.2, Chapter IV in [IW89]), but weak uniqueness suffices for the results in Chapter 3. In abuse of notation, we redefine $\mathbb{P}_\nu^{(\boldsymbol{\vartheta})}$ to be the law induced on $C_T([0,1])$ by the solution to (2.21) when the true diffusion parameters are $\boldsymbol{\vartheta} = (\sigma, \theta_1, \theta_2)$, and $X_0 \sim \nu$, and similarly for the expectation with respect to $\mathbb{P}_\nu^{(\boldsymbol{\vartheta})}$, $\mathbb{E}_\nu^{(\boldsymbol{\vartheta})}$, and with respect to the stationary distribution, $\mathbb{E}^{(\boldsymbol{\vartheta})}$ (the existence of which we discuss below).

We assume that $\theta_1, \theta_2 > 0$, for if at least one is 0 then the diffusion is absorbed in finite time and we are back in the regime studied by Watterson [Wat79]. The boundary behaviour depends on whether the mutation parameters are either less than, or greater or equal to 1, but in either case the diffusion is ergodic as long as $\theta_1, \theta_2 > 0$ (see Lemma 2.1, Chapter 10 in [EK86]).

Substituting μ_{WF} and α_{WF} into (2.2) and simplifying terms leads to the following density for the stationary distribution of the Wright–Fisher diffusion (2.21)

$$f_{\boldsymbol{\vartheta}}(x) = \frac{1}{G_{\boldsymbol{\vartheta}}} e^{\sigma x} x^{\theta_1 - 1}(1-x)^{\theta_2 - 1}, \qquad x \in (0,1), \tag{2.22}$$

where $G_{\boldsymbol{\vartheta}}$ is the normalising constant

$$G_{\boldsymbol{\vartheta}} = \int_0^1 e^{\sigma x} x^{\theta_1 - 1}(1-x)^{\theta_2 - 1} dx \leq \max\{e^{\sigma}, 1\} B(\theta_1, \theta_2) < \infty, \quad (2.23)$$

with

$$B(\theta_1, \theta_2) := \int_0^1 x^{\theta_1 - 1}(1-x)^{\theta_2 - 1} dx \quad (2.24)$$

the beta function. As above we will always assume that $\xi \sim f_{\boldsymbol{\vartheta}}$.

2.2.1 Uniform in the selection and mutation parameters ergodicity

To the best of our knowledge, it has not been shown that the Wright–Fisher diffusion is ergodic *uniformly in its parameters*, which motivates the following corollary to Theorem 2.2.

Corollary 2.5. *The Wright–Fisher diffusion with mutation and selection is uniformly in the selection and mutation parameters* $\boldsymbol{\vartheta} = (\sigma, \theta_1, \theta_2)$ *ergodic for any initial distribution* ν.

Proof. We show that the conditions of Theorem 2.2 hold for the Wright–Fisher diffusion. Positive recurrence follows immediately from (2.23), whilst the existence of a unique strong solution is guaranteed by the Yamada–Watanabe condition. That the boundary points 0 and 1 are either entrance or regular is a consequence of the fact that the mutation parameters are assumed to be strictly positive (see (6.18) and (6.19) in [KT81, Chapter 15, Section 6]). It remains to show that both expressions in (2.4) are bounded away from 0 for any $\mathcal{K} \subset \mathbb{R} \times (0, \infty)^2$

compact. To this end let $\bar{\theta}_1 := \sup_{\boldsymbol{\vartheta}\in\mathcal{K}} \theta_1$, $\bar{\theta}_2 := \sup_{\boldsymbol{\vartheta}\in\mathcal{K}} \theta_2$. Then

$$\begin{aligned}
&\int_a^b e^{-\int_\xi \frac{2\mu_{\mathrm{WF}}(\boldsymbol{\vartheta},y)}{\alpha^2_{\mathrm{WF}}(y)}dy} \int_0^\xi \frac{1}{\sigma^2_{\mathrm{WF}}(\eta)} e^{\int_\eta \frac{2\mu_{\mathrm{WF}}(\boldsymbol{\vartheta},y)}{\alpha^2_{\mathrm{WF}}(y)}dy} d\eta d\xi \\
&\quad = \int_a^b 2e^{-\sigma\xi}\xi^{-\theta_1}(1-\xi)^{-\theta_2} \int_0^\xi e^{\sigma\eta}\eta^{\theta_1-1}(1-\eta)^{\theta_2-1} d\eta d\xi \\
&\quad \geq 2\min\{e^{-\sigma},1\} \int_a^b \xi^{-\theta_1}(1-\xi)^{-\theta_2} d\xi \int_0^a \eta^{\theta_1-1}(1-\eta)^{\theta_2-1} d\eta \\
&\quad \geq 2\min\{e^{-\sigma},1\}(b-a)\frac{a^{\theta_1}}{\theta_1}(1-a)^{\bar{\theta}_2-1}, \qquad (2.25) \\
&\int_a^b e^{-\int_\xi \frac{2\mu_{\mathrm{WF}}(\boldsymbol{\vartheta},y)}{\alpha^2_{\mathrm{WF}}(y)}dy} \int_\xi^1 \frac{1}{\sigma^2_{\mathrm{WF}}(\eta)} e^{\int_\eta \frac{2\mu_{\mathrm{WF}}(\boldsymbol{\vartheta},y)}{\alpha^2_{\mathrm{WF}}(y)}dy} d\eta d\xi \\
&\quad = \int_a^b 2e^{-\sigma\xi}\xi^{-\theta_1}(1-\xi)^{-\theta_2} \int_\xi^1 e^{\sigma\eta}\eta^{\theta_1-1}(1-\eta)^{\theta_2-1} d\eta d\xi \\
&\quad \geq 2\min\{e^{\sigma},1\} \int_a^b \xi^{-\theta_1}(1-\xi)^{-\theta_2} d\xi \int_b^1 \eta^{\theta_1-1}(1-\eta)^{\theta_2-1} d\eta \\
&\quad \geq 2\min\{e^{\sigma},1\}(b-a)\frac{(1-b)^{\theta_2}}{\theta_2} b^{\bar{\theta}_1-1}, \qquad (2.26)
\end{aligned}$$

which follows by observing that

$$\begin{aligned}
\xi^{-\theta_1}(1-\xi)^{-\theta_2} &> 1 && \forall \xi \in (a,b), \forall \theta_1, \theta_2 > 0, \\
(1-\eta)^{\theta_2-1} &\geq (1-a)^{\bar{\theta}_2-1} && \forall \eta \in (0,a), \\
\eta^{\theta_1-1} &\geq b^{\bar{\theta}_1-1} && \forall \eta \in (b,1).
\end{aligned}$$

As the RHS of both (2.25) and (2.26) are bounded away from 0 on $\mathcal{K}$, the result follows by applying Theorem 2.2. □

For the remainder of this chapter we restrict our attention to the parameter space $\boldsymbol{\Theta} \subset \mathbb{R} \times [1,\infty)^2$, where $\boldsymbol{\Theta}$ is open and bounded, for if either of the mutation parameters were less than 1 then the measures $\mathbb{P}_\nu^{(\boldsymbol{\vartheta})}$ within this region would be mutually singular with respect to one another and thus their Radon–Nikodym derivative undefined. Restricting our attention to mutation parameters within the range $[1,\infty)^2$ thus ensures that the family of measures $\{\mathbb{P}_\nu^{(\boldsymbol{\vartheta})}, \boldsymbol{\vartheta} \in \boldsymbol{\Theta}\}$ are equivalent, and

we have that

$$\frac{d\mathbb{P}_\nu^{(\boldsymbol{\vartheta}')}}{d\mathbb{P}_\nu^{(\boldsymbol{\vartheta})}}(X^T) = \frac{\nu(\boldsymbol{\vartheta}', X_0)}{\nu(\boldsymbol{\vartheta}, X_0)} \times \exp\left\{ \int_0^T \left(\frac{\mu_{\mathrm{WF}}(\boldsymbol{\vartheta}', X_t) - \mu_{\mathrm{WF}}(\boldsymbol{\vartheta}, X_t)}{\alpha_{\mathrm{WF}}(X_t)} \right) dW_t - \frac{1}{2} \int_0^T \left(\frac{\mu_{\mathrm{WF}}(\boldsymbol{\vartheta}', X_t) - \mu_{\mathrm{WF}}(\boldsymbol{\vartheta}, X_t)}{\alpha_{\mathrm{WF}}(X_t)} \right)^2 dt \right\} \quad (2.27)$$

with $\mathbb{P}_\nu^{(\boldsymbol{\vartheta})}$-probability 1 when the true value is $\boldsymbol{\vartheta}$, where we assume that the initial distributions $\{\nu(\boldsymbol{\vartheta}, \cdot)\}_{\boldsymbol{\vartheta} \in \boldsymbol{\Theta}}$ are mutually equivalent and admit a density with respect to some common dominating measure $\lambda(\cdot)$, which (in abuse of notation) we denote by $\nu(\boldsymbol{\vartheta}, \cdot)$. Proofs of the above claims regarding the equivalence of the Wright–Fisher diffusion and the form of the Radon–Nikodym derivative can be found in [DMS93], Lemma 7.2.2 and Section 10.1.1. We emphasise here that we have allowed the starting distribution ν to depend on the parameters, as is evident from the first ratio in (2.27). However if there is no such dependence then this ratio is equal to 1 and our results still apply.

Furthermore, restricting to mutation parameters greater than or equal to 1 ensures that the diffusion boundaries now become entrance (see equations (6.18) and (6.19) in [KT81, Chapter 15, Section 6]), and as done in Theorem 2.3, (2.3) can be extended for a particular class of unbounded functions. We focus on two such functions for this class of diffusions, as they turn out to be an essential ingredient necessary to prove the LAN property.

Corollary 2.6. *For the Wright–Fisher diffusion with mutation and selection parameters $\boldsymbol{\vartheta} \in \boldsymbol{\Theta} \subset \mathbb{R} \times [1, \infty)^2$ (for $\boldsymbol{\Theta}$ an open bounded set) with initial distribution ν satisfying* (2.31) *(which is defined in the statement of Theorem 2.8), $\boldsymbol{\vartheta}$-uniform ergodicity* (2.3) *holds also for the functions $h(x) = (1-x)x^{-1}$ and $h(x) = (1-x)^{-1}x$. The result holds in particular for the case $\nu = f_{\boldsymbol{\vartheta}}$.*

Proof. The result follows immediately if we show that all the conditions

of Theorem 2.3 are satisfied for the above functions. In particular, we show that they hold for $h(x) = (1-x)x^{-1}$, as similar arguments apply for the case $h(x) = (1-x)^{-1}x$. The conditions of Theorem 2.2 have already been shown to hold in Corollary 2.5, whilst integrability with respect to the invariant density is guaranteed as we are considering mutation rates $(\theta_1, \theta_2) \in (1, \infty)^2$. That $\sup_{y\in[a,b]} h(y) < \infty$ for any pair $0 < a < b < 1$ is immediate, so it remains to show (2.12), (2.13), and (2.14). Observe that

$$\begin{aligned}
\mathbb{E}_x^{(\boldsymbol{\vartheta})}\left[\int_0^{T_b} \frac{1-X_t}{X_t}dt\right] &= 2\int_x^b e^{-\sigma\xi}\xi^{-\theta_1}(1-\xi)^{-\theta_2} \\
&\qquad \times \int_0^\xi e^{\sigma\eta}\eta^{\theta_1-2}(1-\eta)^{\theta_2}d\eta d\xi \\
&\leq 2\max\{e^{-\sigma},1\}\int_x^b \xi^{-\theta_1}(1-\xi)^{-\theta_2}\int_0^\xi \eta^{\theta_1-2}d\eta d\xi \\
&= 2\max\{e^{-\sigma},1\}\frac{1}{\theta_1-1}\int_x^b \xi^{-1}(1-\xi)^{-\theta_2}d\xi,
\end{aligned}$$

so (2.12) holds as the RHS is continuous in $\boldsymbol{\vartheta}$ and thus can be bounded from above in $\boldsymbol{\vartheta}$ over any compact set $\mathcal{K} \subset \boldsymbol{\Theta}$. For $x > b$

$$\begin{aligned}
\mathbb{E}_x^{(\boldsymbol{\vartheta})}\left[\int_0^{T_b} \frac{1-X_t}{X_t}dt\right] &= 2\int_b^x e^{-\sigma\xi}\xi^{-\theta_1}(1-\xi)^{-\theta_2} \\
&\qquad \times \int_\xi^1 e^{\sigma\eta}\eta^{\theta_1-2}(1-\eta)^{\theta_2}d\eta d\xi \\
&\leq 2\max\{e^{\sigma},1\}\int_b^x \xi^{-\max\{\theta_1,2\}}(1-\xi)^{-\theta_2} \\
&\qquad \times \int_\xi^1 (1-\eta)^{\theta_2}d\eta d\xi \\
&= 2\max\{e^{\sigma},1\}\frac{1}{\theta_2+1}\int_b^x \xi^{-\max\{\theta_1,2\}}(1-\xi)d\xi \\
&\leq 2\max\{e^{\sigma},1\}\frac{1}{\theta_2+1}\int_b^x \xi^{-\max\{\theta_1,2\}}d\xi,
\end{aligned}$$

and thus (2.14) holds in view of condition (2.31). In the case when

$\nu = f_{\boldsymbol{\vartheta}}$, we get that

$$\begin{aligned}
\mathbb{E}_{\nu}^{(\boldsymbol{\vartheta})}\left[\int_0^{T_b}\frac{1-X_t}{X_t}dt\right] &\leq 2\max\{e^{-\sigma},1\}\frac{1}{\theta_1-1}\\
&\qquad\times\int_0^b\int_x^b \xi^{-1}(1-\xi)^{-\theta_2}d\xi f_{\boldsymbol{\vartheta}}(x)dx\\
&\quad+2\max\{e^{\sigma},1\}\frac{1}{\theta_2+1}\\
&\qquad\times\int_b^1\int_b^x \xi^{-\max\{\theta_1,2\}}d\xi f_{\boldsymbol{\vartheta}}(x)dx\\
&\leq 2\max\{e^{\sigma},1\}\frac{1}{\theta_1(\theta_1-1)}\frac{1}{G_{\boldsymbol{\vartheta}}}\int_0^b(1-\xi)^{-\theta_2}d\xi\\
&\quad+2\max\{e^{\sigma},1\}\frac{1}{(\theta_2+1)}\int_b^1\xi^{-\max\{\theta_1,2\}}d\xi,
\end{aligned}$$

which follows from

$$\begin{aligned}
&\int_0^b\int_x^b \xi^{-1}(1-\xi)^{-\theta_2}x^{\theta_1-1}(1-x)^{\theta_2-1}d\xi dx\\
&=\int_0^b\int_0^\xi \xi^{-1}(1-\xi)^{-\theta_2}x^{\theta_1-1}(1-x)^{\theta_2-1}dxd\xi\\
&\leq\frac{1}{\theta_1}\int_0^b\xi^{\theta_1-1}(1-\xi)^{-\theta_2}d\xi\\
&\leq\frac{1}{\theta_1}\int_0^b(1-\xi)^{-\theta_2}d\xi
\end{aligned}$$

because $\theta_1,\theta_2>1$, and

$$\begin{aligned}
\int_b^1\int_b^x \xi^{-\max\{\theta_1,2\}}f_{\boldsymbol{\vartheta}}(x)d\xi dx &= \int_b^1\int_\xi^1 \xi^{-\max\{\theta_1,2\}}f_{\boldsymbol{\vartheta}}(x)dxd\xi\\
&\leq\int_b^1\xi^{-\max\{\theta_1,2\}}d\xi.
\end{aligned}$$

Finally, using the recursions in (2.19) and (2.20), we get that for $x < b$,

$$\begin{aligned}\mathbb{E}_x^{(\boldsymbol{\vartheta})}\left[\left(\int_0^{T_b}\frac{1-X_t}{X_t}dt\right)^2\right] &\leq \frac{2(2\max\{e^{-\sigma},1\})^2}{(\theta_1-1)^2}\int_0^b \gamma^{\theta_1-2}(1-\gamma)^{-\theta_2}d\gamma \\ &\quad\times\int_x^b \xi^{-\theta_1}(1-\xi)^{-\theta_2}d\xi \\ &\leq \frac{2(2\max\{e^{-\sigma},1\})^2}{(\theta_1-1)^2}(1-b)^{-\theta_2}\int_0^b \gamma^{\theta_1-2}d\gamma \\ &\quad\times\int_x^b \xi^{-\theta_1}(1-\xi)^{-\theta_2}d\xi\end{aligned}$$

which follows from

$$\begin{aligned}&\int_0^\xi \eta^{\theta_1-2}(1-\eta)^{\theta_2}\int_\eta^b \gamma^{-1}(1-\gamma)^{-\theta_2}d\gamma d\eta \\ &\leq \int_0^b \eta^{\theta_1-2}(1-\eta)^{\theta_2}\int_\eta^b \gamma^{-1}(1-\gamma)^{-\theta_2}d\gamma d\eta \\ &\leq \int_0^b \gamma^{\theta_1-2}(1-\gamma)^{-\theta_2}d\gamma,\end{aligned}$$

and again the corresponding RHS can be bounded from above over any compact set $\mathcal{K}\subset\boldsymbol{\Theta}$ using continuity in $\boldsymbol{\vartheta}$, such that (2.13) holds and so the result follows by Theorem 2.3. □

2.2.2 Local asymptotic normality

We end this chapter by introducing the concept of *local asymptotic normality* and show that the corresponding family of measures associated to the Wright–Fisher diffusion is uniformly locally asymptotically normal, which will be essential in the next section.

Definition 2.7 (Special case of Definition 2.1 in [Kut04])**.** The family of measures $\{\mathbb{P}_\nu^{(\boldsymbol{\vartheta})},\boldsymbol{\vartheta}\in\boldsymbol{\Theta}\}$ induced by the solution X^T to the SDE (2.1) is said to be *locally asymptotically normal (LAN) at a point* $\boldsymbol{\vartheta}_0\in\boldsymbol{\Theta}$ *at rate* $T^{-1/2}$ if for any $\mathbf{u}\in\mathbb{R}^d$, the likelihood ratio function admits the

representation

$$Z_{T,\boldsymbol{\vartheta}_0}(\mathbf{u}) := \frac{d\mathbb{P}_\nu^{(\boldsymbol{\vartheta}_0+\frac{\mathbf{u}}{\sqrt{T}})}}{d\mathbb{P}_\nu^{(\boldsymbol{\vartheta}_0)}}(X^T)$$
$$= \exp\left\{\left\langle \mathbf{u}, \Delta_T(\boldsymbol{\vartheta}_0, X^T)\right\rangle - \frac{1}{2}\left\langle \mathbf{I}(\boldsymbol{\vartheta}_0)\mathbf{u}, \mathbf{u}\right\rangle + r_T(\boldsymbol{\vartheta}_0, \mathbf{u}, X^T)\right\},$$

where $\langle\cdot,\cdot\rangle$ denotes the Euclidean inner product on $\mathbb{R}^d$, and $\Delta_T(\boldsymbol{\vartheta}_0, X^T)$ is a random variable such that

$$\Delta_T(\boldsymbol{\vartheta}_0, X^T) \xrightarrow{\mathrm{d}} N(\mathbf{0}, \mathbf{I}(\boldsymbol{\vartheta}_0)),$$

with $\mathbf{I}(\boldsymbol{\vartheta}_0)$ the Fisher information matrix evaluated at $\boldsymbol{\vartheta}_0$, i.e.

$$\mathbf{I}(\boldsymbol{\vartheta}_0) := \mathbb{E}^{(\boldsymbol{\vartheta}_0)}\left[\frac{\dot{\boldsymbol{\mu}}(\boldsymbol{\vartheta}_0, \xi)\dot{\boldsymbol{\mu}}(\boldsymbol{\vartheta}_0, \xi)^T}{\alpha^2(\xi)}\right],$$

where $\dot{\boldsymbol{\mu}}(\boldsymbol{\vartheta}, \xi)^T$ is the transpose of the vector of derivatives of $\mu(\boldsymbol{\vartheta}, x)$ with respect to $\boldsymbol{\vartheta}$. Moreover, the function $r_T(\boldsymbol{\vartheta}_0, \mathbf{u}, X^T)$ satisfies

$$\lim_{T\to\infty} r_T(\boldsymbol{\vartheta}_0, \mathbf{u}, X^T) = 0 \quad \text{in } \mathbb{P}_\nu^{(\boldsymbol{\vartheta}_0)}\text{-probability}$$

The family of measures is said to be *LAN on* $\boldsymbol{\Theta}$ if it is LAN at every point $\boldsymbol{\vartheta}_0 \in \boldsymbol{\Theta}$, and further it is said to be *uniformly LAN* on $\boldsymbol{\Theta}$ if either convergence above holds uniformly in $\boldsymbol{\vartheta} \in \mathcal{K}$ for $\mathcal{K} \subset \Theta$ compact, by which we mean that for any compact $\mathcal{K} \subset \boldsymbol{\Theta}$, and any measurable, continuous and bounded function g,

$$\lim_{T\to\infty} \sup_{\boldsymbol{\vartheta}\in\mathcal{K}} \left|\mathbb{E}_\nu^{(\boldsymbol{\vartheta})}\left[g\left(\Delta_T(\boldsymbol{\vartheta}, X^T)\right)\right] - \mathbb{E}\left[g(\zeta)\right]\right| = 0 \tag{2.28}$$

for $\zeta \sim N(\mathbf{0}, \mathbf{I}(\boldsymbol{\vartheta}))$, and $\forall \varepsilon > 0$

$$\lim_{T\to\infty} \sup_{\boldsymbol{\vartheta}\in\mathcal{K}} \mathbb{P}_\nu^{(\boldsymbol{\vartheta})}\left[\left|r_T(\boldsymbol{\vartheta}, \mathbf{u}, X^T)\right| > \varepsilon\right] = 0. \tag{2.29}$$

Theorem 2.8. *The family of measures* $\{\mathbb{P}_\nu^{(\boldsymbol{\vartheta})}, \boldsymbol{\vartheta} \in \boldsymbol{\Theta}\}$ *induced by the*

weak solution to (2.21) with initial distribution satisfying

$$\lim_{|\boldsymbol{\varepsilon}|\to 0} \frac{\nu(\boldsymbol{\vartheta}+\boldsymbol{\varepsilon}, x)}{\nu(\boldsymbol{\vartheta}, x)} = 1, \quad \forall x \in [0,1], \tag{2.30}$$

$$\sup_{\boldsymbol{\vartheta}\in\mathcal{K}} \left\{ \int_0^b \frac{\max\{e^{-\sigma}, 1\}}{\theta_1 - 1} \int_x^b \xi^{-1}(1-\xi)^{-\theta_2} d\xi \nu(\boldsymbol{\vartheta}, dx) \right.$$
$$\left. + \int_b^1 \frac{\max\{e^{\sigma}, 1\}}{\theta_2 + 1} \int_b^x \xi^{-\max\{\theta_1, 2\}} d\xi \nu(\boldsymbol{\vartheta}, dx) \right\} \leq C_{\mathcal{K}} \tag{2.31}$$

on any compact set $\mathcal{K} \subset \Theta$ with $C_{\mathcal{K}} > 0$ constant, is uniformly LAN on $\boldsymbol{\Theta}$, with the likelihood ratio function $Z_{T,\boldsymbol{\vartheta}}(\mathbf{u})$ admitting the representation

$$Z_{T,\boldsymbol{\vartheta}}(\mathbf{u}) = \exp\left\{ \left\langle \mathbf{u}, \boldsymbol{\Delta}_T(\boldsymbol{\vartheta}, X^T) \right\rangle - \frac{1}{2} \left\langle \mathbf{I}(\boldsymbol{\vartheta})\mathbf{u}, \mathbf{u} \right\rangle + r_T(\boldsymbol{\vartheta}, \mathbf{u}, X^T) \right\}$$

for $\mathbf{u} \in \mathcal{U}_{T,\boldsymbol{\vartheta}} = \{\mathbf{u} : \boldsymbol{\vartheta} + \frac{\mathbf{u}}{\sqrt{T}} \in \boldsymbol{\Theta}\}$, where

$$\boldsymbol{\Delta}_T(\boldsymbol{\vartheta}, X^T) = \frac{1}{\sqrt{T}} \int_0^T \frac{\dot{\boldsymbol{\mu}}_{\mathrm{WF}}(\boldsymbol{\vartheta}, X_t)}{\alpha_{\mathrm{WF}}(X_t)} dW_t.$$

In particular the result holds for $\nu = f_{\boldsymbol{\vartheta}}$.

Proof. From (2.27), we have that the log-likelihood ratio is given by

$$\begin{aligned}
\log Z_{T,\boldsymbol{\vartheta}}(\mathbf{u}) = \; & \log \frac{\nu(\boldsymbol{\vartheta} + \frac{\mathbf{u}}{\sqrt{T}}, X_0)}{\nu(\boldsymbol{\vartheta}, X_0)} \\
& + \int_0^T \frac{1}{2}\left(\frac{u_1}{\sqrt{T}} \sqrt{X_t(1-X_t)} + \frac{u_2}{\sqrt{T}} \sqrt{\frac{1-X_t}{X_t}} \right. \\
& \qquad \left. - \frac{u_3}{\sqrt{T}} \sqrt{\frac{X_t}{1-X_t}} \right) dW_t \\
& - \frac{1}{2} \int_0^T \frac{1}{4}\left(\frac{u_1}{\sqrt{T}} \sqrt{X_t(1-X_t)} + \frac{u_2}{\sqrt{T}} \sqrt{\frac{1-X_t}{X_t}} \right. \\
& \qquad \left. - \frac{u_3}{\sqrt{T}} \sqrt{\frac{X_t}{1-X_t}} \right)^2 dt \\
= \; & \log \frac{\nu(\boldsymbol{\vartheta} + \frac{\mathbf{u}}{\sqrt{T}}, X_0)}{\nu(\boldsymbol{\vartheta}, X_0)} + \left\langle \mathbf{u}, \boldsymbol{\Delta}_T(\boldsymbol{\vartheta}, X^T) \right\rangle - \frac{1}{2} \left\langle \mathbf{I}(\boldsymbol{\vartheta})\mathbf{u}, \mathbf{u} \right\rangle \\
& + \frac{1}{2} \left\langle \mathbf{I}(\boldsymbol{\vartheta})\mathbf{u}, \mathbf{u} \right\rangle - \frac{1}{2T} \int_0^T \frac{\left\langle \mathbf{u}, \dot{\boldsymbol{\mu}}_{\mathrm{WF}}(\boldsymbol{\vartheta}, X_t) \right\rangle^2}{\alpha^2_{\mathrm{WF}}(X_t)} dt, \qquad (2.32)
\end{aligned}$$

where

$$\mathbf{I}(\boldsymbol{\vartheta}) = \mathbb{E}^{(\boldsymbol{\vartheta})} \left[\frac{1}{4} \begin{pmatrix} \xi(1-\xi) & 1-\xi & -\xi \\ 1-\xi & \frac{1-\xi}{\xi} & -1 \\ -\xi & -1 & \frac{\xi}{1-\xi} \end{pmatrix} \right].$$

Setting

$$\begin{aligned}
r_T(\boldsymbol{\vartheta}, \mathbf{u}, X^T) := & \log \frac{\nu(\boldsymbol{\vartheta} + \frac{\mathbf{u}}{\sqrt{T}}, X_0)}{\nu(\boldsymbol{\vartheta}, X_0)} + \frac{1}{2} \left\langle \mathbf{I}(\boldsymbol{\vartheta})\mathbf{u}, \mathbf{u} \right\rangle \\
& - \frac{1}{2T} \int_0^T \frac{\left\langle \mathbf{u}, \dot{\boldsymbol{\mu}}_{\mathrm{WF}}(\boldsymbol{\vartheta}, X_t) \right\rangle^2}{\alpha^2_{\mathrm{WF}}(X_t)} dt,
\end{aligned}$$

we show that (2.29) holds. The first term goes to 0 as $T \to \infty$ by (2.30), and in particular $\nu = f_{\boldsymbol{\vartheta}}$ as given in (2.22) is continuous in $\boldsymbol{\vartheta}$.

Thus we deduce that (2.29) follows if we can prove that for any $\varepsilon > 0$

$$\lim_{T\to\infty} \sup_{\boldsymbol{\vartheta}\in\mathcal{K}} \mathbb{P}_{\nu}^{(\boldsymbol{\vartheta})} \left[\left| \frac{1}{T} \int_0^T \frac{\langle \mathbf{u}, \dot{\boldsymbol{\mu}}_{\mathrm{WF}}(\boldsymbol{\vartheta}, X_t)\rangle^2}{\alpha^2_{\mathrm{WF}}(X_t)} dt - \langle \mathbf{I}(\boldsymbol{\vartheta})\mathbf{u}, \mathbf{u}\rangle \right| > \varepsilon \right] = 0. \tag{2.33}$$

Observe that the expression inside the probability in (2.33) is made up of six distinct differences between the averages of the six distinct entries of the Fisher information matrix with respect to time and the stationary density. Thus if we are able to show that each individual difference displays the same convergence as in (2.3), (2.33) follows. Now, as

$$\begin{aligned}
\frac{\langle \mathbf{u}, \dot{\boldsymbol{\mu}}_{\mathrm{WF}}(\boldsymbol{\vartheta}, x)\rangle^2}{\alpha^2_{\mathrm{WF}}(x)} &= \frac{1}{4}\left(u_1\sqrt{x(1-x)} + u_2\sqrt{\frac{1-x}{x}} - u_3\sqrt{\frac{x}{1-x}} \right)^2 \\
&= \frac{1}{4}\Bigg(u_1^2 x(1-x) + 2u_1u_2(1-x) - 2u_1u_3 x - 2u_2u_3 \\
&\qquad + u_2^2\frac{1-x}{x} + u_3^2\frac{x}{1-x} \Bigg)
\end{aligned}$$

using (2.21), we can apply Corollary 2.5 to the first four terms directly. The remaining two differences involve the unbounded functions $(1-x)x^{-1}$ and $x(1-x)^{-1}$, for which (2.33) has been shown to hold in Corollary 2.6 with ν satisfying (2.31). Thus (2.29) holds (we also show in Corollary 2.6 that (2.31) holds in the case $\nu = f_{\boldsymbol{\vartheta}}$), and (2.28) follows from Proposition 1.20 in [Kut04] which we can invoke in view of the above proved (2.33) and the fact that

$$\sup_{\boldsymbol{\vartheta}\in\mathcal{K}} \sqrt{\langle \mathbf{I}(\boldsymbol{\vartheta})\mathbf{u}, \mathbf{u}\rangle} < \infty.$$

□

We point out here that if the mutation parameters are known, condition (2.31) becomes redundant and Theorem 2.8 holds for any initial distribution satisfying $\lim_{\varepsilon\to 0} \nu(\sigma+\varepsilon, x)/\nu(\sigma, x) = 1$ for any $x \in [0,1]$.

Chapter 3

Properties of the ML & Bayesian Estimators

Inference for scalar diffusions, particularly proving consistency of estimators under specific observational regimes, has generated considerable interest over the past few years [GS12, Kut04, NR20, NS17, PvZ09, vdMvZ13, vZ01, Wat79]. However, most of the work so far has considered classes of diffusions which directly exclude the Wright–Fisher diffusion, for instance by imposing periodic boundary conditions on the drift coefficients or by requiring the diffusion coefficient to be strictly positive everywhere. The asymptotic study of a variety of estimators for continuously observed ergodic scalar diffusions has been entertained in great depth in [Kut04]; see in particular Theorems 2.8 and 2.13 in [Kut04], which are respectively adaptations of Theorems I.5.1, I.10.1 and I.5.2, I.10.2 in [IH81]. However Theorems 2.8 and 2.13 in [Kut04] cannot be applied directly to the Wright–Fisher diffusion as certain conditions do not hold, namely the reciprocal of the diffusion coefficient does not have a polynomial majorant. This discrepancy makes replicating the results for the Wright–Fisher diffusion with selection and mutation highly non-trivial.

In this chapter we show how the same set of desirable properties hold for the Maximum Likelihood (ML) and Bayesian estimators

for selection in the continuous observation regime, by exploiting the explicit nature of (2.21), as well as Corollary 2.5 and Theorem 2.8 from the previous chapter. In particular, we show in Theorem 3.2, that these estimators are uniform in the selection parameter over compact sets consistent, asymptotically normal, display moment convergence, and (for a specific class of loss functions) are asymptotically efficient. We achieve this by showing that the conditions of Theorems I.5.1, I.10.1 and I.5.2, I.10.2 in [IH81] (which we have combined and translated to the scalar Wright–Fisher setting in Theorem 3.3) still hold for the Wright–Fisher diffusion. We point out further that the uniformity in our results is particularly useful as it controls the lowest rate (over the true parameters) at which the parameters of interest are being learned by the inferential scheme.

The Wright–Fisher diffusion with selection but without mutation was tackled specifically by Watterson in [Wat79]. Having no mutation ensures that the diffusion is absorbed at either boundary point 0 or 1 in finite time almost surely, and by conditioning on absorption, Watterson computes the moment generating function, proves asymptotic normality, and derives hypothesis tests for the Maximum Likelihood Estimator (MLE). Watterson's work however does not address the Bayesian estimator, nor does it readily extend to the case when mutation is present because then the diffusion is no longer absorbed at the boundaries. In this sense, the results obtained in Theorem 3.2 are complementary to those obtained by Watterson under the assumption that the mutation parameters are known. Although this is a restriction, because we are observing the path continuously over the interval $[0, T]$ and subsequently sending $T \to \infty$, these parameters can be inferred by considering the boundary behaviour of the diffusion. In particular, when either mutation parameter is less than 1, the diffusion hits the corresponding boundary in finite time almost surely. Further, as the diffusion approaches the boundary, the diffusion coefficient (i.e. noise) vanishes, and in fact it vanishes sufficiently quickly on the approach to the boundary that the mutation parameters can be inferred without error as soon as the boundary

is first hit. For mutation parameters greater than or equal to 1, the corresponding boundary point is no longer attainable but the diffusion can get arbitrarily close to it as $T \to \infty$, and a similar argument enables the mutation parameters again to be inferred (see [PY81, Remark 2.2] for a related argument applied to the squared Bessel process).

For the rest of this chapter we shall always assume that observations are available as the entire trajectory $(X_t)_{t=0}^T \in C_T([0,1])$ up to some terminal time T, i.e. a continuous mapping from $[0,T]$ into $[0,1]$, which means that the paths are observed without error. We point out here that this setup allows us to establish and explicitly analyse the statistical error produced by an estimator based on the whole sample path when sending $T \to \infty$, which then clearly illustrates the statistical limitations of alternative estimators based on less informative (e.g. discrete) observations. In a discrete observation setting, in addition to the above mentioned statistical error, one also has to deal with observational error. One certainly cannot hope for an estimator that performs better in a discrete setting than in a continuous one, so our analysis may be viewed as the 'best possible' performance for inference from a discretely observed model.

Apart from the theoretical guarantees that our results provide, we also perform some simulations to illustrate the conclusions of Theorem 3.2. In particular, by making use of the exact algorithm for the Wright–Fisher diffusion ([JS17], see Section 4.2 in the next chapter for a brief overview), we generate datapoints over a time grid, and apply Riemann sum approximations to get an estimate of the MLE (which cannot be evaluated exactly due to the intractable integrals involved). The plots obtained provide empirical evidence that the MLE is consistent, and displays both convergence in distribution and convergence of moments, thereby not only reinforcing the theoretical results derived in Section 3.2, but also suggesting that similar conclusions might hold for the discrete observation case (outside the scope of our results), as long as the sampling regime converges to densely sampled data.

The rest of this chapter is organised as follows: Section 3.1 introduces the relevant notation, defining both the ML and Bayesian estimators, together with the likelihood ratio function which will be the main object of interest. We also further explain why assuming that the mutation parameters are a priori known is not as restrictive an assumption as it might seem in view of the observational regime entertained. In Section 3.2 we state the main result, Theorem 3.2, guaranteeing a set of desirable properties for both the ML and Bayesian estimators, and illustrate how it can be proved by appealing to Theorems I.5.1, I.5.2, I.10.1 and I.10.2 in [IH81] in view of Propositions 3.5, 3.6, 3.7 and Corollary 3.4. Simulations providing empirical evidence of the results proved can be found in Section 3.3, whilst Section 3.4 briefly illustrates the difficulties involved in extending the above framework to allow for the joint inference of selection and mutation parameters.

3.1 Definitions and notation

We henceforth assume that the mutation parameters $\theta_1, \theta_2 > 0$ are known, and thus focus on conducting inference solely on the selection parameter $\sigma \in \mathcal{S} \subset \mathbb{R}$ with $\mathcal{S}$ open and bounded.

Remark 3.1. The continuous observation regime entertained here would enable one to infer the mutation parameters: on $\boldsymbol{\vartheta} \in \mathbb{R} \times (0,1)^2$ this is immediate as the family of measures $\{\mathbb{P}_\nu^{(\boldsymbol{\vartheta})} : \boldsymbol{\vartheta} \in \mathbb{R} \times (0,1)^2\}$ are mutually singular. In particular, when either mutation parameter is less than 1, the diffusion hits the corresponding boundary in finite time almost surely, and as it does so, the diffusion coefficient (i.e. noise) vanishes sufficiently quickly allowing the mutation parameters to be inferred without error. Indeed, by looking at the integrands on the RHS of (2.32), we observe that as the path approaches either boundary, the likelihood ratio explodes. On $\boldsymbol{\vartheta} \in \mathbb{R} \times [1,\infty)^2$ the family of measures $\{\mathbb{P}_\nu^{(\boldsymbol{\vartheta})} : \boldsymbol{\vartheta} \in \mathbb{R} \times [1,\infty)^2\}$ are now mutually absolutely continuous, with both boundary points unattainable. However, the process can get arbitrarily close to either boundary as $T \to \infty$, and again the noise

vanishes sufficiently quickly that the corresponding mutation parameters can be inferred to any required precision. In the case when one mutation parameter is less than 1 and the other is greater than or equal to 1, similar arguments apply.

Actually incorporating the mutation parameter into the inferential setup below leads to some technical difficulties which we discuss briefly at the end of this chapter in Section 3.4, so we will henceforth assume them to be known. Nonetheless all the notation and definitions introduced in the previous chapter carry through by replacing $\boldsymbol{\vartheta}$ by σ.

We start by defining the MLE $\hat{\sigma}_T$ of σ in (2.21) as

$$\hat{\sigma}_T = \arg\sup_{\sigma\in\mathcal{S}} \frac{d\mathbb{P}_\nu^{(\sigma)}}{d\mathbb{P}_\nu^{(\sigma_0)}}(X^T) \tag{3.1}$$

where $\sigma_0 \in \mathcal{S}$ is arbitrary and its only role is to specify a reference measure whose exact value does not matter. Observe that now (2.27) simplifies to

$$\begin{aligned}\frac{d\mathbb{P}_\nu^{(\sigma')}}{d\mathbb{P}_\nu^{(\sigma)}}(X^T) = \frac{\nu(\sigma', X_0)}{\nu(\sigma, X_0)} \\ \times \exp\Bigg\{ \int_0^T (\sigma' - \sigma)\sqrt{X_t(1-X_t)}dW_t \\ - \frac{1}{2}\int_0^T (\sigma' - \sigma)^2 X_t(1-X_t)dt \Bigg\},\end{aligned} \tag{3.2}$$

with $\mathbb{P}_\nu^{(\sigma)}$-probability 1 when σ is the true value, with initial distributions $\{\nu(\sigma,\cdot)\}_{\sigma\in\mathcal{S}}$ admitting a density (which, again in abuse of notation, we denote $\nu(\sigma,\cdot)$) with respect to some common dominating measure $\lambda(\cdot)$. In order to be able to define the Bayesian estimator, we introduce the class $\mathscr{W}_p$ of loss functions $\ell : \mathcal{S} \to \mathbb{R}_+$ for which the following stipulations are satisfied:

A1. $\ell(\cdot)$ is even, non-negative, and continuous at 0 with $\ell(0) = 0$ but not identically zero.

A2. The sets $\{u \in \mathcal{S} : \ell(u) < c\}$ are convex $\forall c > 0$ (and thus $\ell(\cdot)$ is

non-decreasing).

A3. $\ell(\cdot)$ has a polynomial majorant, i.e. there exist strictly positive constants A and b such that for any $u \in \mathcal{S}$,

$$|\ell(u)| \leq A(1 + |u|^b)$$

A4. For any $H > 0$ sufficiently large and for sufficiently small γ, it holds that

$$\inf_{|u|>H} \ell(u) - \sup_{|u|\leq H^\gamma} \ell(u) \geq 0.$$

As remarked above, we assume that $\mathcal{S}$ is an open and bounded subset of $\mathbb{R}$, and we denote by $p(\cdot)$ the prior density on $\mathcal{S}$, which we assume belongs to

$$\mathscr{P}_c := \left\{ p(\cdot) \in C(\bar{\mathcal{S}}, \mathbb{R}_+) : p(u) \leq A(1 + |u|^b) \forall u \in \bar{\mathcal{S}}, \int_{\bar{\mathcal{S}}} p(u)du = 1 \right\},$$

where A and b are some strictly positive constants, and $\bar{\mathcal{S}}$ denotes the closure of $\mathcal{S}$. With $p(\cdot) \in \mathscr{P}_c$ and $\ell(\cdot) \in \mathscr{W}_p$, we define the Bayesian estimator $\tilde{\sigma}_T$ of σ in (2.21) as

$$\tilde{\sigma}_T = \arg\min_{\bar{\sigma}_T} \int_{\mathcal{S}} \mathbb{E}_\nu^{(\sigma)} \left[\ell \left(\sqrt{T} \left(\bar{\sigma}_T - \sigma \right) \right) \right] p(\sigma) ds,$$

where the minimization is over estimators $\bar{\sigma}_T = \bar{\sigma}_T(X^T)$. We introduce the last class of functions we will need, namely denote by $\mathscr{G}$ the class of functions satisfying the following two conditions:

1. For a fixed $T > 0$, $g_T(\cdot)$ is a monotonically increasing function on $[0, \infty)$, with $g_T(y) \to \infty$ as $y \to \infty$.

2. For any $N > 0$,
$$\lim_{\substack{T\to\infty \\ y\to\infty}} y^N e^{-g_T(y)} = 0.$$

Observe that the likelihood ratio function is now given by

$$\begin{aligned} Z_{T,\sigma}(u) :&= \frac{d\mathbb{P}_\nu^{(\sigma+\frac{u}{\sqrt{T}})}}{d\mathbb{P}_\nu^{(\sigma)}}(X^T) \\ &= \frac{\nu(\sigma+\frac{u}{\sqrt{T}}, X_0)}{\nu(\sigma, X_0)} \\ &\quad \times \exp\left\{ \left(\frac{u}{2\sqrt{T}}\right) \int_0^T \sqrt{X_t(1-X_t)} dW_t \right. \\ &\qquad \left. - \frac{1}{2}\left(\frac{u}{2\sqrt{T}}\right)^2 \int_0^T X_t(1-X_t)dt \right\} \end{aligned} \tag{3.3}$$

for

$$u \in \mathcal{U}_{T,\sigma} := \left\{ u \in \mathbb{R} : \sigma + \frac{u}{\sqrt{T}} \in \mathcal{S} \right\}. \tag{3.4}$$

3.2 Properties of the ML and Bayesian estimators for selection in the scalar Wright–Fisher diffusion

We now present the main result of this chapter which states that the ML and Bayesian estimators for σ have a set of desirable properties. We prove this by showing that the conditions of Theorems I.5.1, I.5.2, I.10.1, and I.10.2 in [IH81] are satisfied for the Wright–Fisher diffusion. A similar formulation of the result below for the general case of a continuously observed diffusion on $\mathbb{R}$ can be found in Theorems 2.8 and 2.13 in [Kut04], where the author proves that the conditions necessary to invoke Theorems I.5.1, I.5.2, I.10.1, and I.10.2 in [IH81] hold for a certain class of diffusions. However, this class includes only scalar diffusions for which the inverse of the diffusion coefficient has a polynomial majorant. This fails to hold in our case, forcing us to seek alternative ways to prove that the conditions of the above mentioned theorems hold.

Theorem 3.2. *Let $\bar{\sigma}_T$ be either the ML or Bayesian estimator for the selection parameter $\sigma \in \mathcal{S}$ (for open bounded $\mathcal{S} \subset \mathbb{R}$) in the neutral*

or non-neutral Wright–Fisher diffusion (2.21) with initial distribution satisfying

$$\lim_{\varepsilon\to 0}\frac{\nu(\sigma+\varepsilon,x)}{\nu(\sigma,x)}=1,\quad \forall x\in[0,1],$$

and such that for any $M\geq 2$ *and* $u\in\mathcal{U}_{T,\sigma}$,

$$\mathbb{P}_{\nu}^{(\sigma)}\left[\left|\log\left(\frac{\nu(\sigma+\frac{u}{\sqrt{T}},X_0)}{\nu(\sigma,X_0)}\right)\right|>\mathbb{E}^{(\sigma)}\left[\xi(1-\xi)\right]|u|^2\right]\leq\frac{C_1}{|u|^M}$$

and for any $R>0$ *and* $u,v\in\mathcal{U}_{T,\sigma}$ *with* $|u|<R,|v|<R$

$$\int_0^1\left|\nu\left(\sigma+\frac{u}{\sqrt{T}},x\right)^{\frac{1}{2}}-\nu\left(\sigma+\frac{v}{\sqrt{T}},x\right)^{\frac{1}{2}}\right|^2\lambda(dx)\leq C_2|u-v|^2$$

for some constants $C_1,C_2>0$, *and* $\lambda(\cdot)$ *common dominating measure introduced below* (3.2) *(in particular these conditions hold for the case* $\nu=f_\sigma$, *the stationary density). Then* $\bar{\sigma}_T$ *is uniformly over compact sets* $\mathcal{K}\subset\mathcal{S}$ *consistent, i.e. for any* $\varepsilon>0$

$$\lim_{T\to\infty}\sup_{\sigma\in\mathcal{K}}\mathbb{P}_{\nu}^{(\sigma)}\left[\,|\bar{\sigma}_T-\sigma|>\varepsilon\right]=0;$$

it converges in distribution to a normal random variable

$$\sqrt{T}\,(\bar{\sigma}_T-\sigma)\xrightarrow{\mathrm{d}}N(0,I(\sigma)^{-1}),$$

uniformly in $\sigma\in\mathcal{K}$, *where*

$$I(\sigma)=\frac{1}{4}\mathbb{E}^{(\sigma)}\left[\xi(1-\xi)\right];$$

and it displays moment convergence for any $p>0$

$$\lim_{T\to\infty}\mathbb{E}_{\nu}^{(\sigma)}\left[\left|\sqrt{T}\,(\bar{\sigma}_T-\sigma)\right|^p\right]=\mathbb{E}\left[\left|I(\sigma)^{-\frac{1}{2}}\zeta\right|^p\right],$$

uniformly in $\sigma\in\mathcal{K}$, *where* $\zeta\sim N(0,1)$, *for any compact set* $\mathcal{K}\subset\mathcal{S}$. *Furthermore, if the loss function* $\ell(\cdot)\in\mathscr{W}_p$, *then* $\bar{\sigma}_T$ *is also asymptoti-*

cally efficient, i.e.

$$\lim_{\delta\to 0}\lim_{T\to\infty}\sup_{\sigma:|\sigma-\sigma_0|<\delta}\mathbb{E}_\nu^{(\sigma)}\left[\ell\left(\sqrt{T}\left(\bar{\sigma}_T-\sigma\right)\right)\right]=\mathbb{E}\left[\ell\left(I(\sigma_0)^{-\frac{1}{2}}\zeta\right)\right],$$

holds for all $\sigma_0\in\mathcal{S}$, where $\zeta\sim N(0,1)$.

As mentioned above, the proof relies on Theorems I.5.1, I.5.2, I.10.1, and I.10.2 in [IH81], which for reference we combine together in our notation into Theorem 3.3 below. Establishing that the conditions of Theorem 3.3 hold for the Wright–Fisher diffusion is non-trivial as the standard arguments found in [Kut04] no longer hold, and will thus be the main focus of this chapter. The conclusions of Theorems I.5.1 and I.5.2 guarantee the uniform over compact sets consistency for the MLE and Bayesian estimator respectively, and also give that for any $\varepsilon>0$ and for sufficiently large T

$$\sup_{\sigma\in\mathcal{K}}\mathbb{P}_\nu^{(\sigma)}\left[\left|\sqrt{T}\left(\bar{\sigma}_T-\sigma\right)\right|>\varepsilon\right]\leq\alpha e^{-\beta g_T(\varepsilon)}$$

with α,β strictly positive constants, and $g_T\in\mathscr{G}$. On the other hand, Theorems I.10.1 and I.10.2 provide the necessary conditions to deduce the uniform in $\sigma\in\mathcal{K}$ asymptotic normality and convergence of moments for compact $\mathcal{K}\subset\mathcal{S}$, as well as asymptotic efficiency.

Theorem 3.3 (Ibragimov–Has'minskii)**.** *Let $\bar{\sigma}_T$ denote either the ML or Bayesian estimator for the parameter $\sigma\in\mathcal{S}$, for open bounded $\mathcal{S}\subset\mathbb{R}$, in (2.21), with prior density $p(\cdot)\in\mathscr{P}_c$, and loss function $\ell(\cdot)\in\mathscr{W}_p$. Suppose further that the following conditions are satisfied by the likelihood ratio function $Z_{T,\sigma}(u)$ as defined in (3.3):*

1. *$\forall\mathcal{K}\subset\mathcal{S}$ compact, we can find constants a and B, and functions $g_T(\cdot)\in\mathscr{G}$ (all of which depend on $\mathcal{K}$) such that the following two conditions hold:*
   - *$\forall R>0$, $\forall u,v\in\mathcal{U}_{T,\sigma}$ as defined in (3.4) satisfying $|u|<R$,*

$|v| < R$, *and for some* $m \geq q > 1$

$$\sup_{\sigma \in \mathcal{K}} \mathbb{E}_{\nu}^{(\sigma)} \left[\left| Z_{T,\sigma}(u)^{\frac{1}{m}} - Z_{T,\sigma}(v)^{\frac{1}{m}} \right|^{m} \right] \leq B(1 + R^{a})|u - v|^{q}. \tag{3.5}$$

- $\forall u \in \mathcal{U}_{T,\sigma}$

$$\sup_{\sigma \in \mathcal{K}} \mathbb{E}_{\nu}^{(\sigma)} \left[Z_{T,\sigma}(u)^{\frac{1}{2}} \right] \leq e^{-g_T(|u|)}.$$

2. *The random functions* $Z_{T,\sigma}(u)$ *have marginal distributions which converge uniformly in* $\sigma \in \mathcal{K}$ *as* $T \to \infty$ *to those of the random function* $Z_{\sigma}(u) \in C_0(\mathbb{R})$*, where* $C_0(\mathbb{R})$ *denotes the space of continuous functions on* $\mathbb{R}$ *vanishing at infinity, equipped with the supremum norm and the Borel* σ*-algebra.*

3. *The limit function* $Z_{\sigma}(u)$ *and the random function*

$$\psi(v) = \int_{\mathbb{R}} \ell(v - u) \frac{Z_{\sigma}(u)}{\int_{\mathbb{R}} Z_{\sigma}(y)dy} du$$

attain their maximum and minimum values respectively at a unique point $\bar{u}(\sigma) = u$ *with probability 1.*

Then we have that $\bar{\sigma}_T$ *is: uniformly in* $\sigma \in \mathcal{K}$ *consistent, i.e. for any* $\varepsilon > 0$

$$\lim_{T \to \infty} \sup_{\sigma \in \mathcal{K}} \mathbb{P}_{\nu}^{(\sigma)} \left[\, |\bar{\sigma}_T - \sigma| > \varepsilon \right] = 0,$$

the distributions of the random variables $\bar{u}_T = \sqrt{T}\,(\bar{\sigma}_T - \sigma)$ *converge uniformly in* $\sigma \in \mathcal{K}$ *to the distribution of* $\bar{u}$*, and for any loss function* $\ell \in \mathscr{W}_p$ *uniformly in* $\sigma \in \mathcal{K}$

$$\lim_{T \to \infty} \mathbb{E}_{\nu}^{(\sigma)} \left[\ell \left(\sqrt{T}\,(\bar{\sigma}_T - \sigma) \right) \right] = \mathbb{E}_{\nu}^{(\sigma)} \left[\ell(\bar{u}) \right]. \tag{3.6}$$

For the Bayesian estimator, the requirements for inequality (3.5) can be weakened as it suffices to show that (3.5) holds for $m = 2$ and any $q > 0$.

Proof of Theorem 3.2. Our aim will be to prove that Conditions 1, 2, and 3 in Theorem 3.3 hold for the Wright–Fisher diffusion, for then the ML and Bayesian estimators are uniformly on compact sets consistent. Below, Condition 1 is shown to hold in Propositions 3.5 and 3.6; Condition 2 is shown in Corollary 3.4; and Condition 3 is shown in Proposition 3.7.

It remains to show how uniform in $\sigma \in \mathcal{K}$ asymptotic normality and convergence of moments, as well as asymptotic efficiency (under the right choice of loss function) follow. Given Conditions 1, 2, and 3 of Theorem 3.2, uniform in $\sigma \in \mathcal{K}$ asymptotic normality follows immediately from Proposition 3.7; $\bar{u} = I(\sigma)^{-1}\Delta(\sigma)$, $\Delta(\sigma) \sim N(0, I(\sigma))$, and $\bar{u}_T$ converges uniformly in distribution to $\bar{u}$. Moreover, as stated in Remark I.5.1 in [IH81], the Ibragimov–Has'minskii conditions also give us a bound on the tails of the likelihood ratio, which can be translated into bounds on the tails of $|\hat{u}_T|^p$ for any $p > 0$ (see the display just below (2.27) in [Kut04]). Similar bounds on the tails of $|\tilde{u}_T|^p$ hold for the Bayesian estimator by Theorem I.5.7 in [IH81], and thus we have that the random variables $|\bar{u}_T|^p$ are uniformly integrable for any $p > 0$, uniformly in $\sigma \in \mathcal{K}$ for any compact $\mathcal{K} \subset \mathcal{S}$. Uniform convergence of the moments of the estimators follows from this and the uniform convergence in distribution (by applying a truncation argument).

For loss functions satisfying $\ell(\cdot) \in \mathscr{W}_p$, observe that the uniform convergence in (3.6) allows us to deduce that

$$\lim_{T\to\infty} \sup_{\sigma:|\sigma-\sigma_0|<\delta} \mathbb{E}_\nu^{(\sigma)} \left[\ell\left(\sqrt{T}\,(\bar{\sigma}_T - \sigma)\right)\right] = \sup_{\sigma:|\sigma-\sigma_0|<\delta} \mathbb{E}\left[\ell\left(I(\sigma)^{-\frac{1}{2}}\zeta\right)\right]$$

for $\zeta \sim N(0,1)$. As $I(\sigma)$ is continuous in σ, we have that

$$\lim_{\delta\to 0} \sup_{\sigma:|\sigma-\sigma_0|<\delta} \mathbb{E}\left[\ell\left(I(\sigma)^{-\frac{1}{2}}\zeta\right)\right] = \mathbb{E}\left[\ell\left(I(\sigma_0)^{-\frac{1}{2}}\zeta\right)\right],$$

giving asymptotic efficiency. □

We proceed to show that Conditions 1, 2, and 3 in Theorem 3.3 hold

for the Wright–Fisher diffusion. Theorem 2.8 gives us that the Wright–Fisher diffusion is uniformly LAN, which immediately gives the required marginal convergence of the $Z_{T,\sigma}(u)$ in Condition 2.

Corollary 3.4 (Corollary to Theorem 2.8)**.** *For any initial distribution satisfying*

$$\lim_{\varepsilon \to 0} \frac{\nu(\sigma + \varepsilon, x)}{\nu(\sigma, x)} = 1 \quad \forall x \in [0,1],$$

the random functions $Z_{T,\sigma}(u)$ given by

$$\begin{aligned} Z_{T,\sigma}(u) &= \exp\left\{ \frac{u}{2\sqrt{T}} \int_0^T \sqrt{X_t(1-X_t)} dW_t - \frac{u^2}{8} \mathbb{E}^{(\sigma)}\left[\xi(1-\xi)\right] \right. \\ &\qquad \left. + r_T(\sigma, u, X^T) \right\} \\ &=: \exp\left\{ u\Delta_T(\sigma) - \frac{u^2}{2} I(\sigma) + r_T(\sigma, u, X^T) \right\}, \end{aligned}$$

where

$$\begin{aligned} r_T(\sigma, u, X^T) := \log\left(\frac{\nu(\sigma + \frac{u}{\sqrt{T}}, X_0)}{\nu(\sigma, X_0)} \right) + \frac{u^2}{8} \mathbb{E}^{(\sigma)}\left[\xi(1-\xi)\right] \\ - \frac{1}{2}\left(\frac{u}{2\sqrt{T}} \right)^2 \int_0^T X_t(1-X_t) dt, \end{aligned}$$

have marginal distributions which converge uniformly in $\sigma \in \mathcal{K}$ as $T \to \infty$ to those of the random function $Z_\sigma(u) \in C_0(\mathbb{R})$ given by

$$Z_\sigma(u) := \exp\left\{ u\Delta(\sigma) - \frac{u^2}{2} I(\sigma) \right\},$$

where

$$\Delta(\sigma) := \lim_{T \to \infty} \frac{1}{2\sqrt{T}} \int_0^T \sqrt{X_t(1-X_t)} dW_t \sim N\left(0, I(\sigma)\right).$$

Proof. The result follows immediately from the uniform LAN of the family of measures as shown in Theorem 2.8; see for illustration the display just before Lemma 2.10 in [Kut04]. It is clear that $Z_\sigma(u)$ van-

ishes at infinity and thus is an element of $C_0(\mathbb{R})$. □

The next two results allow us to control the Hellinger distance of the likelihood ratio function as required by Condition 1 in Theorem 3.3.

Proposition 3.5. *Let the initial distribution be such that for any $R > 0$ and for $u, v \in \mathcal{U}_{T,\sigma}$ as defined in (3.4) with $|u| < R, |v| < R$*

$$\int_0^1 \left| \nu\left(\sigma + \frac{u}{\sqrt{T}}, x\right)^{\frac{1}{2}} - \nu\left(\sigma + \frac{v}{\sqrt{T}}, x\right)^{\frac{1}{2}} \right|^2 \lambda(dx) \leq c\,|u - v|^2 \quad (3.7)$$

for some constant $c > 0$ with dominating measure $\lambda(\cdot)$ as specified below (3.2). Then for any $\mathcal{K} \subset \mathcal{S}$ compact, we can find a constant C such that for any $R > 0$, and for any $u, v \in \mathcal{U}_{T,\sigma}$ as defined in (3.4) satisfying $|u| < R$, $|v| < R$, the following holds

$$\sup_{\sigma \in \mathcal{K}} \mathbb{E}_\nu^{(\sigma)} \left[\left| Z_{T,\sigma}(u)^{\frac{1}{2}} - Z_{T,\sigma}(v)^{\frac{1}{2}} \right|^2 \right] \leq C(1 + R^2)|u - v|^2.$$

In particular the result holds for $\nu = f_\sigma$.

Proof. In what follows we denote by C_i, for $i \in \mathbb{N}$, constants which do not depend on u, v, σ, or T. Observe that for any $\sigma', \sigma^* \in \mathcal{S}$ it holds that

$$\begin{aligned}
&\mathbb{E}_\nu^{(\sigma')} \left[\int_0^T \left| \frac{\mu_{\mathrm{WF}}(\sigma', X_t) - \mu_{\mathrm{WF}}(\sigma^*, X_t)}{\alpha_{\mathrm{WF}}(X_t)} \right|^4 dt \right] \\
&\quad = \mathbb{E}_\nu^{(\sigma')} \left[\int_0^T \left| \frac{(\sigma' - \sigma^*)}{2} \sqrt{X_t(1 - X_t)} \right|^4 dt \right] \\
&\quad \leq \left(\frac{\sigma' - \sigma^*}{4} \right)^4 T < \infty,
\end{aligned}$$

and so we can use Lemma 1.13 and Remark 1.14 from [Kut04] (as done

in Lemma 2.10 there) to split the expectation in (3.5) into three

$$\begin{aligned}
&\mathbb{E}_{\nu}^{(\sigma)}\left[\left|Z_{T,\sigma}^{\frac{1}{2}}(u)-Z_{T,\sigma}^{\frac{1}{2}}(v)\right|^2\right] \\
&\quad\leq C_1\int_0^1\left|\nu(\sigma_u,x)^{\frac{1}{2}}-\nu(\sigma_v,x)^{\frac{1}{2}}\right|^2\lambda(dx) \\
&\qquad+C_2\int_0^T\mathbb{E}_{\nu}^{(\sigma_v)}\left[\left(\frac{\mu_{\mathrm{WF}}(\sigma_u,X_t)-\mu_{\mathrm{WF}}(\sigma_v,X_t)}{\alpha_{\mathrm{WF}}(X_t)}\right)^2\right]dt \\
&\qquad+C_3T\int_0^T\mathbb{E}_{\nu}^{(\sigma_v)}\left[\left(\frac{\mu_{\mathrm{WF}}(\sigma_u,X_t)-\mu_{\mathrm{WF}}(\sigma_v,X_t)}{\alpha_{\mathrm{WF}}(X_t)}\right)^4\right]dt, \quad (3.8)
\end{aligned}$$

where we denote $\sigma_u=\sigma+u/\sqrt{T}$ and $\sigma_v=\sigma+v/\sqrt{T}$. The first term on the RHS of (3.8) can be dealt with using (3.7), whilst for the second term observe that

$$\begin{aligned}
&\int_0^T\mathbb{E}_{\nu}^{(\sigma_v)}\left[\left(\frac{\mu_{\mathrm{WF}}(\sigma_u,X_t)-\mu_{\mathrm{WF}}(\sigma_v,X_t)}{\alpha_{\mathrm{WF}}(X_t)}\right)^2\right]dt \\
&\qquad=\frac{|u-v|^2}{4T}\int_0^T\mathbb{E}_{\nu}^{(\sigma_v)}\left[X_t(1-X_t)\right]dt \\
&\qquad\leq\frac{1}{16}|u-v|^2.
\end{aligned}$$

Therefore

$$\begin{aligned}
&C_2\int_0^T\mathbb{E}_{\nu}^{(\sigma_v)}\left[\left(\frac{\mu_{\mathrm{WF}}(\sigma_u,X_t)-\mu_{\mathrm{WF}}(\sigma_v,X_t)}{\alpha_{\mathrm{WF}}(X_t)}\right)^2\right]dt \\
&\qquad\leq C_4|u-v|^2. \quad (3.9)
\end{aligned}$$

A similar calculation can be performed for the third term in (3.8) to get

$$\begin{aligned}
&C_3T\int_0^T\mathbb{E}_{\nu}^{(\sigma_v)}\left[\left(\frac{\mu_{\mathrm{WF}}(\sigma_u,X_t)-\mu_{\mathrm{WF}}(\sigma_v,X_t)}{\alpha_{\mathrm{WF}}(X_t)}\right)^4\right]dt \\
&\qquad\leq C_5|u-v|^4,
\end{aligned}$$

and thus the result holds in view of the fact that $|u|,|v|<R$.
It remains to show that (3.7) holds for $\nu=f_\sigma$. To this end, observe

that

$$\int_0^1 \left| f_{\sigma_u}(x)^{\frac{1}{2}} - f_{\sigma_v}(x)^{\frac{1}{2}} \right|^2 dx$$
$$= \int_0^1 x^{\theta_1 - 1} (1-x)^{\theta_2 - 1} e^{\sigma x} \left| \frac{1}{\sqrt{G_{\sigma_u}}} e^{\frac{ux}{2\sqrt{T}}} - \frac{1}{\sqrt{G_{\sigma_v}}} e^{\frac{vx}{2\sqrt{T}}} \right|^2 dx. \quad (3.10)$$

Now we have that

$$C_6 \min\{e^\sigma, 1\} \leq G_{\sigma_u} := \int_0^1 x^{\theta_1 - 1} (1-x)^{\theta_2 - 1} e^{\left(\sigma + \frac{u}{\sqrt{T}}\right)x} dx$$
$$\leq C_7 \max\{e^\sigma, 1\},$$

where $C_6 = B(\theta_1, \theta_2) e^{-\mathrm{diam}(\mathcal{S})}$, $C_7 = B(\theta_1, \theta_2) e^{\mathrm{diam}(\mathcal{S})}$ are non-zero, positive, and independent of σ and T, since we constrain $u, v \in \mathcal{U}_{T,\sigma}$ and we take $\mathrm{diam}(\mathcal{S})$ to mean $\sup_{w,w' \in \mathcal{S}} |w - w'|$. This allows us to deduce that $G \mapsto 1/\sqrt{G}$ is Lipschitz on the interval $[C_6 \inf_{\sigma \in \mathcal{K}} \min\{e^\sigma, 1\}, C_7 \sup_{\sigma \in \mathcal{K}} \max\{e^\sigma, 1\}]$ with some constant $C_8 > 0$, i.e.

$$\left| \frac{1}{\sqrt{G_{\sigma_u}}} - \frac{1}{\sqrt{G_{\sigma_v}}} \right| \leq C_8 \left| G_{\sigma_u} - G_{\sigma_v} \right|$$
$$= C_8 \int_0^1 x^{\theta_1 - 1} (1-x)^{\theta_2 - 1} e^{\sigma x} \left| e^{\frac{ux}{2\sqrt{T}}} - e^{\frac{vx}{2\sqrt{T}}} \right| dx$$
$$\leq C_8 C_9 \int_0^1 x^{\theta_1 - 1} (1-x)^{\theta_2 - 1} e^{\sigma x} \left| \frac{ux}{2\sqrt{T}} - \frac{vx}{2\sqrt{T}} \right| dx$$
$$= \frac{C_8 C_9}{2\sqrt{T}} |u - v| \int_0^1 x^{\theta_1} (1-x)^{\theta_2 - 1} e^{\sigma x} dx$$
$$\leq \frac{C_{10}}{\sqrt{T}} \max\{e^\sigma, 1\} |u - v|,$$

where in the second inequality we have made use of the fact that e^z is Lipschitz in z on $[-\mathrm{diam}(\mathcal{S}), \mathrm{diam}(\mathcal{S})]$ with some constant $C_9 > 0$.

Thus we deduce that

$$\begin{aligned}
&\Big|\frac{1}{\sqrt{G_{\sigma_u}}}e^{\frac{ux}{2\sqrt{T}}}-\frac{1}{\sqrt{G_{\sigma_v}}}e^{\frac{vx}{2\sqrt{T}}}\Big|^2\\
&=\left|\frac{1}{\sqrt{G_{\sigma_u}}}\left(e^{\frac{ux}{2\sqrt{T}}}-e^{\frac{vx}{2\sqrt{T}}}\right)\right|^2+\left|e^{\frac{vx}{2\sqrt{T}}}\left(\frac{1}{\sqrt{G_{\sigma_u}}}-\frac{1}{\sqrt{G_{\sigma_v}}}\right)\right|^2\\
&\quad+2\frac{e^{\frac{vx}{2\sqrt{T}}}}{\sqrt{G_{\sigma_u}}}\Big|e^{\frac{ux}{2\sqrt{T}}}-e^{\frac{vx}{2\sqrt{T}}}\Big|\Big|\frac{1}{\sqrt{G_{\sigma_u}}}-\frac{1}{\sqrt{G_{\sigma_v}}}\Big|\\
&\le\frac{C_9^2x^2}{4T}\frac{1}{C_6\min\{e^{\sigma},1\}}\big|u-v\big|^2\\
&\quad+e^{\mathrm{diam}(\mathcal{S})x}\frac{C_{10}^2}{T}\max\{e^{2\sigma},1\}\big|u-v\big|^2\\
&\quad+\frac{e^{\mathrm{diam}(\mathcal{S})x}C_9C_{10}x}{T\sqrt{C_6}}\frac{\max\{e^{\sigma},1\}}{\min\{e^{\sigma/2},1\}}\big|u-v\big|^2. \qquad (3.11)
\end{aligned}$$

Putting (3.11) into (3.10) gives us the result

$$\begin{aligned}
&\int_0^1 x^{\theta_1-1}(1-x)^{\theta_2-1}e^{\sigma x}\Big|\frac{1}{\sqrt{G_{\sigma_u}}}e^{\frac{ux}{\sqrt{T}}}-\frac{1}{\sqrt{G_{\sigma_v}}}e^{\frac{vx}{\sqrt{T}}}\Big|^2dx\\
&\qquad\le\frac{C_\sigma}{T}|u-v|^2,
\end{aligned}$$

as

$$C_\sigma:=C_{11}e^{|\sigma|}+C_{12}\max\{e^{3\sigma},1\}+C_{13}\frac{\max\{e^{2\sigma},1\}}{\min\{e^{\sigma/2},1\}},$$

is continuous in σ over any compact set $\mathcal{K}\subset\mathcal{S}$.

□

Proposition 3.6. *Let the initial distribution be such that for any $M\ge 2$ and for $u\in\mathcal{U}_{T,\sigma}$,*

$$\mathbb{P}_\nu^{(\sigma)}\left[\left|\log\left(\frac{\nu(\sigma+\frac{u}{\sqrt{T}},X_0)}{\nu(\sigma,X_0)}\right)\right|>\mathbb{E}^{(\sigma)}\left[\xi(1-\xi)\right]|u|^2\right]\le\frac{C}{|u|^M} \qquad (3.12)$$

for some constant $C>0$. Then for $\mathcal{K}\subset\mathcal{S}$ compact, there exists a function $g_T(\cdot)\in\mathscr{G}$ such that for any $u\in\mathcal{U}_{T,\sigma}$ as defined in (3.4) we

have that

$$\sup_{\sigma\in\mathcal{K}} \mathbb{E}_{\nu}^{(\sigma)}\left[Z_{T,\sigma}(u)^{\frac{1}{2}}\right] \leq e^{-g_T(|u|)}. \tag{3.13}$$

The result holds in particular when $\nu = f_\sigma$.

Proof. Assume for now that for any $M \geq 2$ we have that

$$\mathbb{P}_{\nu}^{(\sigma)}\left[Z_{T,\sigma}(u) > \exp\left\{-\frac{1}{16}\mathbb{E}^{(\sigma)}\left[\xi(1-\xi)\right]|u|^2\right\}\right] \leq \frac{C_{\sigma,M}}{|u|^M} \tag{3.14}$$

for some constant $C_{\sigma,M} > 0$ depending on σ and M. We show that if (3.14) holds, then (3.13) follows. Indeed

$$\begin{aligned}
&\mathbb{E}_{\nu}^{(\sigma)}\left[Z_{T,\sigma}(u)^{\frac{1}{2}}\right] \\
&\quad= \mathbb{E}_{\nu}^{(\sigma)}\left[Z_{T,\sigma}(u)^{\frac{1}{2}} 1_{\left\{Z_{T,\sigma}(u)\leq \exp\left\{-\frac{1}{16}\mathbb{E}^{(\sigma)}[\xi(1-\xi)]|u|^2\right\}\right\}}\right] \\
&\qquad+ \mathbb{E}_{\nu}^{(\sigma)}\left[Z_{T,\sigma}(u)^{\frac{1}{2}} 1_{\left\{Z_{T,\sigma}(u)> \exp\left\{-\frac{1}{16}\mathbb{E}^{(\sigma)}[\xi(1-\xi)]|u|^2\right\}\right\}}\right] \\
&\quad\leq \exp\left\{-\frac{1}{32}\mathbb{E}^{(\sigma)}\left[\xi(1-\xi)\right]|u|^2\right\} \\
&\qquad+ \mathbb{E}_{\nu}^{(\sigma)}\left[Z_{T,\sigma}(u)\right]^{\frac{1}{2}} \\
&\qquad\qquad\times \mathbb{P}_{\nu}^{(\sigma)}\left[Z_{T,\sigma}(u) > \exp\left\{-\frac{1}{16}\mathbb{E}^{(\sigma)}\left[\xi(1-\xi)\right]|u|^2\right\}\right]^{\frac{1}{2}} \\
&\quad\leq \exp\left\{-\frac{1}{32}\mathbb{E}^{(\sigma)}\left[\xi(1-\xi)\right]|u|^2\right\} + \frac{C_{\sigma,M}}{|u|^{\frac{M}{2}}}
\end{aligned}$$

where in the first inequality we have made use of Cauchy-Schwarz, and for the second inequality we have used (3.14). Therefore,

$$\begin{aligned}
&\sup_{\sigma\in\mathcal{K}}\mathbb{E}_{\nu}^{(\sigma)}\left[Z_{T,\sigma}(u)^{\frac{1}{2}}\right] \\
&\quad\leq \sup_{\sigma\in\mathcal{K}}\left\{\exp\left\{-\frac{1}{32}\mathbb{E}^{(\sigma)}\left[\xi(1-\xi)\right]|u|^2\right\} + \frac{C_{\sigma,M}}{|u|^{\frac{M}{2}}}\right\} \\
&\quad= \exp\left\{-\frac{1}{32}\inf_{\sigma\in\mathcal{K}}\mathbb{E}^{(\sigma)}\left[\xi(1-\xi)\right]|u|^2\right\} + \frac{\sup_{\sigma\in\mathcal{K}} C_{\sigma,M}}{|u|^{\frac{M}{2}}} \\
&\quad=: \exp\left\{-g_T(|u|)\right\}.
\end{aligned}$$

It remains to ensure that $g_T(\cdot) \in \mathscr{G}$, that $\inf_{\sigma\in\mathcal{K}} \mathbb{E}^{(\sigma)}[\xi(1-\xi)] \geq \kappa > 0$ for some constant κ, and that for any $M \geq 2$ it holds that $\sup_{\sigma\in\mathcal{K}} C_{\sigma,M} < \infty$. Observe that

$$\min\left\{\inf_{\sigma\in\mathcal{K}} e^{\sigma}, 1\right\} B(\theta_1,\theta_2) \leq G_\sigma \leq \max\left\{\sup_{\sigma\in\mathcal{K}} e^{\sigma}, 1\right\} B(\theta_1,\theta_2).$$

Thus

$$\begin{aligned}
\inf_{\sigma\in\mathcal{K}} \mathbb{E}^{(\sigma)}\left[\xi(1-\xi)\right] &= \inf_{\sigma\in\mathcal{K}} \left\{\int_0^1 \frac{1}{G_\sigma} e^{\sigma\xi}\xi^{\theta_1}(1-\xi)^{\theta_2} d\xi\right\} \\
&\geq \frac{\inf_{\sigma\in\mathcal{K}}\left\{\int_0^1 e^{\sigma\xi}\xi^{\theta_1}(1-\xi)^{\theta_2} d\xi\right\}}{\max\left\{\sup_{\sigma\in\mathcal{K}} e^{\sigma}, 1\right\} B(\theta_1,\theta_2)} \\
&\geq \frac{\min\left\{\inf_{\sigma\in\mathcal{K}} e^{\sigma}, 1\right\} B(\theta_1+1,\theta_2+1)}{\max\left\{\sup_{\sigma\in\mathcal{K}} e^{\sigma}, 1\right\} B(\theta_1,\theta_2)} =: \kappa
\end{aligned}$$

and $\kappa > 0$ because $\mathcal{K}$ is bounded, and thus both $\sup_{\sigma\in\mathcal{K}} e^{\sigma}$ and $\inf_{\sigma\in\mathcal{K}} e^{\sigma}$ are finite and non-zero. We show that $\sup_{\sigma\in\mathcal{K}} C_{\sigma,M}$ is finite $\forall M \geq 2$ in what follows. We now check that $g_T(|u|)$ as defined above is in the class of functions $\mathscr{G}$. To this end, observe that

$$g_T(|u|) = -\log\left(\exp\left\{-\frac{1}{32}\inf_{\sigma\in\mathcal{K}} \mathbb{E}^{(\sigma)}\left[\xi(1-\xi)\right]|u|^2\right\} + \frac{\sup_{\sigma\in\mathcal{K}} C_{\sigma,M}}{|u|^{\frac{M}{2}}}\right).$$

Indeed, for a fixed $T > 0$, $g_T(|u|) \to \infty$ as $|u| \to \infty$, because we have that $\inf_{\sigma\in\mathcal{K}} \mathbb{E}^{(\sigma)}[\xi(1-\xi)] > 0$, and furthermore given any fixed N, we can choose M large enough (note the way we phrased (3.14) allows us to choose our M arbitrarily large, say $M > 2N$) such that

$$\begin{aligned}
&\lim_{\substack{T\to\infty \\ y\to\infty}} y^N e^{-g_T(y)} \\
&= \lim_{\substack{T\to\infty \\ y\to\infty}} y^N \left(\exp\left\{-\frac{1}{32}\inf_{\sigma\in\mathcal{K}} \mathbb{E}^{(\sigma)}\left[\xi(1-\xi)\right]|y|^2\right\} + \frac{\sup_{\sigma\in\mathcal{K}} C_{\sigma,M}}{|y|^{\frac{M}{2}}}\right) = 0,
\end{aligned}$$

where the order in which limits are taken is immaterial since our choice of $g_T(|u|)$ is independent of T. Thus we have proved that if (3.14) holds,

then

$$\sup_{\sigma\in\mathcal{K}} \mathbb{E}_\nu^{(\sigma)}\left[Z_{T,\sigma}(u)^{\frac{1}{2}}\right] \leq e^{-g_T(|u|)}, \qquad g_T(\cdot)\in\mathscr{G}.$$

To show that (3.14) holds, we make use of Markov's inequality as well as Theorem 3.2 in [LLL11]. Indeed, observe that

$$\begin{aligned}
&\mathbb{P}_\nu^{(\sigma)}\left[Z_{T,\sigma}(u) \geq \exp\left\{-\frac{1}{16}\mathbb{E}^{(\sigma)}\left[\xi(1-\xi)\right]|u|^2\right\}\right] \\
&\quad= \mathbb{P}_\nu^{(\sigma)}\Bigg[\frac{\nu(\sigma+\frac{u}{\sqrt{T}}, X_0)}{\nu\left(\sigma, X_0\right)} \\
&\qquad\qquad \times \exp\left\{\frac{u}{2\sqrt{T}}\int_0^T \sqrt{X_t(1-X_t)}dW_t \right. \\
&\qquad\qquad\qquad \left. - \frac{|u|^2}{8}\left(\frac{1}{T}\int_0^T X_t(1-X_t)dt - \mathbb{E}^{(\sigma)}\left[\xi(1-\xi)\right]\right)\right\} \\
&\qquad\qquad > \exp\left\{\frac{1}{16}\mathbb{E}^{(\sigma)}[\xi(1-\xi)]|u|^2\right\}\Bigg] \\
&\quad\leq \mathbb{P}_\nu^{(\sigma)}\left[\left|\log\left(\frac{\nu(\sigma+\frac{u}{\sqrt{T}}, X_0)}{\nu\left(\sigma, X_0\right)}\right)\right| > \frac{1}{48}\mathbb{E}^{(\sigma)}[\xi(1-\xi)]|u|^2\right] \\
&\qquad + \mathbb{P}_\nu^{(\sigma)}\left[\left|\frac{u}{2\sqrt{T}}\int_0^T \sqrt{X_t(1-X_t)}dW_t\right| > \frac{1}{48}\mathbb{E}^{(\sigma)}[\xi(1-\xi)]|u|^2\right] \\
&\qquad + \mathbb{P}_\nu^{(\sigma)}\Bigg[\frac{|u|^2}{8}\left|\frac{1}{T}\int_0^T X_t(1-X_t)dt - \mathbb{E}^{(\sigma)}\left[\xi(1-\xi)\right]\right| \\
&\qquad\qquad\qquad > \frac{1}{48}\mathbb{E}^{(\sigma)}[\xi(1-\xi)]|u|^2\Bigg] \\
&\quad=: A_1 + A_2 + A_3.
\end{aligned}$$

The bound for A_1 follows immediately from (3.12). For the particular

case when $\nu = f_\sigma$, we use Markov's inequality:

$$A_1 = \mathbb{P}_\nu^{(\sigma)}\left[\left|\log\left(\frac{G_\sigma}{G_{\sigma+\frac{u}{\sqrt{T}}}}\right) + \frac{u}{\sqrt{T}}X_0\right| > \frac{1}{48}\mathbb{E}^{(\sigma)}[\xi(1-\xi)]|u|^2\right]$$
$$\leq \left(\frac{48}{\mathbb{E}^{(\sigma)}[\xi(1-\xi)]|u|^2}\right)^M \mathbb{E}_\nu^{(\sigma)}\left[\left|\log\left(\frac{G_\sigma}{G_{\sigma+\frac{u}{\sqrt{T}}}}\right) + \frac{u}{\sqrt{T}}X_0\right|^M\right].$$

But

$$\log\left(\frac{G_\sigma}{G_{\sigma+\frac{u}{\sqrt{T}}}}\right) = \log\left(\frac{\int_0^1 x^{\theta_1-1}(1-x)^{\theta_2-1}e^{\sigma x}dx}{\int_0^1 x^{\theta_1-1}(1-x)^{\theta_2-1}e^{(\sigma+\frac{u}{\sqrt{T}})x}dx}\right) \leq \frac{|u|}{\sqrt{T}},$$

so we have

$$A_1 \leq \left(\frac{48}{\mathbb{E}^{(\sigma)}[\xi(1-\xi)]|u|^2}\right)^M \mathbb{E}_\nu^{(\sigma)}\left[\left|\frac{u}{\sqrt{T}}\right|^M |1+X_0|^M\right]$$
$$= \left(\frac{48}{\mathbb{E}^{(\sigma)}[\xi(1-\xi)]\sqrt{T}|u|}\right)^M \mathbb{E}^{(\sigma)}\left[|1+\xi|^M\right]$$
$$\leq \left(\frac{48 d_\sigma}{\mathbb{E}^{(\sigma)}[\xi(1-\xi)]|u|^2}\right)^M \mathbb{E}^{(\sigma)}\left[|1+\xi|^M\right] =: \frac{C_{\sigma,M}^{(1)}}{|u|^{2M}},$$

where in the second inequality we made use of the fact that $u \in \mathcal{U}_{T,\sigma}$, and thus $|u| \leq d_\sigma\sqrt{T}$ where we define $d_\sigma := \sup_{w\in\partial\mathcal{S}}|\sigma - w|$ (which is strictly positive and bounded as $\mathcal{S}$ is open and bounded). To see that $\sup_{\sigma\in\mathcal{K}} C_{\sigma,M}^{(1)}$ is bounded, observe that

$$\sup_{\sigma\in\mathcal{K}} C_{\sigma,M}^{(1)} = \sup_{\sigma\in\mathcal{K}}\left\{\left(\frac{48 d_\sigma}{\mathbb{E}^{(\sigma)}[\xi(1-\xi)]}\right)^M \mathbb{E}^{(\sigma)}\left[|1+\xi|^M\right]\right\}$$
$$\leq \left(96\frac{B(\theta_1,\theta_2)}{B(\theta_1+1,\theta_2+1)}\sup_{\sigma\in\mathcal{K}} d_\sigma \frac{\max\{e^\sigma,1\}}{\min\{e^\sigma,1\}}\right)^M,$$

which is clearly finite because $\mathcal{K}$ is bounded.

For A_2 we use a similar argument, but now use the fact that we

have a stochastic integral:

$$\begin{aligned}
A_2 &\leq \left(\frac{48}{\mathbb{E}^{(\sigma)}[\xi(1-\xi)]|u|^2}\right)^M \mathbb{E}_\nu^{(\sigma)}\left[\left|\frac{u}{2\sqrt{T}}\int_0^T \sqrt{X_t(1-X_t)}dW_t\right|^M\right] \\
&\leq \left(\frac{24}{\mathbb{E}^{(\sigma)}[\xi(1-\xi)]|u|}\right)^M \left(\frac{M}{2}(M-1)\right)^{\frac{M}{2}} \\
&\quad \times T^{-1}\mathbb{E}_\nu^{(\sigma)}\left[\int_0^T |X_t(1-X_t)|^{\frac{M}{2}}\,dt\right] \\
&\leq \left(\frac{12}{\mathbb{E}^{(\sigma)}[\xi(1-\xi)]|u|}\right)^M \left(\frac{M}{2}(M-1)\right)^{\frac{M}{2}} =: \frac{C^{(2)}_{\sigma,M}}{|u|^M},
\end{aligned}$$

where the first line uses Markov's inequality and the second inequality uses Lemma 1.1 (equation (1.3)) in [Kut04]. That $\sup_{\sigma\in\mathcal{K}} C^{(2)}_{\sigma,M}$ is finite follows from arguments similar to those used for the respective term in A_1.

For A_3 we make use of Theorem 3.2 in [LLL11], which gives us that for $M \geq 2$

$$\begin{aligned}
&\mathbb{P}_\nu^{(\sigma)}\left[\left|\frac{1}{T}\int_0^T X_t(1-X_t)dt - \mathbb{E}^{(\sigma)}\left[\xi(1-\xi)\right]\right| \geq \frac{1}{6}\mathbb{E}^{(\sigma)}\left[\xi(1-\xi)\right]\right] \\
&\qquad \leq K(\sigma,X,M)\frac{\|x(1-x)\|_\infty^M}{\left(\frac{\mathbb{E}^{(\sigma)}[\xi(1-\xi)]}{6}\sqrt{T}\right)^M}. \qquad (3.15)
\end{aligned}$$

For the RHS of (3.15), we have that

$$\begin{aligned}
K(\sigma,X,M)\frac{\|x(1-x)\|_\infty^M}{\left(\frac{\mathbb{E}^{(\sigma)}[\xi(1-\xi)]}{6}\sqrt{T}\right)^M} &\leq K(\sigma,X,M)\left(\frac{6\|x(1-x)\|_\infty d_\sigma}{\mathbb{E}^{(\sigma)}\left[\xi(1-\xi)\right]|u|}\right)^M \\
&=: \frac{C^{(3)}_{\sigma,M}}{|u|^M},
\end{aligned}$$

where $K(\sigma,X,M)$ is a function that depends on M and on the moments of the hitting times of X. Finally we deduce that $\sup_{\sigma\in\mathcal{K}} C^{(3)}_{\sigma,M}$ is finite

by observing that

$$\begin{aligned}\sup_{\sigma\in\mathcal{K}} C^{(3)}_{\sigma,M} &= \sup_{\sigma\in\mathcal{K}} K(\sigma,X,M)\left(\frac{6\|x(1-x)\|_\infty d_\sigma}{\mathbb{E}^{(\sigma)}\left[\xi(1-\xi)\right]}\right)^M \\ &\leq \sup_{\sigma\in\mathcal{K}}\Bigg\{K(\sigma,X,M) \\ &\qquad\times\left(\frac{3}{2}\frac{B(\theta_1,\theta_2)}{B(\theta_1+1,\theta_2+1)}\sup_{\sigma\in\mathcal{K}} d_\sigma \frac{\max\{e^\sigma,1\}}{\min\{e^\sigma,1\}}\right)^M\Bigg\},\end{aligned}$$

which is finite since $\|x(1-x)\|_\infty = 1/4$, $\mathcal{K}$ is compact, and $K(\sigma,X,M)$ is bounded in σ over $\mathcal{K}$ (see Theorem 2.2 and Corollary 2.5 for the corresponding details). □

Finally, we present the result which guarantees that Condition 3 in Theorem 3.3 holds, and thus that the Ibragimov–Has'minskii conditions hold for the Wright–Fisher diffusion.

Proposition 3.7. *The random functions $Z_\sigma(u)$ and*

$$\psi(v) := \int_{\mathbb{R}} \ell\left(v-u\right)\frac{Z_\sigma(u)}{\int_{\mathbb{R}} Z_\sigma(y)dy}du$$

attain their maximum and minimum respectively at the unique point $\bar{u} = \bar{u}(\sigma) = I(\sigma)^{-1}\Delta(\sigma)$ with probability 1.

Proof. The first assertion follows immediately from Corollary 3.4. For the second, a straightforward change of variable coupled with Lemma II.10.2 in [IH81] (which relies on Anderson's Lemma (Lemma II.10.1 in [IH81]) and guarantees the uniqueness of $\psi(v)$) gives the result. A more detailed proof can be found in Theorem III.2.1 in [IH81]. □

3.3 Numerical Simulations

We illustrate the results proved in Section 3 by showing consistency, convergence in distribution and convergence of moments for the MLE when applied to data simulated from the Wright–Fisher diffusion. By making use of the exact algorithm (see [JS17] for full details, or Section 4.2 for a brief review), we obtain exact draws from the Wright–Fisher

diffusion. The generated paths are then used to calculate the MLE, and subsequently kernel smoothed density estimates for the rescaled MLE for various terminal times T are plotted against the density of the limiting distribution. Using the definition in (3.1), the MLE for the selection parameter is given by

$$\hat{\sigma}_T = \frac{X_T - X_0 - \int_0^T \left(-\theta_2 X_t + \theta_1(1 - X_t)\right) dt}{\int_0^T X_t(1 - X_t) dt},$$

which is impossible to calculate exactly in view of the random infinite dimensional paths involved in the integral. Instead we approximate the MLE by using Riemann sums instead of Lebesgue integrals, which gives rise to the approximation of $\hat{\sigma}_T$ given by

$$\check{\sigma}_T = \frac{X_T - X_0 - \sum_{i=1}^N \left(-\theta_2 X_{t_i} + \theta_1(1 - X_{t_i})\right) \Delta_i}{\sum_{i=1}^N X_{t_i}(1 - X_{t_i}) \Delta_i} \tag{3.16}$$

where $\Delta_i := t_i - t_{i-1}$ for a time discretisation grid $\{t_i\}_{i=0}^N$ where $t_0 = 0$ and $t_N = T$, and $N \in \mathbb{N} \setminus \{0\}$. In particular, $\{X_{t_i}\}_{i=0}^N$ denotes the values of the Wright–Fisher path at the times $\{t_i\}_{i=0}^N$, which corresponds to the output generated by the exact algorithm.

To simulate the Wright–Fisher paths, we set the selection parameter $\sigma = 4$, the mutation parameters $\theta_1, \theta_2 = 2$, $\Delta_i = 0.001$, $X_0 = 0.25$ and varied the terminal time $T \in \{1, 2, 10, 50\}$. For each of the 10,000 simulated paths, we computed (3.16), and subsequently for each T we obtained kernel smoothed estimates of the density of $\sqrt{T}(\check{\sigma}_T - \sigma)$ which are plotted against the limiting $N(0, \frac{1}{4}\mathbb{E}^{(\sigma)}[\xi(1-\xi)]^{-1})$ density in Figure 3.3.1.

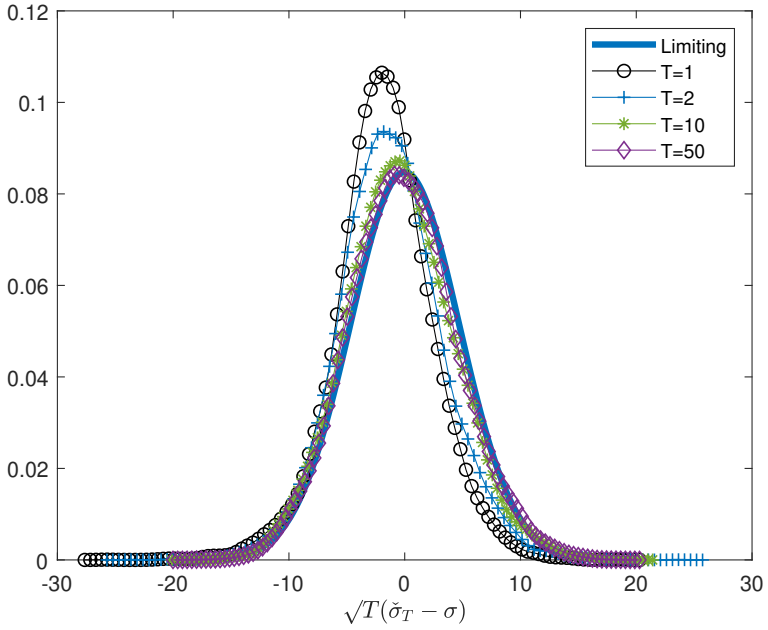


Figure 3.3.1: Plots of the kernel smoothed density estimates for $\sqrt{T}(\check{\sigma}_T - \sigma)$ for $T = 1, 2, 10, 50$, and of the limiting $N(0, I(\sigma)^{-1})$ density.

3.4 Extending the results to $\boldsymbol{\vartheta} = (\sigma, \theta_1, \theta_2)$

As pointed out in Remark 3.1 at the start of this chapter, assuming that the mutation parameters are a priori known is not too restrictive an assumption given the continuous observation regime. Nonetheless, in this section we briefly explore the technical difficulties related to extending the setup in this chapter to allow for the joint inference of the selection and the mutation parameters.

First and foremost we point out that we can only entertain this extension provided the mutation parameters are both larger than or equal to 1, for as seen in the discussion directly preceding and following (2.27) (and further pointed out in Remark 3.1), the Radon–Nikodym derivative can only be defined once we restrict our attention to the parameter space $\mathbb{R} \times [1, \infty)^2$. Given this restriction, we observe that the likelihood ratio $Z_{T,\boldsymbol{\vartheta}}$ would now be given by the first expression on the RHS of (2.32), which implies (as illustrated

in the proof of Theorem 2.8) that we need to be able to deal with the unbounded expressions appearing in the integrands there. It is these extra terms which create problems in establishing Condition 1 in the Ibragimov-Has'minskii conditions via Propositions 3.5 and 3.6.

Observe that Proposition 3.6 relies on (3.14) holding for any $M \geq 2$, which when considering solely selection, follows directly from the bounds derived in Theorem 2.2. However, the approach used in Theorem 2.3 relied on bounds involving $M = 2$ only. Extending the underlying calculations to allow for $M > 2$ requires access to the M^{th} moments of the random variable $\int_0^{T_b} h(Y_t)dt$. Whilst deriving and solving a recursive ODE for these quantities is possible, bounding them from above in terms of $\boldsymbol{\vartheta}$ over compact sets for arbitrary M is non-trivial, as recursive bounds on the respective solutions are no longer guaranteed in general. Moreover, observe that the proof of Proposition 3.5 would also require more delicate (and separate) arguments for the ML and Bayesian estimators, particularly in view of the fact that (3.5) would be required to hold for $m \geq q > 3$ for the MLE (as the parameter $\boldsymbol{\vartheta}$ would now be three-dimensional).

Chapter 4

MCMC Inference for the Wright–Fisher Diffusion

Inferring genetically relevant features such as selection, allele age, mutation and effective population size from population-wide data has been a perennial problem for geneticists. Most of the traditional methods used are based solely on present day genetic information which greatly impairs inference as it constitutes a static snapshot of the population being considered. Recent advances in gene sequencing as well as improvements in the technologies related to ancient DNA (aDNA) retrieval from old remains such as fossils, have allowed for the creation of genetic datasets spanning several centuries ([HSK$^+$05, LPR$^+$09, SCWGSL17, Mat20, F$^+$19]). Such time series data potentially holds a wealth of information with regards to how several genetic factors and phenomena have influenced and helped shape the population upon which they act, but eliciting such information from the observations requires the development of more intricate statistical procedures.

The previous chapter focused on theoretical guarantees associated with the ML and Bayesian estimators in an idealised scenario when one has access to the entire allele frequency trajectory. The results there allowed us to conclude that in such a setting one can conduct inference with the confidence that the estimators being used retain a set of desir-

able properties. In turn, this also sets a baseline from which one could hope to achieve similar guarantees in a discrete observation setting, as observed empirically from the simulations conducted in Subsection 3.3.

In this chapter we develop a practicable inferential framework with the aim of conducting exact inference (as given by Definition 4.1) on the Wright–Fisher diffusion from discrete noisy observations, inspired by similar work done in the context of Brownian motion ([BPRF06, SPR$^+$13]) and in view of the availability of an exact algorithm for the Wright–Fisher class of diffusions ([JS17]). The main challenge faced here is that the likelihood is now no longer explicitly given by the Girsanov theorem in an analytic form, but instead involves a finite product of intractable transition densities.

A standard approach when faced with intractable likelihoods but with access to exact simulation, is to resort to augmentation. By expanding the state space with "extra" variates, the intractability present in the original likelihood is subsumed into the simulation, resulting in a tractable expression for the augmented likelihood. Employing such an augmentation in turn leads to an augmented posterior (over the parameters of interest, as well as the newly introduced auxiliary variables), from which the original posterior can be easily recovered by marginalising over the auxiliary variables through the use of Monte Carlo techniques. In our case, we augment our state space with the values of the latent diffusion path at the observation times and skeleton points (which will be defined in Section 4.3), which in turn allow us to construct a Metropolis-within-Gibbs sampler targetting the joint posterior distribution of the selection parameter, the allele age, and the aforementioned auxiliary variables. We marginalise over the auxiliary variates by employing a combination of exact draws from the posterior and pseudo-marginal Metropolis–Hastings updates as illustrated in Subsections 4.3.1, 4.3.2 and 4.3.3. We specify that in the case when the allele age t_0 is known, the auxiliary variables can be simulated directly from the posterior, however the inclusion of allele age into the inferential framework makes the updating procedure slightly more

intricate and requires a pseudo-marginal Metropolis–Hastings step.

The remainder of this chapter is organised as follows: Section 4.1 gives a brief overview of the current state of the art in terms of inference for selection from time-series data in population genetics, whilst Section 4.2 is a summary of the exact algorithm developed for the Wright–Fisher diffusion (see [JS17] for more details) in both the neutral and non-neutral case. Section 4.3 then focuses on developing the mathematical framework within which we can embed the exact algorithm into a Markov chain Monte Carlo (MCMC) sampler, allowing us to target the posterior of interest. The method was subsequently applied to simulated data, and the resulting output together with details pertaining to the implementation and computational considerations are presented in Section 4.4. The chapter then concludes with Section 4.5 which highlights some extensions that are being looked into, namely developing a suitable proposal kernel to avoid the aforementioned pseudo-marginal update, extending the inference to account for demographic history and include the mutation parameters, and allowing for alleles to emerge from standing variation rather than requiring them to be the result of a de novo mutation.

4.1 Time series inference in population genetics

The lack of a tractable expression for the transition density of a non-neutral Wright–Fisher diffusion has lead to a number of methods in the literature which in some way or another rely on some form of discretisation. This allows for the intractable terms to be suitably approximated, and thus inference can be conducted on the parameters of interest. Although the discretisations and approximations employed vary from one method to another, a common problem faced by all is the fact that it is very hard (and indeed impossible) to quantify the error and bias they introduce into the estimates. In this section we provide a brief (and selective) overview of work that has been done on inferring selection from time-series data, however a more exhaustive and informative review on the subject can be found in [D$^+$20].

Amongst the first concerted efforts at addressing the problem of inferring selection (as well as effective population size) from noisy data driven by a Wright–Fisher diffusion, we mention the method developed in [BYN] which extends the framework first introduced in [WS99]. In the absence of recurrent mutation, the authors gain access to an approximation of the transition density by discretising space and time, and solving the Kolmogorov backward equation numerically via finite differences. Subsequently they employ numerical integration coupled with dynamic programming to approximate the likelihood function, and then conduct inference based on the resulting discretised likelihood surface obtained by repeating the above procedure for each point in the discretised parameter space. In an effort to make this setup less computationally onerous, the authors in [MMES12] approximate the Wright–Fisher diffusion by a discrete time, discrete state space Markov chain which is only allowed moves to adjacent states. This additional approximation avoids any need for numerical integration, as now the transition density is given by matrix exponentiation of the corresponding infinitesimal matrix generator. We point out that the authors here allow for the allele age to be inferred jointly with the selection coefficient and effective population size, however the resulting likelihood function does not allow for an analytic derivation of the MLE. Thus the authors resort to making use of the discretised likelihood surface (as in [BYN]) to obtain estimates of the parameters of interest.

The approach proposed in [MMES12] however still requires a significantly large number of latent states for the underlying process, once the population scaled selection parameter grows. To deal with this computationally infeasible scenario, the authors in [FALJW16], consider an alternative approximation for the Wright–Fisher diffusion by appealing to a discrete state, continuous time process constructed by defining the process's rate matrix in terms of the mean holding times for each state under the original diffusion's dynamics. Although numerical integration becomes necessary to compute the associated

transition densities, the authors are able to derive a forward recursion to compute the associated likelihood and further implement a MCMC scheme targetting the selection parameter, dominance parameter and effective population size. A similar approach was adopted in [MM13] where the underlying Wright–Fisher model dynamics are approximated over a discretised unit interval by integrating the density of a Gaussian (whose mean and variance match those of the true Wright–Fisher process) over subsequent increment mid-points. The authors then analytically derive the MLE for selection for a single population and for a structured population, and make use of the hidden Markov model framework coupled with an expectation maximisation algorithm to compute the MLE. The setup was further generalised to allow for time-varying selection in [Mat20]. Along a similar train of thought, [PSB19] consider (truncated) Gaussian and beta distributions (as well as versions of these with points masses at 0 and 1) as approximating mechanisms for the Wright–Fisher model, again setting the parameters of these parametric distributions by matching their moments to those of the true process. These are then directly used to construct a transition kernel which is subsequently employed within a hidden Markov model framework to obtain an approximation to the likelihood over the discretised parameter space.

Instead of approximating the diffusion process by a simpler one, the approach adopted in [SBS14] is to exploit a spectral decomposition of the non-neutral transition density by looking at the generator of the process directly. In particular, by adopting a suitable basis for the space of square integrable functions (with respect to the stationary density of the diffusion) and considering the eigenfunction expansion of the generator, the authors were able to express the transition density of the non-neutral diffusion in terms of an infinite sum involving computable terms (first developed in [SS12]). Furthermore, they also obtained a dynamic programming algorithm for computing the emission and transmission probabilities of the underlying hidden Markov model setup, allowing for inference to be conducted via a grid search over the likelihood surface generated. Nonetheless as the

transition density is expressed as an infinite series, all computations required a truncation threshold to allow for the quantities involved to be computable in finite time. We mention that this setup was generalised further to allow for varying population sizes in [ŽSSS15]. The last article we mention that explicitly relies on numerically solving the quantities at hand is [HDBY20b], where by making use of filtering recursions, and numerically solving the Kolmogorov backward equation, the authors were able to approximate the likelihood function. Inference for the selection coefficient, dominance parameter and allele age is once again performed via a grid search over the discretised likelihood surface.

A central theme prevalent in all of the above discretisation-based methods is the fact that all the procedures and calculations need to be re-run from scratch for each point in the discretised parameter space. In contrast, in [KPR21] an exact filtering algorithm for the neutral Wright–Fisher diffusion is developed by making use of its dual process, and subsequently inference is conducted via a MCMC scheme. Here the authors are actually interested in estimating the mutation parameters from noisy observations of the diffusion, and rely on the fact that the dual of the neutral Wright–Fisher diffusion is a pure death process, ensuring that all the relevant calculations translate into finite sums. The dual process for the non-neutral Wright–Fisher diffusion however turns out to be a birth and death process, so the associated calculations no longer involve finite sums and thus this approach cannot be extended for this case. Another article focusing on the neutral Wright–Fisher diffusion and inference for the mutation parameters is [GaR17] where the authors make use of Barker's algorithm, Poisson coins and Bernoulli factories to perform exact inference. The former is an alternative to the Metropolis-Hastings algorithm, making use Barker's acceptance probability where instead of taking the quotient of the product of the prior and proposal kernel densities, one considers the ratio between this product evaluated at the proposed value and the sum of the product evaluated at both the proposed and current value of the chain. Although the resulting algorithm produces

a Markov chain with the desired stationary distribution, Barker's acceptance probability leads to a higher asymptotic Monte Carlo variance in comparison to the standard Metropolis-Hastings choice (which minimises this quantity over the space of square integrable functions with respect to the stationary distribution), see Theorem 4 in [aR11] for a formal statement. For more details on the Barker algorithm see [Bar65] or [GaR17, Section 2]. The idea of using Barker's algorithm in [GaR17] is somewhat close to the setup we entertain, with the exception being that the authors make use of a layered Brownian bridge (a Brownian bridge which is simulated jointly with either its maximum or minimum value over the simulation time interval), thus allowing for a Barker's acceptance probability, which can be targetted via a combination of Poisson coins and Bernoulli factories. One downside of this particular approach is that the authors condition on the event that the diffusion avoids the boundary entirely, and secondly, from a computational point of view, the method is quite inefficient as it relies on simulating layered Brownian bridges. This simulation procedure is already known to be quite inefficient in the general case ([PR08]), and will be even more so in the case of a Wright–Fisher diffusion as we require the proposed paths to remain within the interval $[0, 1]$. In [HDBY20a], a sequential Monte Carlo method making use of a particle marginal Metropolis–Hastings algorithm is used to estimate the conditional density of the observations given a particular parameter configuration. One crucial assumption in this setup is that the underlying diffusion is conditioned on avoiding the boundary, thereby allowing the particle filter bootstrap employed to make use of an Euler–Maruyama scheme to simulate the latent path instead of the more computationally intense process of numerically solving the Kolmogorov backward equation.

We conclude this brief overview of time-series inference for population genetics by reviewing the setup and method used in [SES16], as it is perhaps one of the closest to the method we describe in Section 4.3. Here the authors implement a Metropolis-within-Gibbs sampler to sequentially update the selection coefficients (assuming a diploid

population), allele age, and underlying latent path. By augmenting the state space with the latter and choosing a suitable reference measure, the intractable likelihood now becomes amenable to a MCMC scheme. In contrast to the layered Brownian motion used in [GaR17], the authors here make use of the Bessel(0) process as a reference measure, which is particularly suitable for the Wright–Fisher diffusion as it displays a similar behaviour at the 0 boundary. However, the Bessel(0) process is unbounded from above unlike the Wright–Fisher diffusion, and thus this discrepancy ultimately manifests itself in the integrands in the exponent of the corresponding Radon–Nikodym derivative between the two processes being unbounded from below. As will be made clearer in Section 4.2, this means that an exact sampler cannot be entertained as the resulting acceptance probability cannot be associated with a corresponding Poisson point process. This, coupled with the fact that any Lebesgue integrals are necessarily evaluated via Riemann sums, means that the resulting inferential technique is no longer exact. In Section 4.3 we adopt a similar approach, where we make use of a Metropolis-within-Gibbs Sampler to sequentially update the selection coefficient, allele age and latent path, however we advocate for the neutral Wright–Fisher diffusion as being the right reference process to consider in order to obtain an exact inferential scheme. Whilst the resulting likelihood function still involves Lebesgue integrals which cannot be evaluated exactly, they can be associated with a corresponding Poisson point process and thus need not be computed.

4.2 Exact Algorithm for the Wright–Fisher diffusion

The Exact Algorithm, first introduced in [BR05] (commonly termed EA1), and generalised further in [BPR06] and [BPR08] (which are referred to as EA2 and EA3 respectively), allows for the exact simulation from the law of a diffusion by generating candidate paths from the law of a Brownian motion (or Brownian bridge if considering bridge diffusions), followed by a simple accept-reject step. The key

insight into this method is that although the associated acceptance probability is in general intractable due to the presence of Lebesgue integrals involving an infinite dimensional random path, it can be unbiasedly estimated via a Poisson point process provided the resulting terms in the exponent can be bounded from below. Simulation of the Poisson point process is straightforward and allows for a simple rejection routine to return values from the target distribution. The class of diffusions for which this method is applicable is quite large (the most general being the class described in [BPR08]), however an essential feature present in all three variations is that the target diffusion needs to be absolutely continuous with respect to Brownian motion. This is particularly problematic for diffusions which do not share the same underlying state space as Brownian motion (i.e. all of $\mathbb{R}$), for then the laws are not mutually absolutely continuous (unless one conditions on the diffusion avoiding the boundaries), and the Radon–Nikodym derivative is undefined. Although this has been somewhat mitigated by the layered Brownian motion developed in EA3, the approach is very inefficient (it is roughly 10 times slower than EA1 as reported in [PR08]), as any path that crosses a boundary is necessarily discarded. Moreover it is not always the case that the exponents in the Radon–Nikodym derivative can be bounded from below, and thus the link with the Poisson point process breaks.

Rather than pressing on with Brownian motion as a reference measure, perhaps a better candidate would be the measure associated to a process that displays similar behaviour to the target diffusion at the boundaries, retains mutual absolute continuity over the entire state space, and can be simulated exactly reasonably efficiently. Since we are interested in the scalar Wright–Fisher diffusion which has two boundaries at 0 and 1, one natural candidate (as pursued in [SES16]) is the Bessel(0) process which does indeed display similar behaviour as the Wright–Fisher diffusion at 0. However the resulting exponent in the Radon–Nikodym derivative is unbounded from below precluding the use of Poisson point processes to deal with the intractable integrals, and furthermore problems with absolute continuity creep in once the

path exceeds 1 (with the acceptance probability deteriorating once the path gets arbitrarily close to 1). Naturally one might condition on the paths avoiding the upper boundary, however a routine simulating Bessel bridges conditional on their maximum is (as yet) not available, and moreover the problem with bounding the terms in the exponent of the Radon–Nikodym derivative persists.

In [JS17], the authors proved that simulating exact draws from the law of a neutral Wright–Fisher diffusion can be achieved in finite time using the alternating series trick, and subsequently proposed this law as a suitable candidate measure with respect to which one can perform rejection sampling to target a non-neutral Wright–Fisher path. We give a brief description of how exact simulation of the neutral paths can be achieved, before moving on to explain how to relate the acceptance probability to a corresponding Poisson point process in order to target non-neutral paths.

4.2.1 Neutral Wright–Fisher simulation

Although the neutral Wright–Fisher diffusion has a transition density which cannot be evaluated exactly, it can be written in infinite series form as follows (see [Gri79, Tav84, EG93, GS09])

$$p_0^{\boldsymbol{\theta}}(t,x,y) = \sum_{m=0}^{\infty} q_m^{\boldsymbol{\theta}}(t) \sum_{l=0}^{m} \mathcal{B}_{m,x}(l)\mathcal{D}_{\theta_1+l,\theta_2+m-l}(y), \tag{4.1}$$

where

$$q_m^{\boldsymbol{\theta}}(t) = \sum_{k=m}^{\infty} (-1)^{k-m} \frac{|\boldsymbol{\theta}|+2k-1}{m!(k-m)!} \frac{\Gamma(m+|\boldsymbol{\theta}|+k-1)}{\Gamma(m+|\boldsymbol{\theta}|)} e^{\frac{-k(k+|\boldsymbol{\theta}|-1)t}{2}}, \tag{4.2}$$

$|\boldsymbol{\theta}| = \theta_1 + \theta_2$ for $\boldsymbol{\theta} \in \mathbb{R}_+^2$, $x, y \in [0,1]$, $\mathcal{B}_{m,x}$ is the probability mass function for a binomial random variable with parameters m, x, and $\mathcal{D}_{\theta_1+l,\theta_2+m-l}$ is the probability density function for a beta random variable with parameters $\theta_1 + l, \theta_2 + m - l$. We mention here that (4.2) defines a probability mass function on $\mathbb{N}$, and is a manifestation of the duality between the Wright–Fisher diffusion and the Kingman co-

alescent as it describes the number of lineages still alive at time t in a Kingman coalescent started from infinity at time 0. The decomposition into a mixture of beta and binomial random variables in (4.1) allows for a rather simple routine to return draws from the neutral Wright–Fisher diffusion via augmentation:

1. Draw M from (4.2)
2. Conditional on $M = m$, draw $L \sim \mathrm{Bin}(m, x)$
3. Conditional on $M = m, L = l$, draw $Y \sim \mathrm{Beta}(\theta_1 + l, \theta_2 + m - l)$

Steps 2 and 3 are straightforward, however step 1 requires some more work because the probability mass function given in (4.2) is an infinite series and thus cannot be evaluated pointwise. However, as shown in [JS17], one can make use of the alternating series trick (see Chapter 4 in [Dev86] for more details) to return exact draws from this distribution. We point out here that the above decomposition suffers from a numerical instability whenever $t \leq 0.06$, and thus some form of approximation is required to deal with such instances. In particular, [JS17] advocate using a discretised normal distribution to sample from (4.2) in view of Theorem 4 in [Gri84] (note that the statement there is missing a factor β^{-2}) which gives that as $t \to 0$, the number of lineages that survive up to time t is given by a normal distribution with parameters

$$\mu = \frac{2\eta}{t}, \qquad \sigma = \begin{cases} \frac{2\eta}{t}\left(\frac{\eta+\beta}{\beta}\right)^2\left(1 + \frac{\eta}{\eta+\beta} - 2\eta\right) & \text{if } \beta \neq 0 \\ \frac{2}{3t} & \text{if } \beta = 0 \end{cases} \tag{4.3}$$

where $\beta = \frac{t}{2}(|\boldsymbol{\theta}| - 1)$ and

$$\eta = \begin{cases} \frac{\beta}{e^{\beta}-1} & \text{if } \beta \neq 0 \\ 1 & \beta = 0. \end{cases}$$

Using the decomposition in (4.1) as well as the fact that the density of a point $y \in [0, 1]$ sampled at time s from a Wright–Fisher bridge going

from x at time 0 to z at time t (with $s < t$) is given by

$$p_0^{\boldsymbol{\theta},x,t,z}(y;s) = \frac{p_0^{\boldsymbol{\theta}}(s,x,y)p_0^{\boldsymbol{\theta}}(t-s,y,z)}{p_0^{\boldsymbol{\theta}}(t,x,z)}, \tag{4.4}$$

we get that

$$p_0^{\boldsymbol{\theta},x,t,z}(y;s) = \sum_{m=0}^{\infty}\sum_{k=0}^{\infty}\sum_{l=0}^{m}\sum_{j=0}^{k} p_{m,k,l,j}^{(\boldsymbol{\theta},x,t,z,s)}\mathcal{D}_{\theta_1+l+j,\theta_2+m-l+k-j}(y)$$

where

$$\begin{aligned} p_{m,k,l,j}^{(\boldsymbol{\theta},x,t,z,s)} &= \frac{q_m^{\boldsymbol{\theta}}(s)q_k^{\boldsymbol{\theta}}(t-s)}{p_0^{\boldsymbol{\theta}}(t,x,z)}\mathcal{B}_{m,x}(l)\mathcal{D}_{\theta_1+j,\theta_2+k-j}(z) \\ &\quad \times \binom{k}{j}\frac{B(\theta_1+l+j,\theta_2+m-l+k-j)}{B(\theta_1+j,\theta_2+k-j)} \end{aligned}$$

for $m,k,l,j \in \mathbb{N}$ with $l \in \{0,\dots,m\}, j \in \{0,\dots,k\}$, and $B(\cdot,\cdot)$ as defined in (2.24). The alternating series method can be applied here once again as the mixture weights $\{p_{m,k,l,j}^{(\boldsymbol{\theta},x,t,z,s)} : m,k,l,j \in \mathbb{N}\}$ define a probability mass function on $\mathbb{N}^4$, with the only extra complication being the presence of the neutral transition density in the denominator. However, by exploiting the following alternative decomposition of the transition density

$$p_0^{\boldsymbol{\theta}}(t,x,z) = \sum_{m=0}^{\infty} q_m^{\boldsymbol{\theta}}(t)\mathbb{E}\left[\mathcal{D}_{\theta_1+L_m,\theta_2+m-L_m}(z)\right] \tag{4.5}$$

where $L_m \sim \text{Bin}(m,x)$, [JS17, Proposition 3] shows that (4.5) can be targetted via the alternating series method, and thus exact draws can be obtained from the law of a bridge diffusion going from x to y in time t, at time s via the following procedure:

1. Simulate $(M,K,L,J) \sim \{p_{m,k,l,j}^{(\boldsymbol{\theta},x,t,z,s)} : (m,k,l,j) \in \mathbb{N}^4\}$
2. Conditional on $(M,K,L,J) = (m,k,l,j)$, simulate $Y \sim \text{Beta}(\theta_1 + l + j, \theta_2 + m - l + k - j)$.

4.2.2 Non-neutral Wright–Fisher simulation

The above alternating series trick cannot be directly extended to the non-neutral case as the quantities involved are no longer tractable. Instead, [JS17] propose the neutral Wright–Fisher diffusion and bridge as candidates in a rejection sampler to return draws from the corresponding non-neutral process. Let $\mathbb{WF}_{0,\boldsymbol{\theta}}^{(x)}$ denote the law of a neutral Wright–Fisher diffusion started at x with mutation parameters set to $\boldsymbol{\theta}$, and let $\mathbb{WF}_{\sigma,\boldsymbol{\theta}}^{(x)}$ be the non-neutral counterpart with selection coefficient equal to σ. Mutual absolute continuity between these two laws (provided they share the same start point $X_0 = x$ and mutation parameters $\boldsymbol{\theta} = (\theta_1, \theta_2) \in \mathbb{R}_+^2$) is guaranteed (see the paragraph just below (2.27) for references regarding this claim) and in particular the Radon–Nikodym derivative at time $t > 0$ is given by

$$\frac{d\mathbb{WF}_{\sigma,\boldsymbol{\theta}}^{(x)}}{d\mathbb{WF}_{0,\boldsymbol{\theta}}^{(x)}}(X^t) = \exp\left\{\frac{\sigma}{2}\left(X_t - X_0\right) - \int_0^t \varphi_\sigma(X_s)ds\right\} \tag{4.6}$$

where $X^t := (X_s)_0^t$,

$$\varphi_\sigma(x) = \frac{\sigma}{4}\left(-\frac{\sigma}{2}x^2 + \left(\frac{\sigma}{2} - |\boldsymbol{\theta}|\right)x + \theta_1\right), \qquad x \in [0,1],$$

and the term $\frac{\sigma}{2}(X_t - X_0)$ corresponds to the quantity $\tilde{A}(X_t)$ in equation (24) in [JS17]. In view of the fact that $\varphi_\sigma(x)$ is quadratic in x, we can always find $\varphi_\sigma^-, \varphi_\sigma^+$ such that $\varphi_\sigma^- \leq \varphi_\sigma(x) \leq \varphi_\sigma^+$ for $x \in [0,1]$, and thus we can re-write (4.6) as

$$\frac{d\mathbb{WF}_{\sigma,\boldsymbol{\theta}}^{(x)}}{d\mathbb{WF}_{0,\boldsymbol{\theta}}^{(x)}}(X^t) \propto \exp\left\{\frac{\sigma}{2}\left(X_t - 1\right) - \int_0^t \left(\varphi_\sigma(X_s) - \varphi_\sigma^-\right) ds\right\}. \tag{4.7}$$

The fact that $\varphi_\sigma(\cdot)$ can be lower bounded implies that the right-most term above can be viewed as the probability of a unit rate Poisson point process having no points in the epigraph of $t \mapsto \varphi_\sigma(X_t) - \varphi_\sigma^-$. Simulating an event with this probability is straightforward; we simulate the associated Poisson point process and check that the resulting points satisfy the given condition.

To see this, suppose that we simulate the endpoint at time t from (4.1) using the procedure described in Subsection 4.2.1, and assume that the algorithm returned a value y which was accepted by the $e^{\frac{\sigma}{2}(y-1)}$ coin-flip (the first term on the RHS of (4.7)). Then conditional on y, we generate a Poisson point process (Ψ, Γ) of rate $\lambda_\sigma := \varphi_\sigma^+ - \varphi_\sigma^-$ on $(0,t) \times [0,1]$, so $\kappa \sim \text{Pois}(\lambda_\sigma t)$, $\Psi = \{\psi_i\}_{i=1}^\kappa \sim_{iid} \text{Unif}((0,t))$ and $\Gamma = \{\gamma_i\}_{i=1}^\kappa \sim_{iid} \text{Unif}([0,1])$. We then associate a corresponding path $\omega \sim \mathbb{WF}_{0,\boldsymbol{\theta}}^{(t,x,y)}$, by simulating the value of the path at the time stamps given by Ψ, i.e. we draw ω_{ψ_i} for $i = 1, \ldots, \kappa$ from the finite dimensional distributions of $\mathbb{WF}_{0,\boldsymbol{\theta}}^{(t,x,y)}$ at the collections of times Ψ. If $\kappa = 0$, we accept the draw y as coming from the target distribution, otherwise we check whether the generated points are suitable by setting

$$I = \prod_{i=1}^{\kappa} 1_{\left\{ \frac{\varphi_\sigma(\omega_{\psi_i}) - \varphi_\sigma^-}{\varphi_\sigma^+ - \varphi_\sigma^-} < \gamma_i \right\}}.$$

If $I = 1$, we accept y as a draw from the corresponding non-neutral process, whilst if not we discard it and re-draw y from the neutral diffusion, as well as a corresponding Poisson point process (Ψ, Γ) until $I = 1$. The main draw back of this method is that it relies on a rejection sampler, and thus acceptance rates depend greatly on the mismatch between the proposed and target distributions. In the case of a non-neutral Wright–Fisher diffusion, it comes as no surprise that the acceptance rate plummets as the selection coefficient grows. Non-neutral paths with high selection coefficients (both negative and positive) will spend less time in the interior of $[0,1]$ and more time at the boundaries, when compared to their neutral counterparts. We summarise the procedure described above with the following algorithm

Algorithm 1: Generating a path from $\mathbb{WF}^{(x)}_{\sigma,\boldsymbol{\theta}}$ over the time interval $[0,t]$ (Algorithm 7 in [JS17])

repeat
  repeat
    Draw $X_t \sim \mathbb{WF}^{(x,t)}_{0,\boldsymbol{\theta}}$ and $U \sim \text{Unif}([0,1])$;
  until $U \leq \exp\{\frac{\sigma}{2}(X_t - 1)\}$;
  Conditional on $X_t = y$, simulate a Poisson point process (Ψ, Γ) of rate λ_σ on $(0,t) \times [0,1]$, with $\kappa \sim \text{Pois}(\lambda_\sigma t)$;
  for $i = 1, \ldots, \kappa$ **do**
    Simulate $\omega_{\psi_i} \sim \mathbb{WF}^{(t,x,y,\psi_i)}_{0,\boldsymbol{\theta}}$
  end
  if $\frac{\varphi_\sigma(\omega_{\psi_i}) - \varphi^-_\sigma}{\varphi^+_\sigma - \varphi^-_\sigma} \leq \gamma_i$ *for all* $i = 1, \ldots, \kappa$ **then**
    Set $\psi_0 = 0, \psi_{\kappa+1} = t$, $\omega_{\psi_0} = x, \omega_{\psi_{\kappa+1}} = X_t$;
    return $(\psi_i, \omega_{\psi_i})_{i=0}^{\kappa+1}$
  end
until *false*;

The above rejection sampler can be suitably tweaked to allow for generating samples from a non-neutral diffusion bridge. Let $\mathbb{WF}^{(t,x,y)}_{0,\boldsymbol{\theta}}$ denote the law of a neutral Wright–Fisher bridge going from x at time 0 to y at time t with mutation parameter set to $\boldsymbol{\theta}$, and let $\mathbb{WF}^{(t,x,y)}_{\sigma,\boldsymbol{\theta}}$ be the non-neutral counterpart when the selection parameter is set to σ. Then by conditioning on the endpoint $X_t = y$, re-arranging and applying Girsanov's theorem we get that the Radon–Nikodym derivative between these two laws is given by

$$\begin{aligned}\frac{d\mathbb{WF}^{(t,x,y)}_{\sigma,\boldsymbol{\theta}}}{d\mathbb{WF}^{(t,x,y)}_{0,\boldsymbol{\theta}}}(X^t) &= \frac{p^{\boldsymbol{\theta}}_0(t,x,y)}{p^{\boldsymbol{\theta}}_\sigma(t,x,y)} \frac{d\mathbb{WF}^{(x)}_{\sigma,\boldsymbol{\theta}}}{d\mathbb{WF}^{(x)}_{0,\boldsymbol{\theta}}}(X^t) \\ &\propto \exp\left\{ -\int_0^t \left(\varphi_\sigma(X_s) - \varphi^-_\sigma\right) ds \right\} \end{aligned} \tag{4.8}$$

and thus the same procedure as above can be executed to return a path according to the law of a non-neutral Wright–Fisher diffusion bridge, which we summarise into the following algorithm

Algorithm 2: Generating a path from $\mathbb{WF}_{\sigma,\boldsymbol{\theta}}^{(t,x,y)}$ over the time interval $[0,t]$ (Algorithm 8 in [JS17])

repeat
  Simulate a Poisson point process (Ψ, Γ) of rate λ_σ on $(0,t) \times [0,1]$, with $\kappa \sim \text{Pois}(\lambda_\sigma t)$;
  for $i = 1, \ldots, \kappa$ **do**
    Simulate $\omega_{\psi_i} \sim \mathbb{WF}_{0,\boldsymbol{\theta}}^{(t,x,y,\psi_i)}$
  end
  if $\frac{\varphi_\sigma(\omega_{\psi_i}) - \varphi_\sigma^-}{\varphi_\sigma^+ - \varphi_\sigma^-} \leq \gamma_i$ *for all* $i = 1, \ldots, \kappa$ **then**
    Set $\psi_0 = 0, \psi_{\kappa+1} = t$, $\omega_{\psi_0} = x, \omega_{\psi_{\kappa+1}} = y$;
    return $(\psi_i, \omega_{\psi_i})_{i=0}^{\kappa+1}$
  end
until *false*;

We now introduce the setup within which we want to embed the above exact sampling algorithms in order to devise a suitable MCMC scheme with which to conduct joint inference on the selection coefficient and the allele age based on allele frequency time-series data.

4.3 Exact MCMC Inference for Selection & Allele age

We start this section by defining precisely what we mean when we say "exact inference" by means of the following definition:

Definition 4.1. An inferential scheme shall be referred to as being *exact* if the only errors present in the method are Monte Carlo or machine precision errors.

Using this definition, it becomes apparent that none of the methods reviewed in Section 4.1 are exact as the discretisations and subsequent approximations used there result in the inferential procedure targetting some approximation to the true model, and thus the inference suffers from model error too. In this section we show how by considering a suitable state space augmentation and by resorting to the exact algorithms described in the previous section we can develop an exact

MCMC scheme targetting the joint posterior of the allele age t_0 and the selection coefficient σ.

We are interested in conducting inference on the selection parameter and allele age given that the underlying allele frequency dynamics are driven by a non-neutral Wright–Fisher diffusion, so the main problem is once again the intractability of the transition density which leads to an intractable likelihood. Assume that our observations are given by binomial draws $\mathbf{Y} = \{Y_{t_i}\}_{i=1}^n$ at observation times $\mathbf{t} = \{t_i\}_{i=1}^n$ with known sample sizes $\mathbf{n} := \{n_{t_i}\}_{i=1}^n$, and success probability driven by the underlying Wright–Fisher diffusion $(X_t)_{t_0}^{t_n}$ satisfying the same SDE as in (2.21), where t_0 is the birth of the allele such that $X_s \equiv 0 \; \forall s \leq t_0$, and

$$dX_t = \frac{1}{2}\Big(\sigma X_t(1-X_t) - \theta_2 X_t + \theta_1(1-X_t)\Big)dt + \sqrt{X_t(1-X_t)}dW_t,$$

holds $\forall t > t_0$. Such a setup is rather natural when considering aDNA datasets (see for instance [BYN, SBS14, ŽSSS15, GaR17, KPR21, SES16] amongst others); the underlying allele frequency dynamics are driven by a process which is well-approximated by the Wright–Fisher diffusion, and the binomial sampling encodes the noisy observations of the latent path coming from the samples collected from the archaeological remains. In such a setting, a straightforward expression for the likelihood of the data $\mathbf{Y}$ given the selection coefficient σ and the allele age t_0 is

$$\ell(\mathbf{Y}|\sigma, t_0) = \int_{[0,1]^n} \prod_{i=1}^n \mathcal{B}_{n_{t_i}, x_i}(Y_{t_i})\, p_\sigma^{\boldsymbol{\theta}}(t_i - t_{i-1}, x_{i-1}, x_i) d\mathbf{x}$$

where we are marginalising over the latent diffusion values at the observation times. As previously mentioned, the transition density $p_\sigma^{\boldsymbol{\theta}}(t_i - t_{i-1}, x_{i-1}, x_i)$ is intractable, and thus so is the posterior of interest. In view of the fact that we can simulate draws from the non-neutral Wright–Fisher diffusion and diffusion bridge, one could augment the state space with the values of the diffusion at the sampling

times $\mathbf{X} = \{X_{t_i}\}_{i=1}^n$ leading to the following augmented likelihood

$$\ell(\mathbf{Y}|\sigma, t_0, \mathbf{X}) = \prod_{i=1}^{n} \mathcal{B}_{n_{t_i}, X_{t_i}}(Y_{t_i})\, p_\sigma^{\boldsymbol{\theta}}(t_i - t_{i-1}, X_{t_{i-1}}, X_{t_i}),$$

and then apply a Metropolis-within-Gibbs scheme which sequentially updates the underlying path and the parameters of interest. The main difficulty in such a setup is the evaluation of the acceptance probabilities for the updates involving the selection coefficient and the allele age, as the Metropolis–Hastings acceptance probability would involve an intractable ratio. To see this, recall that the Radon–Nikodym derivative between the law of a neutral and non-neutral Wright–Fisher bridge sharing the same mutation parameters and endpoints is given by (4.8). Re-arranging and integrating both sides with respect to the bridge measure $\mathbb{WF}_{0,\boldsymbol{\theta}}^{(t,x,y)}$, we get that

$$\begin{aligned} p_\sigma^{\boldsymbol{\theta}}(t,x,y) &= p_0^{\boldsymbol{\theta}}(t,x,y) e^{\frac{\sigma}{2}(y-x) - t\varphi_\sigma^-} \mathbb{E}_{\mathbb{WF}_{0,\boldsymbol{\theta}}^{(t,x,y)}} \left[e^{-\int_0^t \left(\varphi_\sigma(X_s) - \varphi_\sigma^-\right) ds} \right] \\ &=: p_0^{\boldsymbol{\theta}}(t,x,y) e^{\frac{\sigma}{2}(y-x) - t\varphi_\sigma^-} a(t,x,y,\sigma). \end{aligned} \tag{4.9}$$

Note that whilst the first two quantities on the RHS can be dealt with, the term $a(t,x,y,\sigma)$ is intractable in view of the Lebesgue integral involved and the fact that it is an average over the space of continuous functions with respect to the measure $\mathbb{WF}_{0,\boldsymbol{\theta}}^{(t,x,y)}$. Thus if we were to augment our state space solely with the values of the latent diffusion at the observation times, the acceptance probabilities for the selection coefficient and allele age updates would involve ratios of these intractable terms. The above decomposition (4.9) also sheds light on why extending the alternating series method from Subsection 4.2.1 does not extend to the non-neutral case, as the $a(t,x,y,\sigma)$ terms cannot be targetted via (eventually) monotone upper and lower bounds.

We point out here that one could make use of a pseudo-marginal algorithm at this point, where instead of the quantities $a(t,x,y,\sigma)$, unbiased estimates could be used (which are readily available by making use of the Poisson estimator, which will be discussed in more

detail in Subsection 4.3.3). However, pseudo-marginal algorithms tend to lead to sticky behaviour in the chain updates (see [AR09]), and more importantly augmenting our state space further leads to a tractable expression for the likelihood. The augmentation we consider is inspired by the one used in [BPRF06] and [SPR$^+$13] where the authors respectively embed EA1 and EA3 into a MCMC scheme. The crux of this augmentation is the fact that the intractable terms described above turn out to be the normalising constants of the extra random variates introduced, allowing for a tractable expression of the likelihood to be derived. These extra variates are related to the procedure described in Subsection 4.2.2, in particular they are the time stamps Ψ and associated values of the path ω at these times (together with an extra variate which will be described more precisely later), collectively termed the "skeleton points", which one generates when deciding whether or not a generated proposal comes from the target non-neutral distribution. We give a detailed derivation of the likelihood contribution stemming from these terms, but emphasise here that the main difference between our setup and that in [BPRF06] and [SPR$^+$13] is that location invariance does not hold for the Wright–Fisher diffusion. Indeed this property is exploited quite heavily in [BPRF06] and [SPR$^+$13] to decouple the dependence of the dominating measure from the values of the diffusion at the sampling times. This is particularly important in their setting in view of the Lamperti transform they use which leads to the endpoints of this transformed diffusion being informative about the parameter being inferred. In view of the fact that we are making use of a pseudo-marginal Metropolis-within-Gibbs scheme, the choice of a dominating measure is a significantly more involved task for the Wright–Fisher case as we do not have access to location invariance, but at least we need not worry about the diffusion values at the observation times being informative for the parameters of interest. Furthermore, the above also implies that our setting requires a carefully chosen updating procedure which allows for the auxiliary quantities to be updated without running into problems relating to mutual absolute continuity.

We now derive the likelihood contribution these skeleton points generate and show how this leads to a tractable likelihood which can be targetted via a MCMC procedure. To this end, suppose, for illustrative purposes, that we only have one observation interval given by $[0, t]$, and that at the current iteration the latent diffusion takes the values $X_0 = x$ and $X_t = y$. Let (Ψ, Γ, Ξ) define a unit rate Poisson point process on $(0, t) \times [0, 1] \times (0, \infty)$, such that for any bounded $A \subset (0, \infty)$, the restriction of this process to $(0, t) \times [0, 1] \times A$ (which we denote $(\Psi, \Gamma, \Xi)|_A$) gives that $\Psi|_A := \{\psi_j : \xi_j \in A\} \sim_{iid} \text{Unif}((0, t))$, $\Gamma|_A = \{\gamma_j : \xi_j \in A\} \sim \text{Unif}([0, 1])$, $\Xi|_A = \{\xi_j : \xi_j \in A\} \sim_{iid} \text{Unif}(A)$. Further let $\omega \sim \mathbb{WF}_{0,\boldsymbol{\theta}}^{(t,x,y)}$ denote a neutral path from a Wright–Fisher bridge going from x to y in time t. The role of the extra variate Ξ here is to allow for a non-centred re-parametrisation of the problem, where the Poisson point process is decoupled from the parameter of interest (in this case σ). If we omit Ξ, the Poisson point process we generate would depend on σ through its rate $\lambda_\sigma (= \varphi_\sigma^+ - \varphi_\sigma^-)$, thus leading to a less efficient sampler. The main idea is that by including Ξ we can precompute any additional points needed for the proposed value of the selection coefficient by simulating a Poisson point process whose rate is given by the maximum between the rate under the current value of σ and the proposed one. So, if $\sigma^{(k)}$ is the current value of the selection coefficient and σ' is the proposed value of a selection coefficient update, then we set $\lambda_{\max} := \max\{\lambda_{\sigma^{(k)}}, \lambda_{\sigma'}\}$. The thinning property of the Poisson point process then implies that we can invoke the "correct" points only when required whilst retaining the structure of the Poisson point process. For more details refer to Section 4 in [SPR+13].

Given a realisation of (Ψ, Γ, Ξ) and ω, one checks that the generated points lie below the epigraph of the function $s \mapsto \frac{\varphi_\sigma(X_s) - \varphi_\sigma^-}{\varphi_\sigma^+ - \varphi_\sigma^-}$ as in Subsection 4.2.2, and if this is the case, one defines $\omega^\Psi := \{\omega_{\psi_j}\}$ (i.e. ω^Ψ corresponds to the values of the path ω at the timestamps given by Ψ), and stores the generated points by setting $\Phi = (\Psi, \Xi, \omega^\Psi)$ which will be henceforth termed the skeleton points. An accepted

configuration $(\Psi, \Gamma, \Xi, \omega^{\Psi})$ has density given by

$$\frac{\prod_{\{j:\xi_j \leq \lambda_\sigma\}} 1_{\left\{\frac{\varphi_\sigma(\omega_{\psi_j}) - \varphi_\sigma^-}{\varphi_\sigma^+ - \varphi_\sigma^-} < \gamma_j\right\}}}{a(t, x, y, \sigma)}$$

with respect to $\mathbb{PP}^{(t,1)} \otimes \mathbb{WF}_{0,\boldsymbol{\theta}}^{(t,x,y)}$, where $\mathbb{PP}^{(t,1)}$ denotes the law of a unit rate Poisson point process on $(0, t) \times [0, 1] \times (0, \infty)$. Marginalising over the uniform marks $\{\gamma_j\}$, gives that the skeleton points Φ have density

$$\frac{\prod_{\{j:\xi_j \leq \lambda_\sigma\}} \frac{\varphi_\sigma^+ - \varphi_\sigma(\omega_{\psi_j})}{\varphi_\sigma^+ - \varphi_\sigma^-}}{a(t, x, y, \sigma)} \tag{4.10}$$

with respect to $\mathbb{PP}^{(t)} \otimes \mathbb{WF}_{0,\boldsymbol{\theta}}^{(t,x,y)}$, where $\mathbb{PP}^{(t)}$ denotes the law of a unit rate Poisson point process on $(0, t) \times (0, \infty)$. It is clear now that the denominator in (4.10) will cancel out with the same term appearing in the numerator of (4.9).

The last thing to note is that we need to ensure that the dominating measure used above is independent of the quantities we will be updating in our Metropolis-within-Gibbs sampler. There are two potential problems with the dominating measure $\mathbb{PP}^{(t)} \otimes \mathbb{WF}_{0,\boldsymbol{\theta}}^{(t,x,y)}$ we derived for (4.10): the diffusion bridge endpoints and the time interval t. The latter will only be problematic when t_0 is involved as all other timestamps t_i for $i \in \{1, \ldots, n\}$ will be fixed, and can easily be tackled by considering a "bigger" Poisson point process and invoking Poisson thinning. For the time interval whose left endpoints is given by t_0, we consider a unit rate Poisson point process on $(0, \infty)^2$. We denote the law of this process by $\mathbb{PP}$, and use it as a dominating measure for the skeleton points over this interval such that the resulting density of the accepted skeleton points is analogous to the quantities derived in (4.10) with the only difference being an additional condition in the product's running index given by $\{\psi_j < t_1 - t_0\}$, which allows for the "correct" points to be invoked and retains the correct Poisson

structure in view of the thinning property of the Poisson point process.

To deal with the dependence of the reference diffusion bridge measure on the endpoints, we consider changing the dominating measure from that of a diffusion bridge to that of a diffusion. Indeed observe that by conditioning on the endpoint X_t, we have that

$$\frac{d\mathbb{WF}_{0,\boldsymbol{\theta}}^{(t,x,X_t)}}{d\mathbb{WF}_{0,\boldsymbol{\theta}}^{(x)}}\left((X_s)_0^t\right) = \frac{1}{p_0^{\boldsymbol{\theta}}(t,x,X_t)} \tag{4.11}$$

and thus by multiplying (4.10) by the resulting Radon–Nikodym derivative we get that the dominating Wright–Fisher measure for (4.10) is now $\mathbb{WF}_{0,\boldsymbol{\theta}}^{(x)}$. This still depends on the left hand bridge endpoint x, however as these measures will be sequentially chained to one another, we will ultimately end up with the dominating measure $\mathbb{WF}_{0,\boldsymbol{\theta}}^{(0)}$ as shown shortly. Note further that the joint density of X_t and the skeleton points Φ is now given by

$$e^{\frac{\sigma}{2}(X_t-x)-\varphi_\sigma^- t}\prod_{\{j:\xi_j\leq\lambda_\sigma\}}\frac{\varphi_\sigma^+-\varphi_\sigma(\omega_{\psi_j})}{\varphi_\sigma^+-\varphi_\sigma^-}$$

with respect to $\mathbb{PP}\otimes\mathbb{WF}_{0,\boldsymbol{\theta}}^{(x)}\otimes\mathrm{Leb}([0,1])$, where the intractable $a(t,x,y,\sigma)$ and non-neutral transition density $p_0^{\boldsymbol{\theta}}(t,x,y)$ cancel out when combining (4.9), (4.10) and (4.11).

Now denote by $\Phi_i = (\Psi_i, \Xi_i, \omega_i^{\Psi_i})$ the skeleton points over the time interval $[t_{i-1}, t_i]$ for $i = 1,\ldots,n$. Putting all of the above together, and assuming for the moment that $Y_{t_1} > 0$, we can formulate the joint density of the data $\mathbf{Y} = \{Y_{t_i}\}_{i=1}^n$, the selection coefficient σ, the allele age t_0, the value of the diffusion at the sampling times

$\mathbf{X} = \{X_{t_i}\}_{i=1}^n$, and the collection of skeleton points $\{\Phi_i\}_{i=1}^n$ as

$$p_1(\sigma)p_2(t_0)\prod_{i=1}^{n}\mathcal{B}_{n_{t_i},X_{t_i}}(Y_{t_i})\,e^{\frac{\sigma}{2}X_{t_n}-\varphi_\sigma^-(t_n-t_0)}$$
$$\times\prod_{i=1}^{n}\prod_{\substack{\{j:\xi_{i,j}\leq\lambda_\sigma,\\ \psi_{1,j}<t_1-t_0\}}}\frac{\varphi_\sigma^+-\varphi_\sigma(\omega_{i,\psi_{i,j}})}{\varphi_\sigma^+-\varphi_\sigma^-}$$

with respect to the dominating measure

$$\mathbb{PP}\otimes\mathbb{PP}^{(t_n-t_1)}\otimes\mathbb{WF}_{0,\boldsymbol{\theta}}^{(0)}\otimes\mathrm{Leb}\left(E_{t_1}^n\right)\otimes\Sigma(\otimes_{i=1}^n n_{t_i}),$$

where we set $E_{t_1}^n := [0,1]^n\times\mathbb{R}\times(-\infty,t_1)$, p_1, p_2 represent the prior densities on σ and t_0 respectively with respect to Lebesgue measure, $\mathbb{WF}_{0,\boldsymbol{\theta}}^{(0)}$ denotes the law of a neutral Wright–Fisher diffusion started from 0, $\mathbb{PP}^{(t_n-t_1)}$ is the law of a unit rate Poisson point process on $(t_1,t_n)\times(0,\infty)$, $\mathbb{PP}$ is the law of a unit rate Poisson point process on $(0,\infty)^2$, and $\Sigma(\otimes_{i=1}^n n_{t_i})$ is the counting measure over $\otimes_{i=1}^n\{0,\ldots,n_{t_i}\}$. Note that we applied the following change of measure to ensure that the dominating measure for the skeleton points does not depend on the values $\mathbf{X}$

$$\frac{d\left(\bigotimes_{i=1}^{n}\mathbb{WF}_{0,\boldsymbol{\theta}}^{(t_i-t_{i-1},X_{t_{i-1}},X_{t_i})}\right)}{d\mathbb{WF}_{0,\boldsymbol{\theta}}^{(0)}}\left((X_t)_{t_0}^{t_n}\right)$$
$$=\prod_{i=1}^{n}\frac{1}{p_0^{\boldsymbol{\theta}}(t_i-t_{i-1},X_{t_{i-1}},X_{t_i})}.$$

which follows by iterating (4.11) over the values of the diffusion at subsequent observation times and deriving the corresponding Radon–Nikodym derivative.

Although it is possible to update all the auxiliary variables in one go, the resulting acceptance probability would be quite low, and would require a pseudo-marginal Metropolis–Hastings step in view of the more problematic updates involving the allele age. Instead, we propose splitting this update into several subroutines

which consider individual path segments in such a way that the dominating measure remains independent of the simulated quantities and the updating procedure retains a reasonable acceptance rate. To this end, we calculate the likelihood contribution for each path segment separately below, starting with an interior path segment.

Take $i \in \{2, \ldots, n-1\}$, fix $X_{t_{i-1}} = x_{i-1}$ and $X_{t_{i+1}} = x_{i+1}$, and denote by Φ_i, Φ_{i+1} the collection of skeleton points over the time intervals $[t_{i-1}, t_i], [t_i, t_{i+1}]$ respectively, such that the joint density of the data Y_{t_i}, the selection coefficient σ, the value of the diffusion X_{t_i} and corresponding skeleton points Φ_i, Φ_{i+1} is given by

$$p_1(\sigma)\mathcal{B}_{n_{t_i}, X_{t_i}}(Y_{t_i})\, e^{\frac{\sigma}{2}(x_{i+1}-x_{i-1})-\varphi_\sigma^-(t_{i+1}-t_{i-1})} \times \prod_{k=i}^{i+1} \prod_{\{j:\xi_{k,j}\leq\lambda_\sigma\}} \frac{\varphi_\sigma^+ - \varphi_\sigma(\omega_{k,\psi_{k,j}})}{\varphi_\sigma^+ - \varphi_\sigma^-} \tag{4.12}$$

with respect to the dominating measure

$$\mathbb{PP}^{(t_{i+1}-t_{i-1})} \otimes \mathbb{WF}_{0,\boldsymbol{\theta}}^{(x_{i-1})} \otimes \mathrm{Leb}\,([0,1]\times\mathbb{R}) \otimes \Sigma(n_{t_i}),$$

where $\mathbb{WF}_{0,\boldsymbol{\theta}}^{(x_{i-1})}$ denotes the law of a neutral Wright–Fisher diffusion started from x_{i-1}, $\mathbb{PP}^{(t_{i+1}-t_{i-1})}$ is the law of a unit rate Poisson point process on $(t_{i-1}, t_{i+1}) \times (0, \infty)$, and $\Sigma(n_{t_i})$ is the counting measure on the integers $\{0, \ldots, n_{t_i}\}$. Note that the fact that we considered two neighbouring time intervals and conditioned on the left-most and right-most endpoints allows us to use the law of a diffusion as a reference measure. In this manner, the dominating measure for such a local update does not depend on any of the parameters of interest nor on the auxiliary variables within the time interval (t_{i-1}, t_{i+1}). Moreover, all of the terms on the RHS of (4.12) are computable in finite time, suggesting that we have derived the correct augmentation within which to frame our MCMC scheme.

We note that the above only applies for what we call interior path segments, that is intervals of the form $[t_{i-1}, t_{i+1}]$ with $i \in \{2, \ldots, n-1\}$.

A similar approach gives us that fixing $X_{t_{n-1}} = x_{n-1}$, the joint density of the data Y_{t_n}, the selection coefficient σ, the diffusion endpoint X_{t_n} and corresponding skeleton points Φ_n is given by

$$p_1(\sigma)\mathcal{B}_{n_{t_n},X_{t_n}}\left(Y_{t_n}\right) e^{\frac{\sigma}{2}(X_{t_n}-x_{n-1})-\varphi_\sigma^-(t_n-t_{n-1})} \times \prod_{\{j:\xi_{n,j}\leq\lambda_\sigma\}} \frac{\varphi_\sigma^+ - \varphi_\sigma(\omega_{n,\psi_{n,j}})}{\varphi_\sigma^+ - \varphi_\sigma^-} \tag{4.13}$$

with respect to the dominating measure

$$\mathbb{PP}^{(t_n-t_{n-1})} \otimes \mathbb{WF}_{0,\boldsymbol{\theta}}^{(x_{n-1})} \otimes \mathrm{Leb}\left([0,1]\times\mathbb{R}\right) \otimes \Sigma(n_{t_n}).$$

where $\mathbb{WF}_{0,\boldsymbol{\theta}}^{(x_{n-1})}$ denotes the law of a neutral Wright–Fisher diffusion started at x_{n-1}, $\mathbb{PP}^{(t_n-t_{n-1})}$ is the law of a unit rate Poisson point process on $(t_{n-1},t_n)\times(0,\infty)$, and $\Sigma(n_{t_n})$ is the counting measure on $\{0,\ldots,n_{t_n}\}$.

The last section of path we need to tackle is the initial one, where (as mentioned earlier) we need to be slightly more careful in the choice of the dominating measure for the Poisson point process. Fixing the right endpoint $X_{t_2} = x_2$, the joint density of the data Y_{t_1}, the selection coefficient σ, the allele age t_0, the diffusion value at the first observation time X_{t_1} and corresponding skeleton points Φ_1, Φ_2 is given by

$$p_1(\sigma)p_2(t_0)\mathcal{B}_{n_{t_1},X_{t_1}}\left(Y_{t_1}\right) e^{\frac{\sigma}{2}x_2-\varphi_\sigma^-(t_2-t_0)} \times \prod_{i=1}^{2} \prod_{\substack{\{j:\xi_{i,j}\leq\lambda_\sigma,\\ \psi_{1,j}<t_1-t_0\}}} \frac{\varphi_\sigma^+ - \varphi_\sigma(\omega_{i,\psi_{i,j}})}{\varphi_\sigma^+ - \varphi_\sigma^-} \tag{4.14}$$

with respect to the dominating measure

$$\mathbb{PP} \otimes \mathbb{PP}^{(t_2-t_1)} \otimes \mathbb{WF}_{0,\boldsymbol{\theta}}^{(0)} \otimes \mathrm{Leb}\left(E_{t_1}^1\right) \otimes \Sigma(n_{t_1}),$$

where we set $E_{t_1}^1 := [0,1]\times\mathbb{R}\times(-\infty,t_1)$, $\mathbb{WF}_{0,\boldsymbol{\theta}}^{(0)}$ denotes the law of a neutral Wright–Fisher diffusion started at 0, whilst $\mathbb{PP}$ is the law of a

unit rate Poisson point process on $(0,\infty)^2$ and $\Sigma(n_{t_1})$ is the counting measure on $\{0,\ldots,n_{t_1}\}$.

In light of the above we have found an augmentation which grants us access to an expression for the likelihood which we can evaluate pointwise, and that admits a dominating measure which is independent of the parameters and the auxiliary variables being used. We are now left with the task of devising an appropriate proposal mechanism which does not lead to any extra intractable terms cropping up in the acceptance probabilities or dominating measures. As mentioned earlier, in [BPRF06] and [SPR$^+$13] the authors exploit location invariance of Brownian bridges to ensure that the dominating measure is that of a Brownian bridge starting and ending at 0. This allows for the decoupling of the dependence between the bridge measure and its endpoints (which is necessary in their setting due to the Lamperti transform they employ), and for a straightforward proposal mechanism. In our case, we propose a piecewise path updating procedure which ensures that locally the updates can be executed without any of the aforementioned problems relating to mutual absolute continuity, retaining reasonable acceptance rates, and ensuring that the global dominating measure does not depend on any of the parameters or auxiliary variables.

The Metropolis-within-Gibbs Sampler we employ is essentially split into two main update steps: updating the selection parameter conditional on the allele age, latent values of the diffusion at the observation times, and skeleton points; and updating the allele age, latent values of the diffusion at the observation times and skeleton points conditional on the selection parameter. The former is a simple Metropolis–Hastings step (necessarily so because the resulting posterior is not amenable to any standard simulation techniques), whilst for the latter we split the global update into several local piecewise updates to improve mixing. In principle one could update the entire path in one go, however in view of the pseudo-marginal step required for the initial path segment, there would be long runs where the chain

would remain stuck at the same values. Instead, we split the path into an initial segment, several interior segments and an end segment, and subsequently:

1. Fix the diffusion value at t_2, propose an allele age t_0 together with associated latent path value at the first observation time X_{t_1} and skeleton points Φ_1, Φ_2, and run a pseudo-marginal step,

2. For $i = 2, \ldots, n-2$, fix the diffusion values at t_{i-1} and t_{i+1}, and update X_{t_i} and Φ_i, Φ_{i+1} by directly sampling from the posterior,

3. Fix the diffusion value at t_{n-1}, and update X_{t_n} and Φ_n by directly sampling from the posterior.

Figure 4.3.1 below illustrates one sequential run of the algorithm, where one starts from the first segment A and performs step 1 in the above list. The algorithm then moves on to the first (left-most) of the blocks labelled B, and updates each segment using step 2 above. Once all of the interior segments labelled B are updated (going from left to right), the algorithm updates the end segment C as detailed in step 3. Note that by using bridge proposals and imposing overlapping segments in subsequent update steps, we allow for the whole path to be updated in one iteration. We point out here that the last segment C is updated via a diffusion update rather than a bridge update to allow for the terminal diffusion value X_{t_n} to be updated. Note further that whilst the interior and end path segments can be updated without an accept-reject step, the initial segment requires a pseudo-marginal Metropolis–Hastings step due to the dependence on the allele age of the proposed path. The precise details for each step above can be found in Subsection 4.3.1 for the interior path segment, 4.3.2 for the end path segment and 4.3.3 for the slightly more intricate initial path segment. The selection parameter update is illustrated in Subsection 4.3.4.

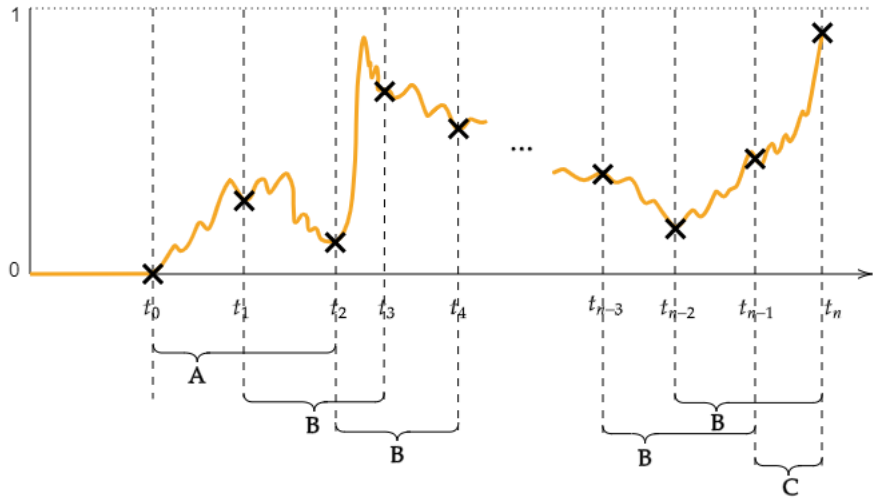


Figure 4.3.1: Diagram illustrating the update procedure. The latent path at the current iteration is in orange, crosses denote the values of the latent path at the observation times. The update procedure starts with the segment A which is the initial path segment, then proceeds to update the interior segments B sequentially from left to right and finishes with updating the end segment C.

Note that in the above we have assumed that the first observation $Y_{t_1} > 0$, which restricts the range of values t_0 can take to $(-\infty, t_1)$. However there is no a priori reason why this should be the case, and in general, if we set $c := \min\{i : Y_{t_i} > 0\}$, then t_0 should be able to assume any value in the range $(-\infty, t_c)$. This setting however requires a more careful proposal mechanism for the initial path segment, because if for instance $t_c = t_2$, we need to have two separate initial path segment proposal methods - one for when $t_0 < t_1$ and one for the case $t_0 \in (t_1, t_2)$. The sampler needs to be able to switch between and compare these two path segments when computing the acceptance probability, and thus we need to extend the initial path segment to include the observation interval $[t_0, t_3]$. If we have $t_0 \in (t_1, t_2)$, then we set $X_{t_1} = 0, \Phi_1 = \emptyset$ and use the same update procedure as above, using a bridge proposal over the interval $[t_0, t_3]$. If however $t_0 < t_1$, then we need to first simulate the value X_{t_1} using a diffusion proposal, and then for the remaining time interval $[t_1, t_3]$ we use a bridge update as above conditional on the output generated via the diffusion proposal. This

construction easily extends to accommodate an arbitrary t_c, however the larger the time interval over which the initial path segment is defined (it will necessarily be equal to $[t_0, t_{c+1}]$), the sharper the fall in the acceptance probabilities in virtue of the increase in time interval length which leads to larger numbers of Poisson points being favoured. In what follows we therefore allow for at most $t_c = t_3$ (which is in line with what we observe in real datasets such as the one in [LPR$^+$09]), however below we illustrate our method for the case $t_c = t_1$ for ease of exposition. The update procedures for when $t_c = t_2$ (or $t_c = t_3$) follow the same pattern, with the only difference being the inclusion of one (respectively two) "extra" diffusion proposals at the start, which we have therefore moved to Appendix A together with the relevant calculations of acceptance probabilities. Algorithm 3 summarises the above discussed update procedure, where each of the individual latent path updates are explained in more detail in the following subsections.

4.3.1 Updating an inner path segment

To update the path segment over the observation interval $[t_{i-1}, t_{i+1}]$, we first fix $X_{t_{i-1}} = x_{i-1}$ and $X_{t_{i+1}} = x_{i+1}$, and conditional on these values:

1. Draw $U \sim \text{Unif}([0,1])$, and

$$X_{t_i} \sim \mathbb{WF}_{\sigma,\boldsymbol{\theta}}^{(t_{i+1}-t_{i-1}, x_{i-1}, x_{i+1}, t_i - t_{i-1})}$$

   (i.e. draw X_{t_i} from the law of a non-neutral Wright–Fisher bridge going from x_{i-1} to x_{i+1} in time $t_{i+1} - t_{i-1}$, sampled at time $t_i - t_{i-1}$), and check whether

$$U < \frac{1}{M_i} {X_{t_i}}^{Y_{t_i}} (1 - X_{t_i})^{n_{t_i} - Y_{t_i}} \tag{4.15}$$

   where $M_i := \sup_{z \in [0,1]} z^{Y_{t_i}} (1-z)^{n_{t_i} - Y_{t_i}}$. If (4.15) holds, we proceed to 2, otherwise we keep drawing X_{t_i} and U as above until (4.15) is true.

2. Conditional on $X_{t_i} = x$, draw

$$\begin{aligned}
\kappa_i &\sim \text{Pois}(\lambda_{\max}(t_i - t_{i-1})),\\
\{\psi_{i,j}\}_{j=1}^{\kappa_i} &\sim_{iid} \text{Unif}((t_{i-1}, t_i)),\\
\{\xi_{i,j}\}_{j=1}^{\kappa_i} &\sim_{iid} \text{Unif}((0, \lambda_{\max})),\\
\{\gamma_{i,j}\}_{j=1}^{\kappa_i} &\sim_{iid} \text{Unif}([0,1]),\\
\omega_i &\sim \mathbb{WF}_{0,\boldsymbol{\theta}}^{(t_i - t_{i-1}, x_{i-1}, x)}
\end{aligned}$$

and

$$\begin{aligned}
\kappa_{i+1} &\sim \text{Pois}(\lambda_{\max}(t_{i+1} - t_i)),\\
\{\psi_{i+1,j}\}_{j=1}^{\kappa_{i+1}} &\sim_{iid} \text{Unif}((t_i, t_{i+1})),\\
\{\xi_{i+1,j}\}_{j=1}^{\kappa_{i+1}} &\sim_{iid} \text{Unif}((0, \lambda_{\max})),\\
\{\gamma_{i+1,j}\}_{j=1}^{\kappa_i} &\sim_{iid} \text{Unif}([0,1]),\\
\omega_{i+1} &\sim \mathbb{WF}_{0,\boldsymbol{\theta}}^{(t_{i+1} - t_i, x, x_{i+1})}
\end{aligned}$$

3. If

$$\prod_{k=i}^{i+1} \prod_{\{j:\xi_{k,j} \le \lambda_\sigma\}} 1_{\left\{\frac{\left(\varphi_\sigma(\omega_{k,\psi_{k,j}}) - \varphi_\sigma^-\right)}{\varphi_\sigma^+ - \varphi_\sigma^-} < \gamma_{k,j}\right\}} = 1$$

set $\Psi_k = \{\psi_{k,j}\}_{j=1}^{\kappa_k}$, $\Xi_k = \{\xi_{k,j}\}_{j=1}^{\kappa_k}$, $\omega_k^\Psi = \{\omega_{k,\psi_{k,j}}\}_{j=1}^{\kappa_k}$, $\Phi_k = (\Psi_k, \Xi_k, \omega_k^\Psi)$ for $k = i, i+1$, else go to 2.

A proposal $(X_{t_i}, \Phi_i, \Phi_{i+1})$ generated by the above mechanism has density given by

$$\begin{aligned}
&X_{t_i}^{Y_{t_i}}(1 - X_{t_i})^{n_{t_i} - Y_{t_i}} \frac{e^{\frac{\sigma}{2}(x_{i+1} - x_{i-1}) - \varphi_\sigma^-(t_{i+1} - t_{i-1})}}{p_\sigma^{\boldsymbol{\theta}}(t_{i+1} - t_{i-1}, x_{i-1}, x_{i+1})}\\
&\quad \times \prod_{k=i}^{i+1} \prod_{\{j:\xi_{k,j} \le \lambda_\sigma\}} \frac{\varphi_\sigma^+ - \varphi_\sigma(\omega_{k,\psi_{k,j}})}{\varphi_\sigma^+ - \varphi_\sigma^-}
\end{aligned} \tag{4.16}$$

with respect to $\mathbb{PP}^{(t_{i+1} - t_{i-1})} \otimes \mathbb{WF}_{0,\boldsymbol{\theta}}^{(x_{i-1})} \otimes \text{Leb}([0,1])$, which is the same dominating measure as that for (4.12). The only problematic term in (4.16) would be the non-neutral transition density appearing in the

denominator, however this quantity does not depend on any of the simulated quantities. Comparing (4.12) with (4.16), we thus conclude that we have access to exact updates for the interior path segments without need to pass a Metropolis–Hastings step. Note that despite the fact that we are taking a unit rate Poisson point process on an unbounded space as reference measure (which therefore has infinitely many points and cannot be simulated using finite computation), we only ever require computing finitely many points in view of Poisson thinning.

4.3.2 Updating an end path segment

The end path segment over the observation interval $[t_{n-1}, t_n]$ can be updated in a similar manner, however instead of bridge updates we make use of diffusion proposals. In particular, fix $X_{t_{n-1}} = x_{n-1}$ and

1. Draw $U \sim \text{Unif}([0,1])$,

$$X_{t_n} \sim \mathbb{WF}_{\sigma,\boldsymbol{\theta}}^{(x_{n-1}, t_n - t_{n-1})}$$

   (i.e. draw X_{t_n} from the law of a non-neutral Wright–Fisher diffusion started from x_{n-1}, sampled at time $t_n - t_{n-1}$) and check whether $U < \frac{1}{M_n} X_{t_n}^{Y_{t_n}} (1 - X_{t_n})^{n_{t_n} - Y_{t_n}}$, with $M_n := sup_{z \in [0,1]} z^{Y_{t_n}} (1-z)^{n_{t_n} - Y_{t_n}}$. If it is, then proceed to 2, otherwise keep drawing X_{t_n} and U until it is.

2. Conditional on $X_{t_n} = x$, draw

$$\begin{aligned}
\kappa_n &\sim \text{Pois}(\lambda_{\max}(t_n - t_{n-1})), \\
\{\psi_{n,j}\}_{j=1}^{\kappa_n} &\sim_{iid} \text{Unif}((t_{n-1}, t_n)), \\
\{\xi_{n,j}\}_{j=1}^{\kappa_n} &\sim_{iid} \text{Unif}((0, \lambda_{\max})), \\
\{\gamma_{n,j}\}_{j=1}^{\kappa_n} &\sim_{iid} \text{Unif}([0,1]), \\
\omega_n &\sim \mathbb{WF}_{0,\boldsymbol{\theta}}^{(t_n - t_{n-1}, x_{n-1}, x)}
\end{aligned}$$

3. If

$$\prod_{\{j:\xi_{n,j}\leq\lambda_\sigma\}} 1_{\left\{\frac{\varphi_\sigma(\omega_{n,\psi_{n,j}})-\varphi_\sigma^-}{\varphi_\sigma^+-\varphi_\sigma^-}<\gamma_{n,j}\right\}} = 1$$

set $\Psi_n = \{\psi_{n,j}\}$, $\Xi_n = \{\xi_{n,j}\}$, $\omega_n^\Psi = \{\omega_{n,\psi_{n,j}}\}$, $\Phi_n = (\Psi_n, \Xi_n, \omega_n^\Psi)$, else go to 2.

A proposal (X_{t_n}, Φ_n) generated by the above procedure has density

$$\begin{aligned} &X_{t_n}^{Y_{t_n}}(1-X_{t_n})^{n_{t_n}-Y_{t_n}} e^{\frac{\sigma}{2}(X_{t_n}-x_{n-1})-\varphi_\sigma^-(t_n-t_{n-1})} \\ &\quad\times \prod_{\{j:\xi_{n,j}\leq\lambda_\sigma\}} \frac{\varphi_\sigma^+-\varphi_\sigma(\omega_{n,\psi_{n,j}})}{\varphi_\sigma^+-\varphi_\sigma^-} \end{aligned} \tag{4.17}$$

with respect to the dominating measure $\mathbb{PP}^{(t_n-t_{n-1})} \otimes \mathbb{WF}_{0,\boldsymbol{\theta}}^{(x_{n-1})} \otimes \text{Leb}([0,1])$, where $\mathbb{WF}_{0,\boldsymbol{\theta}}^{(x_{n-1})}$ is the law of a Wright–Fisher diffusion started at x_{n-1}. Note that this dominating measure is the same one as for (4.12), and upon comparing (4.13) and (4.17) we deduce that we can update the corresponding auxiliary variables from their posterior distribution directly.

4.3.3 Updating the initial path segment

Updating the initial path segment is slightly more involved due to the dependence on both the allele age and the timestamp associated to the first non-zero observation in the generating mechanism. In particular, we first draw a proposal for the allele age, conditional on this draw generate proposals for the value of the diffusion at the first sampling time, and subsequently draw skeleton points conditional on both the allele age and the diffusion value. The length of path being proposed depends on $t_c (= \min\{t_i : Y_{t_i} > 0\})$, as does the proposal mechanism, however for simplicity we outline the procedure for the case when $t_c = t_1$, and defer the cases $t_c = t_2$ and $t_c = t_3$ (with the relevant calculations and computations of associated acceptance probabilities) to Appendix A. Naturally t_c can take on any value out of the observation times, however as pointed out previously, the later t_c appears, the longer the proposed path needs to be and the less effi-

cient the algorithm becomes due to significantly lower acceptance rates.

So assume that $t_c = t_1$, and thus the initial path segment is set to be over the observation interval $[t_0, t_2]$. Fixing $X_{t_2} = x_2$, we

1. Draw $t_0 \sim q_2(\cdot)$ for some proposal distribution having density q_2 with respect to $\mathrm{Leb}((-\infty, t_1))$

2. Conditional on $t_0 = t$, draw $U \sim \mathrm{Unif}([0,1])$,

$$X_{t_1} \sim \mathbb{WF}_{\sigma,\boldsymbol{\theta}}^{(t_2-t,0,x_2,t_1-t)}$$

   (i.e. draw X_{t_1} from the law of a non-neutral Wright–Fisher bridge going from 0 to x_2 in time $t_2 - t$, sampled at $t_1 - t$), and check whether $U < \frac{1}{M_1} X_{t_1}^{Y_{t_1}} (1 - X_{t_1})^{n_{t_1} - Y_{t_1}}$ with $M_1 := \sup_{z \in [0,1]} z^{Y_{t_1}} (1 - z)^{n_{t_1} - Y_{t_1}}$. If it is, then proceed to 3, otherwise redraw X_{t_1} and U as above until the condition holds.

3. Conditional on $t_0 = t$, $X_{t_1} = x$, draw

$$\begin{aligned}
\kappa_1 &\sim \mathrm{Pois}(\lambda_{\max}(t_1 - t)), \\
\{\psi_{1,j}\}_{j=1}^{\kappa_1} &\sim_{iid} \mathrm{Unif}((t, t_1)), \\
\{\xi_{1,j}\}_{j=1}^{\kappa_1} &\sim_{iid} \mathrm{Unif}((0, \lambda_{\max})), \\
\{\gamma_{1,j}\}_{j=1}^{\kappa_1} &\sim_{iid} \mathrm{Unif}([0,1]), \\
\omega_1 &\sim \mathbb{WF}_{0,\boldsymbol{\theta}}^{(t_1-t,0,x)}
\end{aligned}$$

   and

$$\begin{aligned}
\kappa_2 &\sim \mathrm{Pois}(\lambda_{\max}(t_2 - t_1)), \\
\{\psi_{2,j}\}_{j=1}^{\kappa_2} &\sim_{iid} \mathrm{Unif}((t_1, t_2)), \\
\{\xi_{2,j}\}_{j=1}^{\kappa_2} &\sim_{iid} \mathrm{Unif}((0, \lambda_{\max})), \\
\{\gamma_{2,j}\}_{j=1}^{\kappa_2} &\sim_{iid} \mathrm{Unif}([0,1]), \\
\omega_2 &\sim \mathbb{WF}_{0,\boldsymbol{\theta}}^{(t_2-t_1,x,x_2)}
\end{aligned}$$

4. If

$$\prod_{i=1}^{2} \prod_{\substack{\{j:\xi_{i,j}\le\lambda_\sigma,\\ \psi_{1,j}<t_1-t\}}} 1_{\left\{\frac{\varphi_\sigma(\omega_{i,\psi_{i,j}})-\varphi_\sigma^-}{\varphi_\sigma^+-\varphi_\sigma^-}<\gamma_{i,j}\right\}} = 1$$

   proceed, else go back to 3.

5. Compute α as in (4.19), and run a Metropolis–Hastings accept-reject step. If we accept, set $\Psi_k = \{\psi_{k,j}\}$, $\Xi_k = \{\xi_{k,j}\}$, $\omega_k^\Psi = \{\omega_{k,\psi_{k,j}}\}$, $\Phi_k = (\Psi_k, \Xi_k, \omega_k^\Psi)$ for $k = 1, 2$, else keep the old values.

A proposal generated according to steps 1-4 above has density

$$\begin{aligned} q_2(t_0) X_{t_1}^{Y_{t_1}} (1 - X_{t_1})^{n_{t_1} - Y_{t_1}} \frac{e^{\frac{\sigma}{2}x_2 - \varphi_\sigma^-(t_2 - t_0)}}{p_\sigma^{\boldsymbol{\theta}}(t_2 - t_0, 0, x_2)} \\ \times \prod_{i=1}^{2} \prod_{\substack{\{j:\xi_{i,j}\le\lambda_\sigma,\\ \psi_{1,j}<t_1-t_0\}}} \frac{\varphi_\sigma^+ - \varphi_\sigma(\omega_{i,\psi_{i,j}})}{\varphi_\sigma^+ - \varphi_\sigma^-} \end{aligned} \tag{4.18}$$

with respect to

$$\mathbb{PP} \otimes \mathbb{PP}^{(t_2 - t_1)} \otimes \mathbb{WF}_{0,\boldsymbol{\theta}}^{(0)} \otimes \mathrm{Leb}((-\infty, t_1) \times [0, 1]),$$

where $\mathbb{PP}$ is the law of a unit rate Poisson point process on $(0, \infty)^2$, and $\mathbb{WF}_{0,\boldsymbol{\theta}}^{(0)}$ is the law of a Wright–Fisher diffusion started from 0, and again the dominating measure used here matches that for (4.14).

Combining (4.14) with the above (4.18), we get that if the current values of the chain are $t_0^{(k)}, X_{t_1}^{(k)}, \Phi_1^{(k)}$ and $\Phi_2^{(k)}$, then a proposal $(t_0, X_{t_1}, \Phi_1, \Phi_2)$ leads to the acceptance probability to be computed in

step 5 above to be given by

$$
\begin{aligned}
&\alpha\left((t_0, X_{t_1}, \Phi_1, \Phi_2)|(t_0^{(k)} X_{t_1}^{(k)}, \Phi_1^{(k)}, \Phi_2^{(k)})\right) \\
&= \min\left\{1, \frac{p(t_0, X_{t_1}, \Phi_1, \Phi_2|\mathbf{Y}, \sigma^{(k)}, \mathbf{X}_{-1}^{(k)}, \mathbf{\Phi}_{-1:2}^{(k)})}{p(t_0^{(k)}, X_{t_1}^{(k)}, \Phi_1^{(k)}, \Phi_2^{(k)}|\mathbf{Y}, \sigma^{(k)}, \mathbf{X}_{-1}^{(k)}, \mathbf{\Phi}_{-1:2}^{(k)})}\right. \\
&\qquad\left. \times \frac{q((t_0^{(k)}, X_{t_1}^{(k)}, \Phi_1^{(k)}, \Phi_2^{(k)})|(t_0, X_{t_1}, \Phi_1, \Phi_2))}{q((t_0, X_{t_1}, \Phi_1, \Phi_2)|(t_0^{(k)}, X_{t_1}^{(k)}, \Phi_1^{(k)}, \Phi_2^{(k)}))}\right\} \\
&= \min\left\{1, \frac{p_2(t_0)}{p_2(t_0^{(k)})} \frac{q_2(t_0|t_0^{(k)})}{q_2(t_0^{(k)}|t_0)} e^{-\varphi_\sigma^-\left(t_0^{(k)} - t_0\right)} \frac{p_0^{\boldsymbol{\theta}}(t_2 - t_0, 0, x_2)}{p_0^{\boldsymbol{\theta}}(t_2 - t_0^{(k)}, 0, x_2)}\right. \\
&\qquad\left. \times \frac{a(t_2 - t_0, 0, x_2, \sigma^{(k)})}{a(t_2 - t_0^{(k)}, 0, x_2, \sigma^{(k)})}\right\} \qquad (4.19)
\end{aligned}
$$

where $\mathbf{\Phi}_{-i:j}^{(k)}$ means the vector $\mathbf{\Phi}^{(k)}$ excluding the i^{th} up to the j^{th} entries, $p(t_0, X_{t_1}, \Phi_1, \Phi_2|\mathbf{Y}, \sigma^{(k)}, \mathbf{X}_{-1}^{(k)}, \mathbf{\Phi}_{-1:2}^{(k)})$ denotes the conditional density of $(t_0, X_{t_1}, \Phi_1, \Phi_2)$ given $\mathbf{Y}, \sigma^{(k)}, \mathbf{X}_{-1}^{(k)}, \mathbf{\Phi}_{-1:2}^{(k)}$, $p_2(t_0)$ is the prior on the allele age, $q_1(t_0|t_0^{(k)})$ is the proposal kernel for the allele age, and the remaining terms in the last expression on the RHS of (4.19) arise due to the ratio of non-neutral transition densities which have mismatching start times t_0 and $t_0^{(k)}$ in the denominator of (4.18). The problematic term here is the ratio of intractable quantities of the form $a(t, x, y, \sigma)$ which cannot be evaluated exactly, whilst the ratio of neutral transition densities can be easily targetted via a refinement scheme. In particular, one can show that for a sufficiently large number of terms, there exist monotonic upper and lower bounds on this quantity which converge to the true value (for exact details see Proposition 4 in [JS17]), and thus one can keep on adding terms to either bound until a decision can be made on whether to accept or reject the proposed move. All other terms are all computable (subject to choosing a suitable prior and proposal density for the allele age), and thus the final task is to deal with the intractable ratio $\frac{a(t_2 - t_0, 0, x_2, \sigma)}{a(t_2 - t_0^{(k)}, 0, x_2, \sigma)}$. To this end, we make use of a pseudo-marginal step for this update, where we replace the intractable terms with unbiased estimators provided by the Poisson estimator.

We now provide some detail on how the Poisson estimator for the quantity $a(t, x, y, \sigma)$ will be implemented in the above MCMC scheme. We start by introducing the quantities $\kappa \sim \text{Pois}(\lambda t)$, $\tau = \{\tau_j\}_{j=1}^{\kappa} \sim_{iid} \text{Unif}((0, t))$ and $\zeta \sim \mathbb{WF}_{0,\boldsymbol{\theta}}^{(t,x,y)}$, which we combine together to obtain an unbiased estimate for $a(t, x, y, \sigma)$ given by

$$e^{(\lambda - c)t} \prod_{j=1}^{\kappa} \left(\frac{c - \left(\varphi_\sigma(\zeta_{\tau_j}) - \varphi_\sigma^- \right)}{\lambda} \right),$$

where $\lambda \in \mathbb{R}, c > 0$ are two arbitrary constants. The estimator is unbiased and has second moment given by

$$e^{(\lambda - 2c)t} \mathbb{E}_{\mathbb{WF}_{0,\boldsymbol{\theta}}^{(t,x,y)}} \left[\exp \left\{ \int_0^t \frac{(c - \varphi_\sigma(X_s) - \varphi_\sigma^-)^2}{\lambda} ds \right\} \right], \tag{4.20}$$

which makes choosing a pair (λ, c) which minimises the variance non-trivial due to the associated Lebesgue integrals being incomputable. Instead, as discussed in Section 7 of [BPRF06], we upper bound the integrand in (4.20) and minimise the resulting quantity to get $\lambda = \varphi_\sigma^+ - \varphi_\sigma^-$, and $c = \lambda$. Setting $\zeta_\tau := \{\zeta_{\tau_j}\}_{j=1}^{\kappa}$, this choice leads to the unbiased estimator

$$\tilde{a}(t, x, y, \sigma, \tau, \zeta_\tau) = \prod_{j=1}^{\kappa} \left(\frac{\varphi_\sigma^+ - \varphi_\sigma(\zeta_{\tau_j})}{\varphi_\sigma^+ - \varphi_\sigma^-} \right). \tag{4.21}$$

It now remains to detail how and when we employ the above unbiased estimate. As illustrated in [AR09, Bea03] (see in particular Table 1 in [AR09]), if one re-estimates $a(t, x, y, \sigma)$ using (4.21) each time the quantity needs to be evaluated, the resulting MCMC scheme does not target the correct posterior of interest. Instead, if the estimator is re-calculated solely when a new proposal is introduced with the resulting estimate being stored for an accepted proposal to be used in the subsequent iterations (so we augment the state space with the corresponding τ and ζ_τ and recycle these in the next iteration), then the resulting pseudo-marginal scheme does indeed target the correct distribution. By making use of this pseudo-marginal formulation and expression for $\tilde{a}(t, x, y, \sigma, \tau, \zeta_\tau)$, the pseudo-marginal Metropolis–Hastings acceptance

probability becomes computable:

$$\alpha = \min\left\{1, \frac{p_2(t_0)}{p_2(t_0^{(k)})}\frac{q_2(t_0^{(k)}|t_0)}{q_2(t_0|t_0^{(k)})}e^{-\varphi_\sigma^-\left(t_0^{(k)}-t_0\right)}\frac{p_0^{\boldsymbol{\theta}}(t_2-t_0,0,x_2)}{p_0^{\boldsymbol{\theta}}(t_2-t_0^{(k)},0,x_2)} \times \frac{\tilde{a}(t_2-t_0,0,x_2,\sigma,\tau,\zeta_\tau)}{\tilde{a}(t_2-t_0^{(k)},0,x_2,\sigma,\tau^{(k)},\zeta_{\tau^{(k)}}^{(k)})}\right\} \tag{4.22}$$

where τ, ζ_τ are the variables generated jointly with the proposals t_0, X_{t_1}, Φ_1 and Φ_2 to unbiasedly estimate $a(t_2 - t_0, 0, x_2, \sigma)$, whilst $\tau^{(k)}, \zeta_{\tau^{(k)}}^{(k)}$ are the stored values which were used in the previous iteration. So replacing the acceptance probability (4.19) by (4.22) leads to an implementable pseudo-marginal Metropolis–Hastings update over the initial path segment.

A full detailed account of the proposal procedures for the cases when $t_c = t_2, t_3$ together with associated computation of acceptance probabilities can be found in Appendix A, but regardless of the value of t_0 and t_c the above illustrated problem involving intractable terms persists and we resort to a pseudo-marginal Metropolis–Hastings update.

4.3.4 Selection coefficient update

Finally, we illustrate the selection coefficient update which is a straightforward Metropolis–Hastings step, with the joint density of the data, selection coefficient, allele age, latent path at the observation times, and skeleton points in this case being given by

$$p_1(\sigma)p_2(t_0)\prod_{i=1}^{n}\mathcal{B}_{n_{t_i},X_{t_i}}(Y_{t_i})e^{\frac{\sigma}{2}X_{t_n}-\varphi_\sigma^-(t_n-t_0)} \times \prod_{i=c+1}^{n}\prod_{\substack{\{j:\xi_{i,j}\le\lambda_\sigma,\\ \psi_{1,j}<t_1-t_0\}}}\frac{\varphi_\sigma^+-\varphi_\sigma(\omega_{i,\psi_{i,j}})}{\varphi_\sigma^+-\varphi_\sigma^-}.$$

with respect to the dominating measure

$$\mathbb{PP} \otimes \mathbb{PP}^{(t_n - t_c)} \otimes \mathbb{WF}^{(0)}_{0,\boldsymbol{\theta}} \otimes \mathrm{Leb}(E^n_{t_1}) \otimes \Sigma\left(\otimes_{i=1}^n n_{t_i}\right),$$

where we recall that $E^n_{t_1} = [0,1]^n \times \mathbb{R} \times (-\infty, t_1)$. On the other hand, we propose new values of the selection coefficient according to some proposal kernel having density q_1 with respect to Lebesgue measure on $\mathbb{R}$, such that the acceptance probability for a proposed move to σ' given that the current value of the selection coefficient is σ evaluates to

$$\alpha = \min\left\{1, \frac{q_1(\sigma|\sigma')}{q_1(\sigma'|\sigma)} \frac{p_1(\sigma')}{p_1(\sigma)} e^{\frac{\sigma'-\sigma}{2} X_{t_n} - (t_n - t_0)\left(\varphi^-_{\sigma'} - \varphi^-_{\sigma}\right)} \times \frac{\prod_{i=c}^n \prod\limits_{\substack{\{j:\xi_{i,j} \le \lambda_\sigma,\\ \psi_{c,j} < t_1 - t_0\}}} \frac{\varphi^+_{\sigma'} - \varphi_{\sigma'}(\omega_{i,\psi_{i,j}})}{\varphi^+_{\sigma'} - \varphi^-_{\sigma'}}}{\prod_{i=c}^n \prod\limits_{\substack{\{j:\xi_{i,j} \le \lambda_\sigma,\\ \psi_{c,j} < t_1 - t_0\}}} \frac{\varphi^+_{\sigma} - \varphi_{\sigma}(\omega_{i,\psi_{i,j}})}{\varphi^+_{\sigma} - \varphi^-_{\sigma}}}\right\} \tag{4.23}$$

where all the terms involved are computable, and $p_1(\sigma)$ is the prior we impose on the selection coefficient.

Algorithm 3: Metropolis-within-Gibbs Sampler to conduct exact inference on σ, t_0 given noisy observations $\mathbf{Y}$

Initialise $\sigma^{(0)}, t_0^{(0)}, \mathbf{X}^{(0)}, \mathbf{\Phi}^{(0)}, \tau^{(0)}, \zeta_\tau^{(0)}$
repeat
  Draw $\sigma' \sim q$, compute $\lambda_{\max} := \max\{\lambda_{\sigma^{(k)}}, \lambda_{\sigma'}\}$
  repeat
    repeat
      Propose $(t_0, X_{t_1}, \{\Phi_i\}_{i=1}^2, \tau, \zeta_\tau)$ conditional on
        $\sigma^{(k)}, \lambda_{\max}, X_{t_2}^{(k)}, n_{t_1}, Y_{t_1}$ as in Section 4.3.3
      Compute α as in (4.19), draw $U_1 \sim \text{Unif}([0,1])$
      if $\alpha < U_1$ **then**
        $(t_0^{(k+1)}, X_{t_1}^{(k+1)}, \{\Phi_i^{(k+1)}\}_{i=1}^2) \leftarrow (t_0, X_{t_1}, \{\Phi_i\}_{i=1}^2)$
        $(\tau^{(k+1)}, \zeta_\tau^{(k+1)}) \leftarrow (\tau, \zeta_\tau)$
      else
        $(t_0^{(k+1)}, X_{t_1}^{(k+1)}, \{\Phi_i^{(k+1)}\}_{i=1}^2) \leftarrow$
          $(t_0^{(k)}, X_{t_1}^{(k)}, \{\Phi_i^{(k)}\}_{i=1}^2)$
        $(\tau^{(k+1)}, \zeta_\tau^{(k+1)}) \leftarrow (\tau^{(k)}, \zeta_\tau^{(k)})$
      end
    for $i = 2, \ldots, n-1$ **do**
      Update $(X_{t_i}, \{\Phi_k\}_{k=i-1}^{k=i})$ conditional on
        $X_{t_{i-1}}^{(k+1)}, X_{t_{i+1}}^{(k)}, \lambda_{\max}, \sigma^{(k)}, n_{t_i}, Y_{t_i}$ as in Section 4.3.1
    end
    Update (X_{t_n}, Φ_n) conditional on
      $X_{t_{n-1}}^{(k+1)}, \lambda_{\max}, \sigma^{(k)}, n_{t_n}, Y_{t_n}$ as in Section 4.3.2
  Compute $\alpha(\sigma'|\sigma^{(k)})$ as in (4.23), draw $U \sim \text{Unif}([0,1])$
  if $\alpha(\sigma'|\sigma^{(k)}) < U$ **then**
    $\sigma^{(k+1)} \leftarrow \sigma'$
  else
    $\sigma^{(k+1)} \leftarrow \sigma^{(k)}$
  end
until *convergence*

4.4 Simulation results

The inferential framework presented above has been implemented in C++ [1], and tested on simulated data for the case when $t_c = t_1$. By making use of the exact algorithm, we were able to simulate a non-neutral Wright–Fisher diffusion and superimpose binomial sampling to create a synthetic dataset. We tested our method on a set of different prior distributions for both the selection coefficient σ and the allele age t_0, whilst maintaining the same simulated data as input, in order to better assess the influence of the prior on the inferential results. In simulating the data, we set the following parameters:

Datapoints	σ	θ_1	θ_2	t_0	t_1	$t_i - t_{i-1}$	n_{t_i}
6	10	0.1	0.1	0.2	0.5	0.1	20

Table 4.4.1: List of the parameter configurations for the simulated dataset.

where the small number of datapoints and sample sizes n_{t_i}, as well as the order of magnitude of the mutation rates was chosen to be in line with similar datasets in the literature (see for instance [LPR$^+$09, BYN, W$^+$16, F$^+$19]). In choosing the observation time spacing, a big discrepancy was noted in the effective population size ranges used in different studies for the same dataset (see for instance the horse coat colouration dataset from [LPR$^+$09] where suitable ranges for the effective population size vary from $(10^3, 10^6)$ in [LPR$^+$09], to $(2500, 10^3)$ in [SBS14], to $(200, 5000)$ in [MMES12]). In view of the fact that this directly affects the diffusion time scaling (which is recovered by dividing time in years by a factor of $2N_e g$, where N_e is the effective population size and g is the generation gap), we chose a spacing of 0.1 such that the resulting observations are not spaced too far apart, nor too closely together. As for the choice of selection coefficient, we wanted to emulate the case when the selection coefficient is significantly different from 0, but not too large such that a diffusion approximation is no longer an appropriate model for the underlying dynamics. We envisaged that there would not be sufficient signal in the data to allow for the allele age to be accurately picked

[1] Available for download from `https://github.com/JaroSant/MCMC4WF`

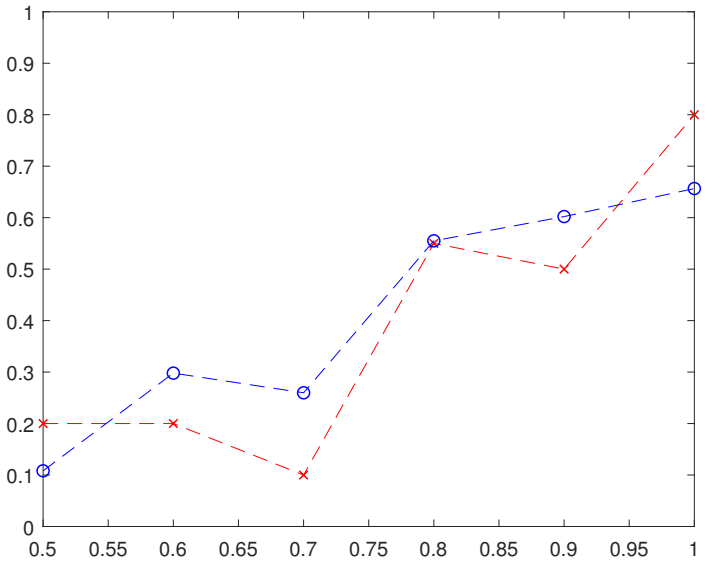


Figure 4.4.1: The blue circles denote the exact draws from the Wright–Fisher diffusion, whilst the red crosses are the binomial draws obtained. The straight lines in between observations are a linear interpolation.

up on, and thus chose a value which was relatively close to the first sampling time.

Using the setup in Table 4.4, the path generated is given in Figure 4.4.1 in blue, whilst the superimposed binomial sampling is given by the red plot. Tables 4.4.2 and 4.4.3 lists the different configurations of priors imposed on the selection coefficient and allele age respectively. Gaussian distributions were chosen as priors for the selection coefficient, whereas for the allele age we made use of suitably transformed exponential and gamma distributions. In view of the fact that the allele age cannot take on values later than the first non-zero observation time t_c, we set $t_0 = t_c - \varepsilon - z$, for ε a small tolerance parameter (in the simulations this was set to 0.05 to ensure that the method does not run into computational issues - see Subsection 4.4.3 for more details) and z the random variable which takes on either an exponential or gamma distribution.

Case	σ Priors
A	$\mathrm{N}(-10, 10)$
B	$\mathrm{N}(-5, 10)$
C	$\mathrm{N}(0, 10)$
D	$\mathrm{N}(5, 5)$

Table 4.4.2: Priors for the selection coefficient for each different case.

Case	t_0 Priors
E	$\mathrm{Exp}(3)$
F	$\mathrm{Exp}(1.5)$
G	$\Gamma(2, 0.2)$
H	$\Gamma(1.5, 0.5)$

Table 4.4.3: Priors for the allele age for each different case.

With regards to proposals, we used a Gaussian proposal centred at the previous value with a standard deviation of 10 (which was chosen after some tuning) for the selection parameter, whilst for the allele age we made use of a truncated Gaussian proposal (again centred at the previous value) with standard deviation 0.25 (with the truncation threshold being set to a small value just below the first non-zero observation $t_c - \varepsilon$, where the ε was chosen small enough such that no computational issues would hamper the inferential scheme - refer to Subsection 4.4.3).

Parameter	Proposals
σ	$\mathrm{N}(\cdot, 10)$
t_0	$\mathrm{tN}(\cdot, 0.25, t_c - \varepsilon)$

Table 4.4.4: Proposal distributions for the selection coefficient and allele age, where tN denotes a truncated normal distribution.

Convergence was monitored by checking whether the maximal difference in the value of the mean (correspondingly standard deviation) of the generated samples in the last k iterations fell below a user specified threshold level $\varepsilon_{\sigma,m}$ (respectively $\varepsilon_{\sigma,s}$) for both the selection parameter and allele age. So at iteration N, say, we checked whether

$$\max_{i,j\in\{N-k,\ldots,N\}} |\bar{\sigma}_i - \bar{\sigma}_j| < \varepsilon_{\sigma,m}, \qquad \max_{i,j\in\{N-k,\ldots,N\}} |\hat{\sigma}_i - \hat{\sigma}_j| < \varepsilon_{\sigma,s} \tag{4.24}$$

where $\bar{\sigma}_i := i^{-1}\sum_{j=1}^{i}\sigma^{(j)}$ and $\hat{\sigma}_i := \sqrt{(i-1)^{-1}\sum_{j=1}^{i}(\sigma^{(j)} - \bar{\sigma}_j)^2}$, the sample mean and standard deviation, with a similar criterion in place for the allele age. In our runs, we required $k = 10,000$ iterations,

$\varepsilon_{\sigma,m} = 0.01$, $\varepsilon_{\sigma,s} = 0.1$, $\varepsilon_{t_0,m} = 0.001$, and $\varepsilon_{t_0,s} = 0.01$. We found these bounds to be quite stringent and several of the runs plotted here had still not converged according to the above criteria, however the output seems to indicate that the draws are indeed coming from (a reasonable approximation to) the stationary distribution of the Markov chain.

We now proceed to consider the performance of the method when varying the prior for the selection parameter through cases A to D in Subsection 4.4.1, and then do the same with the allele age by considering cases E through to H in Subsection 4.4.2. Throughout Subsection 4.4.1, we assume that the prior on the allele age is an Exp(3), whilst in Subsection 4.4.2 we set the prior on the selection parameter to be $\mathrm{N}(-5, 10)$.

4.4.1 Different priors for the selection coefficient

As evidenced in Figure 4.4.2, the selection parameter is mixing relatively well with an acceptance rate in the range $(0.3, 0.4)$ across all test cases, with the autocorrelation plots in Figure 4.4.3 further confirming a sharp decay in the autocorrelation between returned samples within a relatively short time lag.

By comparing the kernel smoothed posterior (black line), the prior (red dotted line), likelihood evaluations (orange circles) and truth (magenta) in Figure 4.4.4, we conclude that the method seems to be doing reasonably well in detecting signatures of selection even in cases when the prior assigns most of its mass away from the truth. The posterior is concentrating around the true value despite having a relatively diffuse prior, and thus seems to suggest that the influence of the data (via the likelihood) is filtering through. We point out that the smoothed posterior and prior are plotted against the left axis, whilst the likelihood points (i.e. the contributions to the likelihood which depend on σ) are plotted against the right axis. Note that comparing all three on the same axis is infeasible due to the normalising constant for the likelihood evaluations being intractable and hard to estimate via Riemann sum type approximations, as it is very noisy in view of

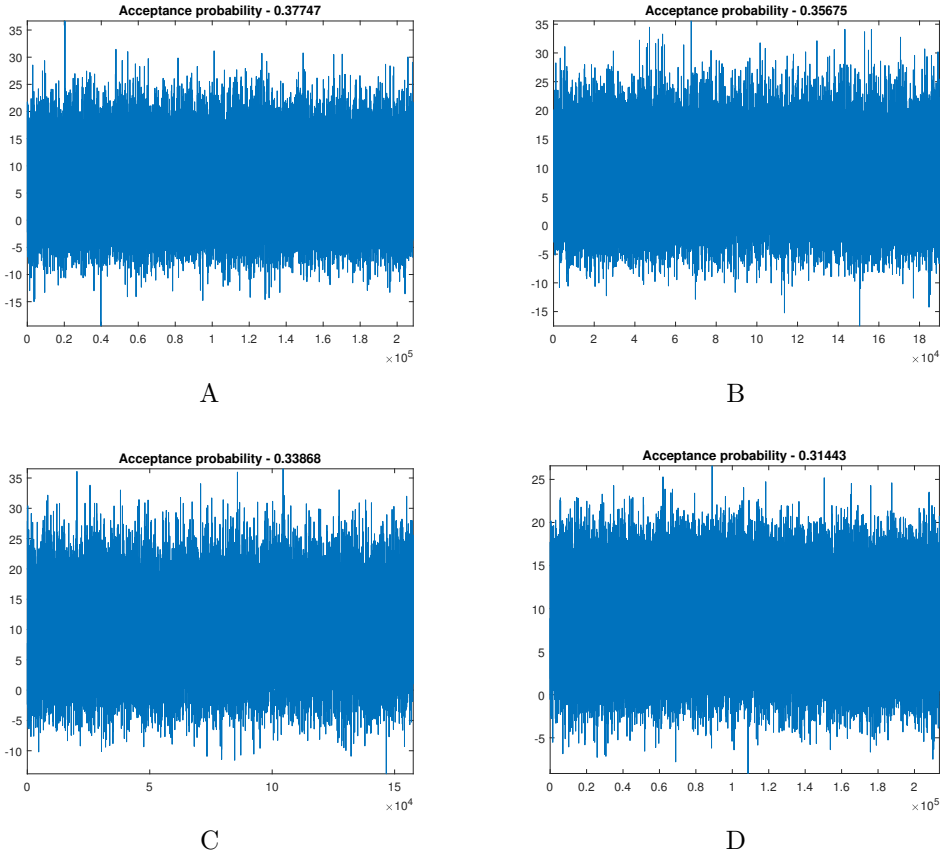


Figure 4.4.2: Traceplots for σ for the 4 cases

its dependence on $\mathbf{X}, \mathbf{\Phi}$ and t_0. Finally, the plots for the mean and standard deviation of σ over the last 10,000 iterations are presented below. Although they did not fall below the user specified threshold (4.24), they seem to display sufficiently constant behaviour to indicate that convergence may have been reasonably achieved.

4.4.2 Different priors for the allele age

We now inspect the output obtained for the allele age t_0, by first looking at the traceplots and autocorrelation functions. In this case the acceptance rates are relatively high at roughly 0.65 across all test cases, and the ACF plots indicate that any autocorrelation is decaying reasonably quickly for cases E and G. We point out that cases F and H have more diffuse priors, leading to values of the allele age straying further into

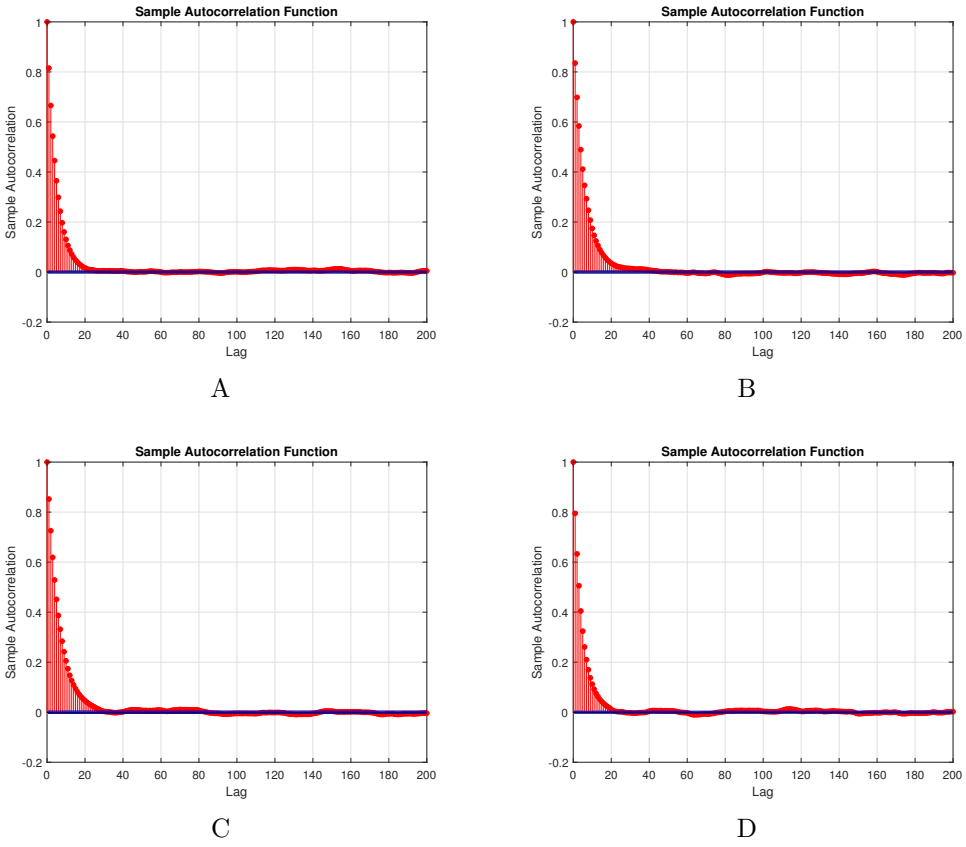


Figure 4.4.3: Autocorrelation plots for σ for the 4 cases

the negative values (as evidenced in Figures 4.4.7 B and D). This in turn leads to a larger number of Poisson points being favoured in the simulations, and thus the algorithm in these two cases resulted to run significantly slower than in cases E and G, explaining the higher autocorrelations observed in Figure 4.4.8. Turning to the comparison of priors, likelihood evaluations and smoothed posteriors in Figure 4.4.9, the situation is quite interesting - for relatively restrictive prior distributions that concentrate most of their mass around the true value, the kernel smoothed posterior (black line) concentrates well around the truth but is also quite close to the prior distribution (red dotted line). On the other hand, the likelihood evaluations (orange circles) seem to suggest that the likelihood does not die out the further back in time the allele age is. This apparent flatness in the likelihood comes as no surprise, as one would expect there to be relatively little information

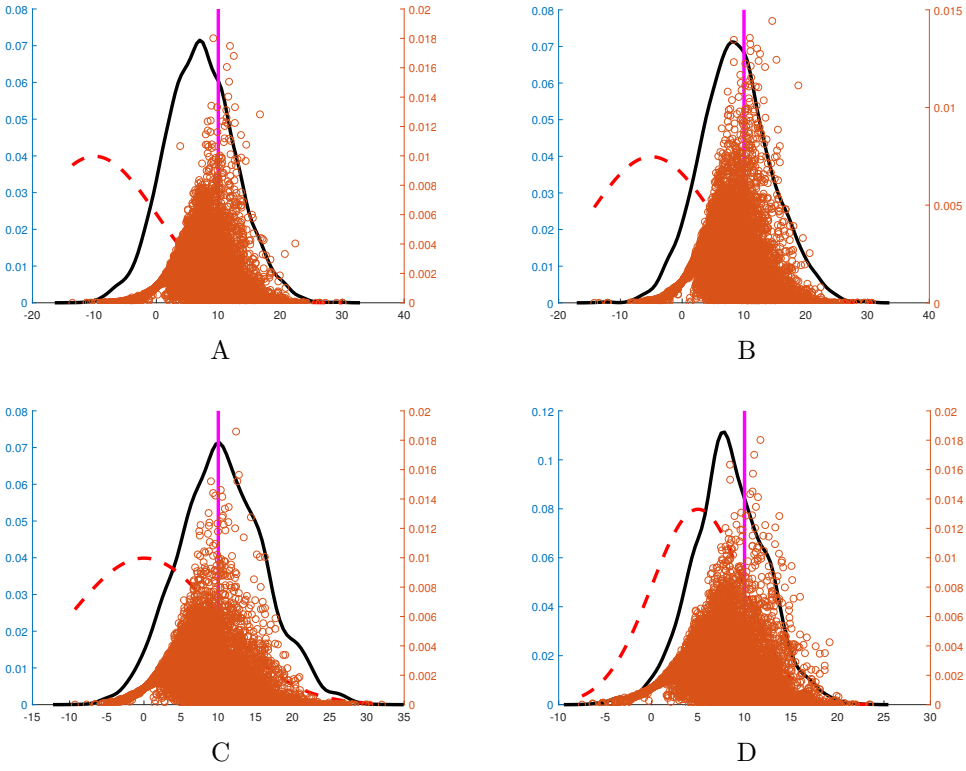


Figure 4.4.4: Plots of the prior (dotted red line), likelihood (orange circles), kernel smoothed posterior (black line) and the truth (magenta). The prior and smoothed posterior are plotted against the left axis whilst the likelihood is plotted against the right axis.

in the data about the allele age to start with. However the persistence displayed by the likelihood could potentially be explained by the fact that the transition density for the diffusion gets arbitrarily close to the invariant density (which is independent of t_0 and thus contributes towards the observed invariance in time). It also suggests that having relatively diffuse priors on the allele age is not necessarily the best option, as for a sufficiently large time increment, the diffusion might well have been started from stationarity. In light of this, perhaps a better strategy would be to restrict the state space for the allele age to a finite continuous range immediately before the first non-zero observation time, say $[t_c - \delta, t_c - \varepsilon]$, for some threshold value δ, together with some state Δ which will capture all the time before $t_c - \delta$, i.e. $t_0 \in \Delta \cup [t_c - \delta, t_c - \varepsilon]$. Naturally the threshold δ would need to be

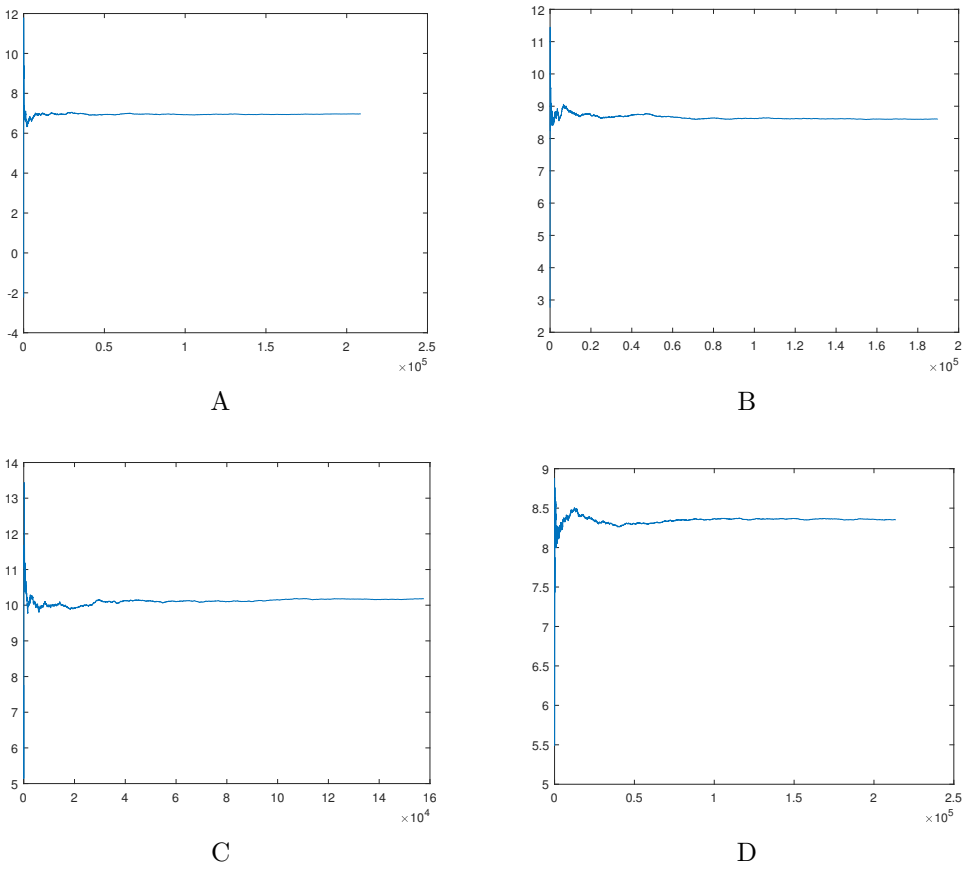


Figure 4.4.5: Plot of the mean of σ for the last 10,000 iterations

chosen large enough to allow the diffusion to reach equilibrium within this time interval, such that allele ages that reach further back in time lead to a time homogeneous diffusion. By appealing to the duality between the neutral Wright–Fisher diffusion and the Kingman coalescent ([Gri79, Gri80, Tav84]), a first approximation for δ would be the value 2 as it turns out to be the expected time it takes for the Kingman coalescent to get to the most recent common ancestor, thereby ensuring that the diffusion is at equilibrium. We point out however that these considerations only apply for the neutral Wright–Fisher case, so the appropriate dual process for the non-neutral Wright–Fisher diffusion (that is, the ancestral selection graph [KN97, NK97]) would need to be consulted for a more appropriate choice of δ. As done in the case of selection, convergence was diagnosed by checking whether the quan-

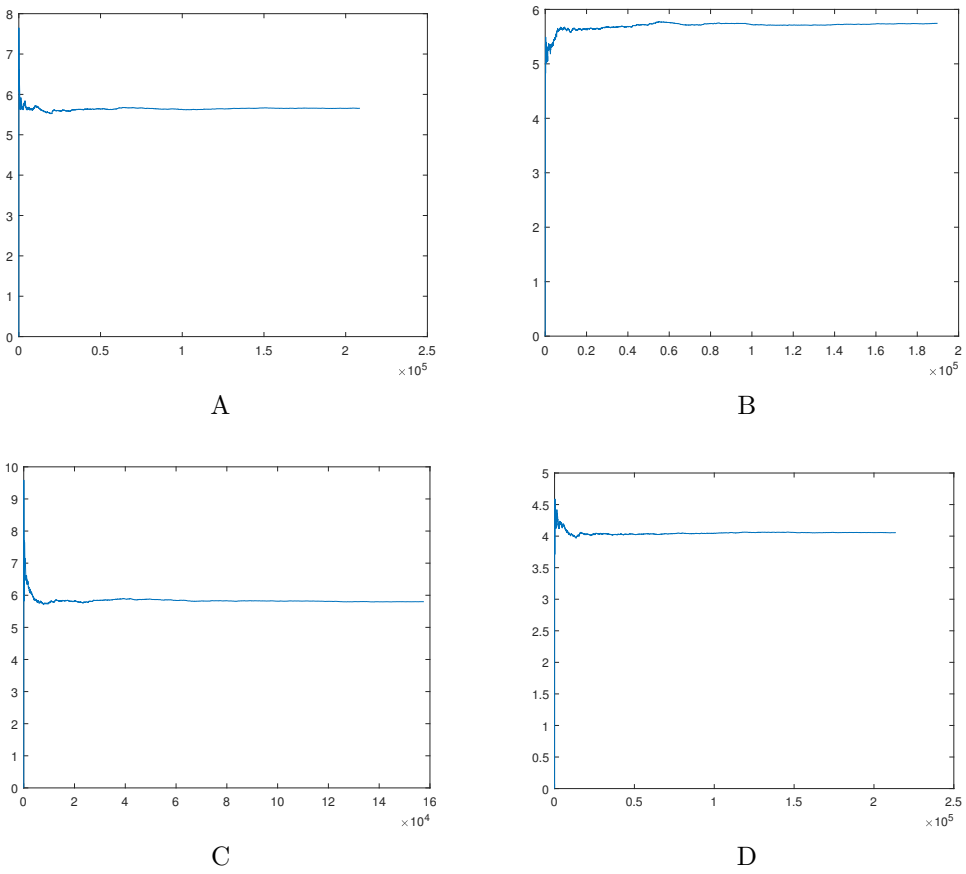


Figure 4.4.6: Plot of the standard deviation of σ for the last 10,000 iterations

tities (4.24) (suitably translated for the allele age) fell below the user specified thresholds. Figures 4.4.10 and 4.4.11 again suggest that the samples being returned can be treated as coming from the invariant measure, despite the fact that the convergence criteria set out have as yet not been met.

4.4.3 Computational considerations

The algorithm was implemented in C++ and run on an Intel Xeon E5-2670 v2 CPU with a dual multithreaded deca-core and 64GB of memory. Runtime lasted roughly 28 days which is quite a long time, however the convergence criteria we set were quite strict, and one could reasonably argue (based on Figures 4.4.5, 4.4.6, 4.4.10 and 4.4.11) that samples are coming from the posterior distribution despite

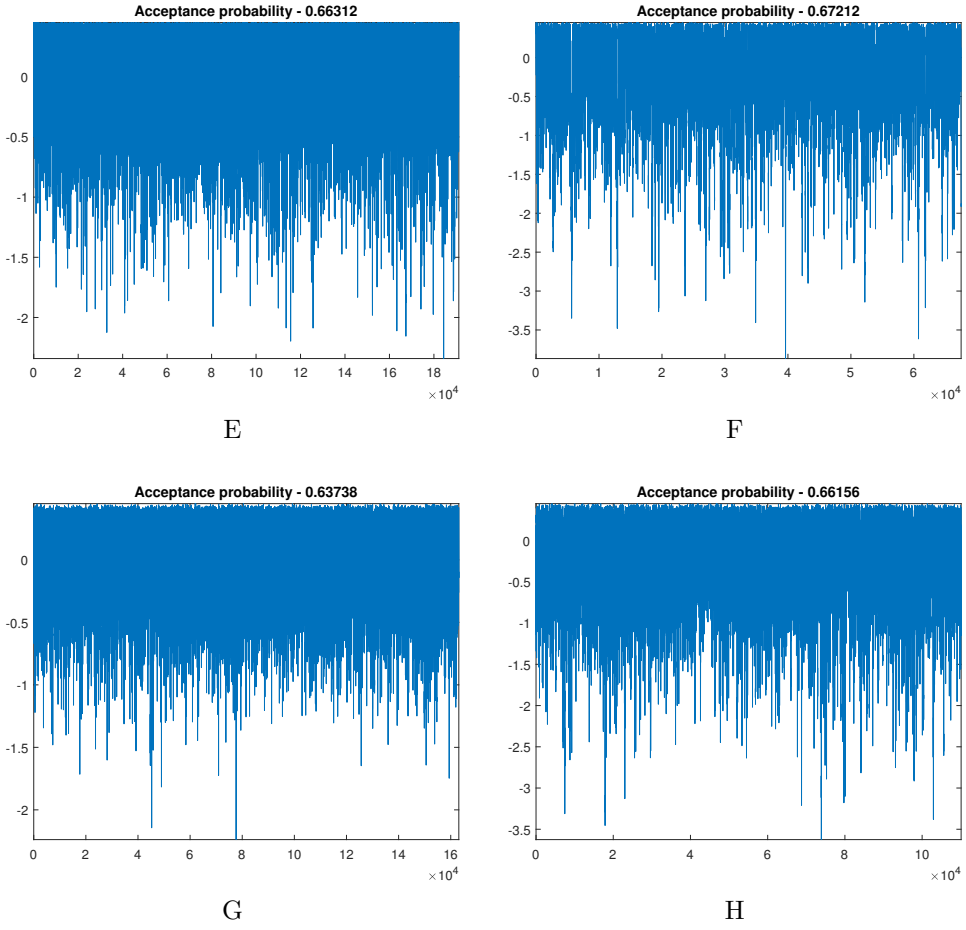


Figure 4.4.7: Traceplots for t_0 for the 4 cases

the chain not having formally converged. The main bottleneck in the updating procedure is the simulation making use of non-neutral bridge proposals together with the skeleton point proposals, in virtue of the rejection sampler being used there. With larger selection coefficients, the non-neutral bridge sampler becomes less efficient because the neutral proposals become less suitable candidates in comparison to the target paths. To this end, we decided to test our method on simulated data where the selection coefficient was not too large but also significantly different from 0, thereby ensuring that there was sufficient signal in the data for the method to pick up on.

Although the proposed method is computationally demanding,

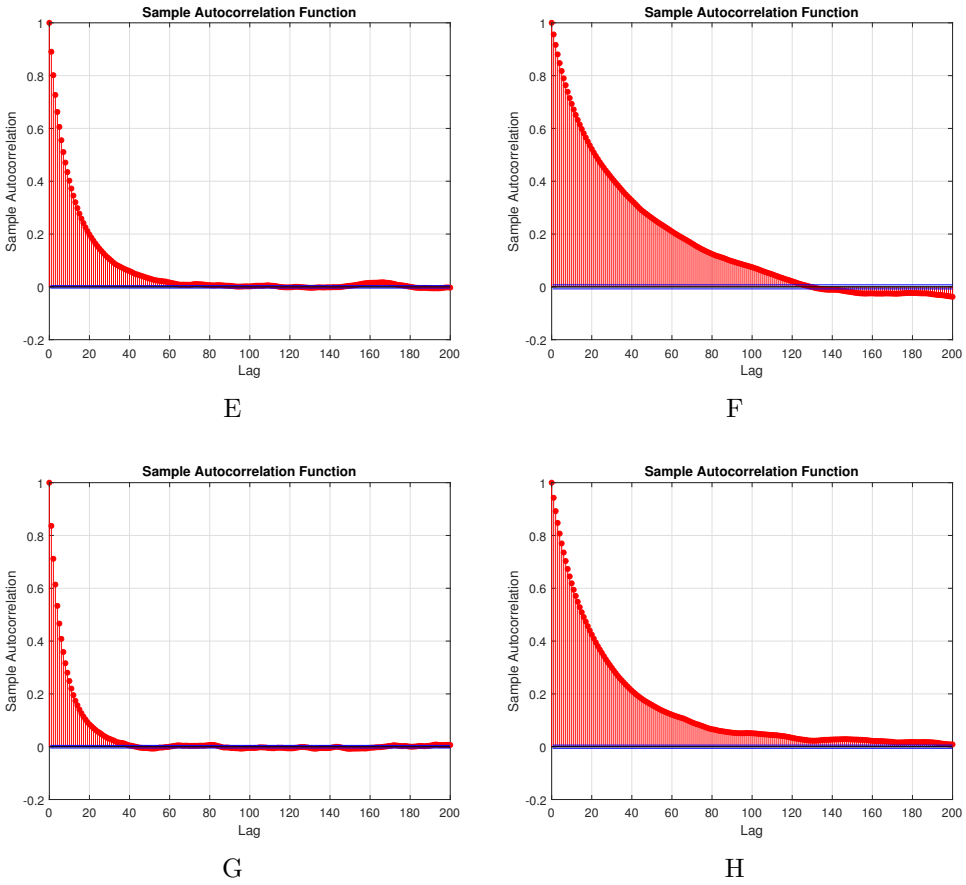


Figure 4.4.8: Autocorrelation plots for t_0 for the 4 cases

we emphasise that when dealing with aDNA time series data we are not in the standard statistical setting where we need to devise efficient statistical techniques which can both produce reliable results and deal with thousands of datapoints. On the contrary one can only hope for a handful of readings and a limited number of sampling times, as these aDNA datasets rely on sufficiently well-preserved fossils and remains found in archaeological excavation sites, which further need to be amenable to the extraction of genetic material. In addition, although advances in aDNA extraction techniques have lead to a significant increase in the number of such datasets over the past decade or so, this has not been accompanied by a considerable improvement in the number of observations being recovered or an increase in the number of sampling times. In this context, one should therefore concentrate on

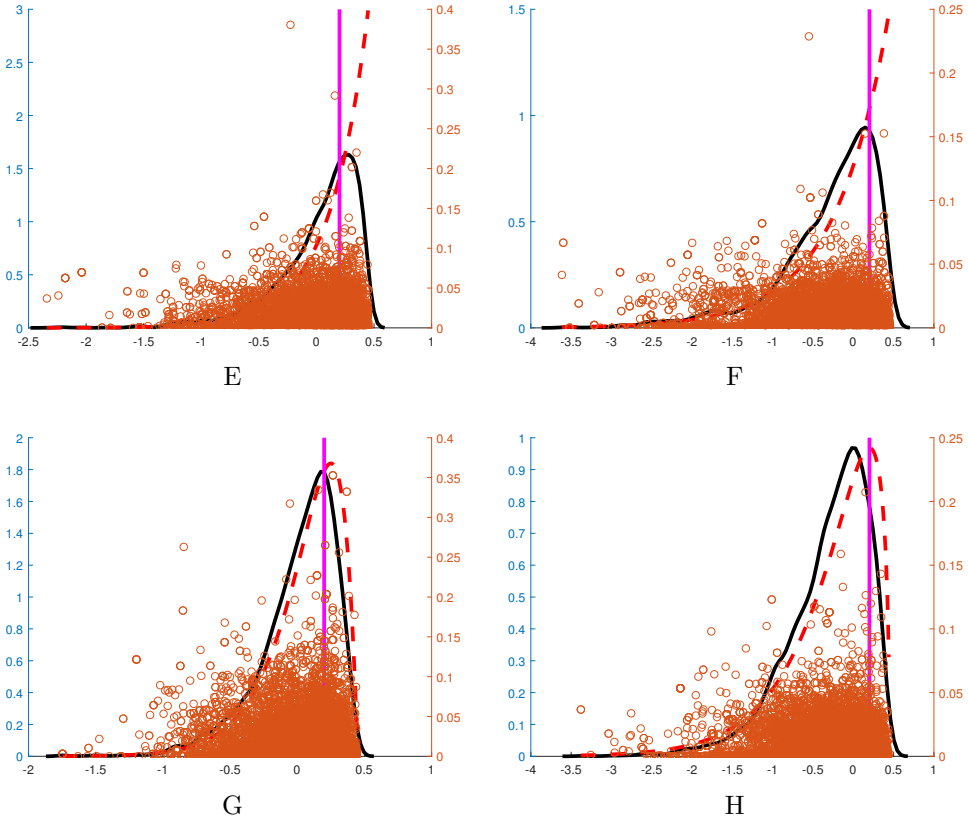


Figure 4.4.9: Plots of the prior (dotted red line), likelihood (orange circles), kernel smoothed posterior (black line) and the truth (magenta). The prior and smoothed posterior are plotted against the left axis whilst the likelihood is plotted against the right axis.

developing statistical procedures which elicit as much information as possible from the few available datapoints at the cost of entertaining inferential schemes which are more computationally onerous.

As mentioned at the end of Subsection 4.2.1, the exact algorithm runs into some computational issues whenever the time increment becomes smaller than 0.06. To account for this and to ensure that no numerical instabilities are present in the method, we employed the approximations in (4.3), such that a discretised Gaussian is used in the exact simulation approach as detailed in [JS17]. We point out here that this approximation is not being applied to the underlying diffusion itself, but rather to the ancestral block counting process

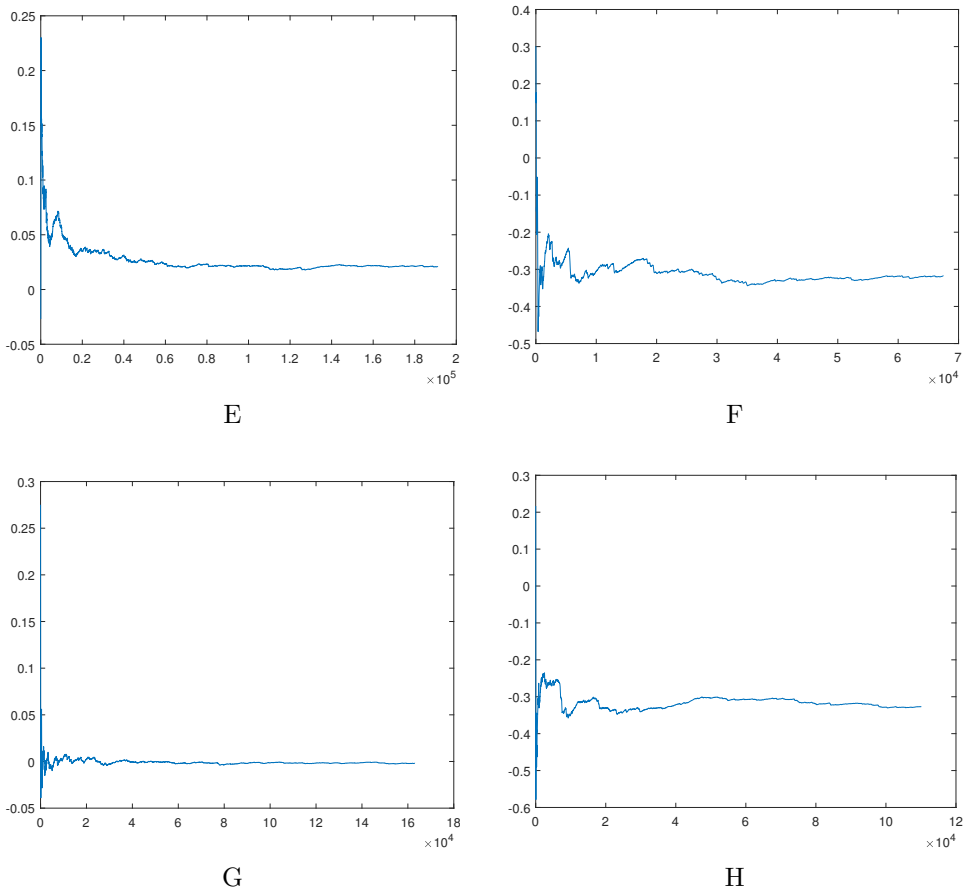


Figure 4.4.10: Plot of the mean of t_0 for the last 10,000 iterations

of the dual process to the neutral Wright–Fisher diffusion (i.e. the Kingman coalescent). Although this approximation does well for time increments between $(0.01, 0.06)$, further computational issues arises if the time increment goes significantly below 0.01 in the bridge sampler. Observe that μ in (4.3) grows proportionally to t^{-1}, thus for very small time increments the approximating discretised Gaussian returns a very large value which we need to sum over twice in the bridge simulation. In turn, whenever the allele age being proposed was very close to the time corresponding to the first non-zero observation t_c, the resulting algorithm would spend a large amount of time computing the corresponding sums. To avoid this computational bottleneck, and speed up the approximation procedure, we allow the allele age to vary within the interval $(-\infty, t_c - \varepsilon)$ for some small tolerance parameter

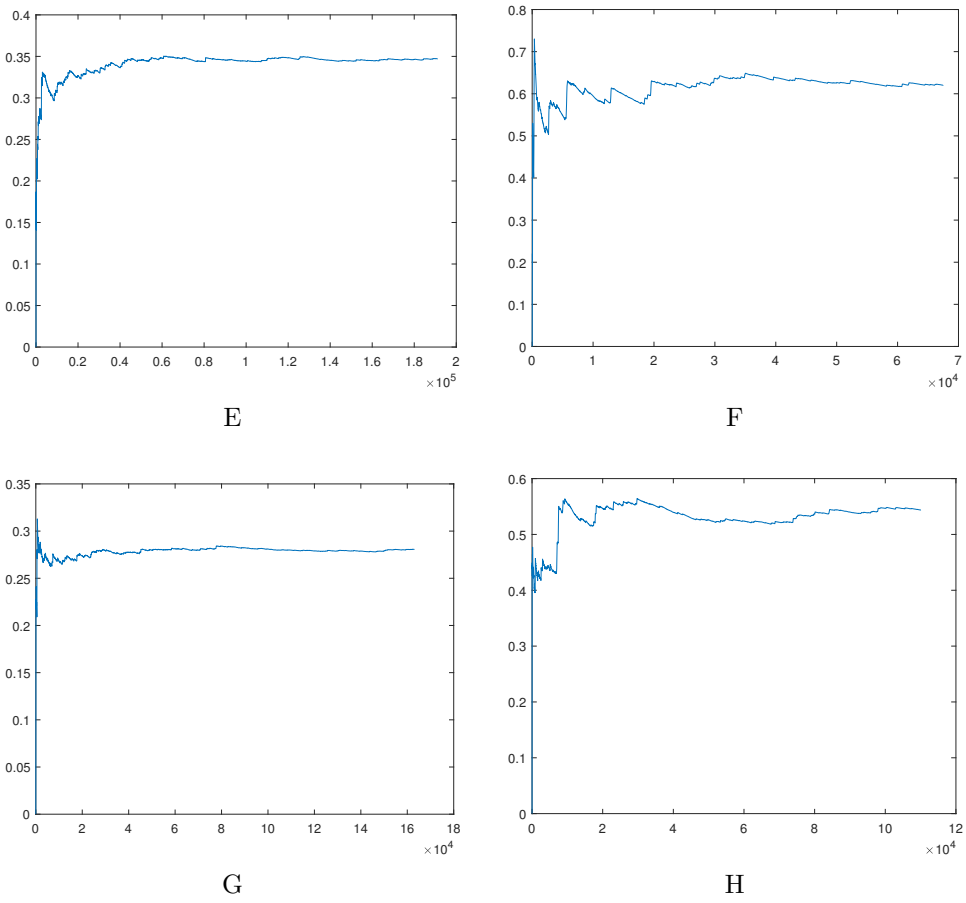


Figure 4.4.11: Plot of the standard deviation of t_0 for the last 10,000 iterations

ε which ensures that the approximation in (4.3) does not lead to a blow up in the algorithm's run time. For the above simulations we set $\varepsilon = 0.05$. For the case when such small time increments presented themselves in the simulation of the skeleton points, we made use of a diffusion approximation (conditioning only on one of the endpoints).

4.5 Extensions

The results reported here are restricted to simulated data, when the underlying dynamics are given by the Wright–Fisher diffusion. We are currently looking into applying the method to the horse coat colouration dataset in [LPR+09], such that we can compare our scheme's

performance vis-à-vis the other methods applied to this same dataset, thereby providing a verifiable benchmark which can be used to compare the efficacy of our approach. The dataset in [LPR$^+$09] however requires update procedures for the initial path segment that account for the cases when $t_c = t_2$ and $t_c = t_3$ for the MC1R and ASIP alleles. As detailed in Appendix A, the proposal mechanisms for such cases lead to significantly more intricate updating procedures in comparison to the straightforward setup adopted in the above simulation study.

In our analysis, we restricted ourselves to inferring the population rescaled selection parameter and allele age, i.e. σ and t_0, whilst assuming a constant demography for the population being considered. It would be desirable to incorporate the latter into our scheme as it would allow for the possibility to correctly decouple the influence of changes in effective population size on the corresponding allele trajectory from those induced by selection (see for instance [SES16] where the authors compare the result for when demography is, and is not accounted for). In essence this can be straightforwardly achieved by introducing a scaling function $\rho(t)$ which measures the size of the current population relative to the size at some reference point in time, say at the allele birth t_0, such that $\rho(t) = N(t)/N(t_0)$. The mathematical formulation and implementation used above can easily be amended to incorporate relatively simple changes in demography given by deterministic functions ρ.

The main drawback of the approach presented above is the fact that the initial path segment update relies on a pseudo-marginal step. This however is an unfortunate by-product of the updating procedure we use, namely the fact that we are as yet not able to jointly simulate a time increment t and an endpoint y from the bridge transition density $p_\sigma^{(\boldsymbol{\theta},x,t,z)}(y;s)$ (see (4.4) for the analogous neutral bridge density) of a Wright–Fisher diffusion. Although this might seem infeasible, we are currently looking into developing a technique which would allow us to simulate the allele age according to a distribution which is proportional to the intractable $a(t_2 - t_0, 0, x_2, \sigma)^{-1}$ appearing on the

RHS of (4.18), which would then enable us to construct an appropriate proposal mechanism in order to forego the pseudo-marginal update.

Throughout this chapter we have assumed that the mutation parameters were known a priori and thus fixed throughout. In reality this is seldom the case as often the mutation parameters for certain organisms are unknown and thus estimated using the parameters observed and documented for other species (see for instance [MMES12] where the authors use human mutation rates for horses). It thus would be natural to extend our framework to allow for the mutation parameters $\boldsymbol{\theta}$ to be jointly inferred with selection and allele age, however as is evident throughout Section 4.2, all the densities derived there are with respect to a product measure involving a neutral Wright–Fisher diffusion with fixed mutation parameter $\boldsymbol{\theta}$. To allow for varying mutation parameters, we would need to decouple this dependence in the dominating measure by finding a common reference measure independent of the mutation parameter $\boldsymbol{\theta}$, and with respect to which all relevant laws admit a density. This is relatively straightforward for the case when the mutation parameters are assumed to be greater than or equal to 1 (see the discussion just before and after equation (2.27)) as the corresponding laws of Wright–Fisher diffusions are mutually absolutely continuous with respect to one another and admit a Radon–Nikodym derivative which is readily available via Girsanov's theorem. However, when either mutation parameter is less than 1, the above is no longer true, and thus the above inferential scheme can only be extended for mutation parameters $\boldsymbol{\theta} \in [1, \infty)^2$. There is no real biological reason why this should be a priori true for any organism considered, and thus the method would be much more useful and realistic if one could instead allow for any strictly positive value of the mutation parameters. Note further that even when $\boldsymbol{\theta} \in [1, \infty)^2$, the resulting Radon–Nikodym derivatives will feature unbounded functions (from both above and below) in the exponent, such that the link with the Poisson point process is lost.

Whereas extending the methodology to allow for inferring muta-

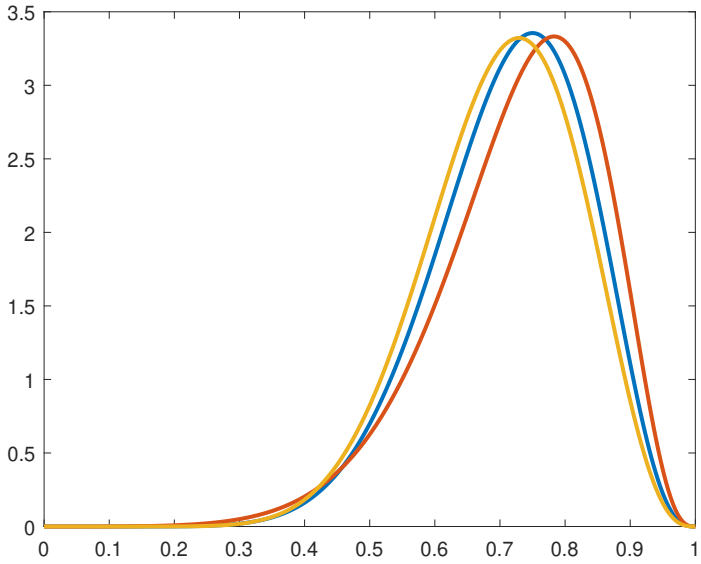


Figure 4.5.1: Plots of the invariant density for the scalar Wright-Fisher diffusion with parameter setups given in Table 4.5

tion would be beneficial from a biological and modelling perspective, one particular caveat presents itself: non-identifiability. By increasing the number of parameters to be inferred, one needs to keep in mind that the resulting method might not be able to distinguish between different parameter setups which produce very similar output. To illustrate this point, consider the following three cases:

Plot	σ	θ_1	θ_2
Red	0	10	4
Blue	10	4	4
Yellow	-4	12	4

Table 4.5.1: Three parameter setups for $(\sigma, \theta_1, \theta_2)$ used to illustrate the problem of non-identifiability once the analysis is extended to include the mutation rates.

The resulting invariant density plots for the Wright–Fisher diffusion are illustrated in Figure 4.5.1. Given just a dataset and without access to informative priors, any algorithm would struggle in deciding

which setup from the three above is most adequate, and the resulting procedure would most likely keep switching between the three configurations (as well as other configurations giving rise to similar density plots). Of course the above plots are only meant as an illustration of a potential issue as they depict the diffusion's invariant density rather than the transition density, but the fact that we are dealing with temporally spaced draws can at most only partially mitigate this non-identifiability. One possible way around this problem would be to obtain separate estimates or bounds on the distributions of the mutation parameters independently of the data collected, and subsequently reflect these estimates into the priors used for the MCMC implementation, however obtaining such estimates for aDNA data is not so straightforward and would require further investigation.

Lastly, as alluded to at the end of Subsection 4.4.2, if the allele age reaches far back in time, then it might be conceivable that the allele was present in the population for some time before it started being selected for. Thus the allele would be present in a proportion of the population which is distributed according to the invariant measure of the neutral diffusion when the selective switch occurs, enabling it to arise from standing genetic variation. This change (which might be induced for example by changes in the environment) is commonly referred to as a 'soft sweep' and is well studied in the biology literature, see for instance [HP05, PH06b, PH06a]. The setup presented above could be easily amended to allow for this phenomenon by changing the initial path dynamics. Instead of necessarily relying on de novo mutations to give rise to a selected allele, we would allow for the case when alleles rise from standing variation by sampling the initial value of the diffusion at the time at which the selective pressures shifted in its favour from the invariant measure of the diffusion. One way to allow both de novo mutations as well as soft sweeps is to adopt $\Delta \cup [t_c - \delta, t_c - \varepsilon]$ as potential values for the quantity t_0, such that if $t_0 \in [t_c - \delta, t_c - \varepsilon]$, then the allele arose as a de novo mutation (and therefore $X_{t_0} = 0$ and we proceed as described in Section 4.3, with t_0 retaining its interpretation as the time at which the allele was

born), whilst if $t_0 \in \Delta$, then the allele rose from standing variation (and we model X_{t_0} as a draw from the invariant density $f_{\boldsymbol{\vartheta}}$ given in (2.22)). One would then be interested in estimating both the selection coefficient and the time t_0 at which the switch from neutral to selected happened (note that we can no longer treat t_0 as the birth time of the allele), as this might help shed light on to why the change in selectiveness occurred if other information related to the organism is available. We point out however that the method would still require a pseudo-marginal step for the initial path segment update.

Chapter 5

Extending $\boldsymbol{\vartheta}$-uniform ergodicity to the Multidimensional Wright–Fisher Diffusion

Thus far we have concerned ourselves solely with the one-dimensional Wright–Fisher diffusion which allows for the frequency of a single allele to be tracked over time. This scalar setting is particularly neat as a number of properties of one-dimensional diffusions (such as point recurrence and the notions of speed and scale) allow for the calculations involved to be significantly simplified. In this chapter we present a brief overview of how the results derived in Chapter 2 can be extended for the K-dimensional Wright–Fisher diffusion, in particular we provide a rough sketch of the first steps towards proving $\boldsymbol{\vartheta}$-uniform ergodicity (as given in Definition 2.1). This property is a key ingredient in establishing uniform local asymptotic normality for the family of laws induced by the solutions to the associated SDEs, which in turn unlocks the door to deriving results similar to those in Chapter 3 for the ML and Bayesian estimators.

The main challenge now is that it is no longer straightforward to decompose integrals similar to those on the LHS of (2.3) by making

use of the hitting times of sets (as opposed to points as done in the scalar case). Whilst the hitting time of a set is well-defined and can be controlled sufficiently well via Lyapunov functions (see Subsection 5.3.2), it does not offer the correct framework within which the diffusion path can be split into i.i.d. chunks as in the scalar case. Instead, we show how Nummelin splitting can be applied in the continuous time setting (via the construction first presented in [LL08]), to assemble a richer stochastic process which allows for a set of suitable stopping times to be defined. These times (referred to as regeneration times) then provide the natural generalisation for the hitting times used in Theorem 2.2, and thus allow us to perform the necessary path decomposition in such a way that the LHS of (5.8) (which is the multidimensional analogue of (2.3)) can be bounded from above by the moments of these regeneration times. The above detailed technique was successfully employed by [LL13] for a general class of Harris recurrent processes with general state space in their Theorem 5.2, where one can explicitly control the rate of convergence in the ergodic theorem in terms of the initial starting point $\mathbf{x}$ via a suitable Lyapunov function. The key to achieving this final bound in terms of Lyapunov functions is Theorem 4.1 therein, where results from [DFMS04, DFG09] for the modulated moments of the resolvent chain under the drift condition implied by Assumption 2.2 in [LL13] (which we reproduce below for reference and also guarantees Harris recurrence) allow for the moments of the regeneration times to be bounded from above.

Assumption 5.1 (Assumption 2.2 in [LL13])**.** *There exists a closed petite set B, a continuous function $V : \Delta_K \mapsto [1, \infty)$, an increasing differentiable concave positive function $\Phi : [1, \infty) \mapsto (0, \infty)$, and constant $b < \infty$ such that $\forall s \geq 0$*

$$\begin{aligned} &\mathbb{E}^{(\boldsymbol{\vartheta})}_{\mathbf{X},\mathbf{x}}\left[V(\mathbf{X}_s)\right] + \mathbb{E}^{(\boldsymbol{\vartheta})}_{\mathbf{X},\mathbf{x}}\left[\int_0^s \Phi \circ V(\mathbf{X}_u)du\right] \\ &\qquad \leq V(\mathbf{x}) + b\mathbb{E}^{(\boldsymbol{\vartheta})}_{\mathbf{X},\mathbf{x}}\left[\int_0^s 1_B(\mathbf{X}_u)du\right]. \end{aligned} \tag{5.1}$$

We further present bound (2.5) in [LL13] which can be shown to imply

the above assumption:

$$\mathscr{G}V(\mathbf{x}) \leq -\Phi \circ V(\mathbf{x}) + b1_B(\mathbf{x}). \tag{5.2}$$

Verifying Assumption 5.1 (or indeed (5.2)) for the K-dimensional Wright–Fisher diffusion is non-trivial and remains an open problem. We note that it might indeed be the case that these conditions are not achievable for the Wright–Fisher case much in the same way that standard drift conditions for scalar diffusions in the literature fail for the analogous Wright–Fisher diffusion. For instance, Assumptions 5.1 and 5.2 in [LLL11] guarantee the existence of (and allow for bounds on) the moments of hitting times of scalar diffusions, but they immediately fail to hold for the scalar Wright–Fisher diffusion as the diffusion coefficient is zero at the boundaries. Nonetheless, as we have seen in Corollaries 2.5 and 2.6, these quantities can be bounded through alternative techniques, which however do not extend for the multidimensional case.

In light of the above, whilst we pursue the approach presented in [LL13] to decompose the the LHS of (5.8), we stop one step short and present all bounds in terms of the moments of the regeneration times. At the end of this chapter, we point out how the moments of hitting times of a specific set for the original diffusion can be controlled via standard multidimensional diffusion process theory and a suitable choice of Lyapunov function (for a certain set of parameter configurations) by means of Theorem 3.9 in [Kha12]. Whilst this does not necessarily offer a direct way with which to tackle the above mentioned problem, it does offer some insight in the way of choosing the correct Lyapunov functions and serves to highlight particular features pertinent to the specific case of the Wright–Fisher diffusion, thereby providing some insight into how a result similar to Theorem 4.1 in [LL13] might be derived for the Wright–Fisher case.

The rest of this chapter is organised as follows: In Section 5.1 we give an overview of the K-dimensional Wright–Fisher diffusion,

illustrating several known properties of the process. We then introduce the necessary notation and quantities to construct the richer stochastic process via Nummelin splitting in Section 5.2, and use it to define the regeneration times. We subsequently outline how these allow for the diffusion to be split up into i.i.d. segments such that the LHS of (5.8) can be bounded from above by the moments of these regeneration times. We end the chapter by focusing on how to handle these moments of the regeneration times, giving an overview of the approach adopted in Theorem 4.1 in [LL13] and how it relies on Assumption 2.2 therein, before showing how standard multidimensional theory and Lyapunov functions allow us to control the hitting times of sets for the Wright–Fisher diffusion.

5.1 The K-dimensional Wright–Fisher diffusion

The K-dimensional Wright–Fisher diffusion is defined on the $K-1$ dimensional simplex $\Delta_K := \{\mathbf{x} \in [0,1]^K : \sum_{i=1}^K x_i = 1\}$, with generator $\mathscr{G}$ given by

$$\begin{aligned}\mathscr{G} :=& \frac{1}{2}\sum_{i,j=1}^{K} x_i\left(\delta_{ij} - x_j\right)\frac{\partial^2}{\partial x_i \partial x_j} \\ &+ \sum_{i=1}^{K}\left(\sum_{j=1}^{K}\gamma_{ji}x_j + x_i\left(\sum_{j=1}^{K}\sigma_{ij}x_j - \sum_{k,l=1}^{K}\sigma_{kl}x_k x_l\right)\right)\frac{\partial}{\partial x_i} \end{aligned} \tag{5.3}$$

where $(\gamma_{ij})_{i,j=1}^K$ is the infinitesimal matrix of mutation parameters, and $(\sigma_{ij})_{i,j=1}^K$ is a real symmetric matrix describing the selection parameters. Note in particular that the above framework allows for both haploid ($\sigma_{ij} = \sigma_i + \sigma_j$), as well as diploid selection patterns. The domain of $\mathscr{G}$ is the space of all twice continuously differentiable functions defined on the $K-1$-dimensional simplex with boundary behaviour depending on the values of the mutation matrix $(\gamma_{ij})_{i,j=1}^K$. We point out further that one could express the multidimensional diffusion $\mathbf{X}^T$ as the solution to a corresponding multidimensional SDE, however such a formulation is rather cumbersome and offers little in the way of insight or techniques in comparison to the generator

approach adopted here.

If one assumes parent independent mutation (PIM), that is when the mutation parameter is independent of the parent's type, leading to $\gamma_{ij} = 2^{-1}\theta_j$ for $\theta_j > 0$, then we have that the multidimensional non-neutral Wright–Fisher diffusion admits a stationary distribution (see [Wri49, BEG00]), whose density is given

$$f_{\boldsymbol{\vartheta}}(\mathbf{x}) = \frac{1}{G_{\boldsymbol{\vartheta}}} e^{\sum_{i,j=1}^{K} \sigma_{ij} x_i x_j} \prod_{i=1}^{K} x_i^{\theta_i - 1}, \tag{5.4}$$

where $G_{\boldsymbol{\vartheta}}$ is a normalising constant such that

$$G_{\boldsymbol{\vartheta}} = \int_{\Delta_K} e^{\sum_{i,j=1}^{K} \sigma_{ij} x_i x_j} \prod_{i=1}^{K} x_i^{\theta_i - 1} d\mathbf{x}.$$

We point out the slight abuse of notation above, as we have redefined our parameter of interest $\boldsymbol{\vartheta}$, invariant density $f_{\boldsymbol{\vartheta}}$ and normalising constant $G_{\boldsymbol{\vartheta}}$. In particular $\boldsymbol{\vartheta}$ is now set to be both the selection matrix and mutation vector, i.e. $\boldsymbol{\vartheta} = (\boldsymbol{\sigma}, \boldsymbol{\theta}) = (\{(\sigma_{ij})_{i,j=1}^{K}, (\theta_j)_{j=1}^{K}\})$ (assuming PIM), in contrast to the selection coefficient and mutation parameters ($\boldsymbol{\vartheta} = (\sigma, \theta_1, \theta_2)$) used in Chapter 2. Note that from now on we shall always assume PIM, such that whenever we say mutation it is understood that we are referring to PIM, unless otherwise specified.

Denote by $p_{\mathbf{0}}^{\boldsymbol{\theta}}(t, \mathbf{x}, \mathbf{y})$ the transition density associated with the neutral multidimensional Wright–Fisher diffusion having generator (5.3) (i.e. with $\sigma_{ij} = 0$ for all $i, j \in \{1, \ldots, K\}$), and by $p_{\boldsymbol{\sigma}}^{\boldsymbol{\theta}}(t, \mathbf{x}, \mathbf{y})$ the corresponding non-neutral transition density (note that the bold subscripts are used to differentiate from the scalar transition densities introduced in Chapter 2). Then we have the following transition

density expansion for the neutral case (see [Shi76, Gri79, GL83, Tav84])

$$p_{\mathbf{0}}^{\boldsymbol{\theta}}(t,\mathbf{x},\mathbf{y}) = \sum_{m=0}^{\infty}\sum_{k=m}^{\infty}(-1)^{k-m}\frac{|\boldsymbol{\theta}|+2k-1}{m!(k-m)!}\frac{\Gamma(|\boldsymbol{\theta}|+k+m-1)}{\Gamma(|\boldsymbol{\theta}|+m)} \times e^{-\frac{k(k+|\boldsymbol{\theta}|-1)t}{2}}\sum_{\substack{\mathbf{l}\\|\mathbf{l}|=m}}\mathcal{M}_{m,\mathbf{x}}(\mathbf{l})\mathcal{D}_{\boldsymbol{\theta}+\mathbf{l}}(\mathbf{y}), \tag{5.5}$$

where $|\cdot|$ here denotes the L^1-norm, and $\mathcal{M}_{m,\mathbf{x}}(\cdot)$ denotes the probability mass function of a multinomial distribution with parameters m and $\mathbf{x}$. We can derive a decomposition for the non-neutral transition density much in the same way as done in (4.9) by conditioning and using Girsanov's theorem. Let $\mathbb{WF}_{\mathbf{0},\boldsymbol{\theta}}^{(\mathbf{x})}$ and $\mathbb{WF}_{\mathbf{0},\boldsymbol{\theta}}^{(t,\mathbf{x},\mathbf{y})}$ denote the law of a K-dimensional neutral Wright–Fisher diffusion started at $\mathbf{x}$ and with mutation parameter set to the vector $\boldsymbol{\theta}$, and the law of a K-dimensional neutral Wright–Fisher diffusion bridge started at $\mathbf{x}$ and ending at $\mathbf{y}$ in time t respectively. Further let $\mathbb{WF}_{\boldsymbol{\sigma},\boldsymbol{\theta}}^{(\mathbf{x})}$ and $\mathbb{WF}_{\boldsymbol{\sigma},\boldsymbol{\theta}}^{(t,\mathbf{x},\mathbf{y})}$ denote the laws of the corresponding non-neutral processes. Then by conditioning on the endpoint $\mathbf{X}_t = \mathbf{y}$ and re-arranging we get that

$$\frac{d\mathbb{WF}_{\boldsymbol{\sigma},\boldsymbol{\theta}}^{(t,\mathbf{x},\mathbf{y})}}{d\mathbb{WF}_{\mathbf{0},\boldsymbol{\theta}}^{(t,\mathbf{x},\mathbf{y})}}(\mathbf{X}^t) = \frac{p_{\mathbf{0}}^{\boldsymbol{\theta}}(t,\mathbf{x},\mathbf{y})}{p_{\boldsymbol{\sigma}}^{\boldsymbol{\theta}}(t,\mathbf{x},\mathbf{y})}\frac{d\mathbb{WF}_{\boldsymbol{\sigma},\boldsymbol{\theta}}^{(\mathbf{x})}}{d\mathbb{WF}_{\mathbf{0},\boldsymbol{\theta}}^{(\mathbf{x})}}(\mathbf{X}^t), \tag{5.6}$$

and the multidimensional Girsanov theorem gives us that

$$\begin{aligned}\frac{d\mathbb{WF}_{\boldsymbol{\sigma},\boldsymbol{\theta}}^{(\mathbf{x})}}{d\mathbb{WF}_{\mathbf{0},\boldsymbol{\theta}}^{(\mathbf{x})}}(\mathbf{X}^t) &= \exp\left\{\int_0^t (\boldsymbol{\mu}_{\boldsymbol{\sigma}}-\boldsymbol{\mu}_{\mathbf{0}})^T\boldsymbol{\Sigma}^{-1}(\mathbf{X}_t)\,d\mathbf{X}_t \right.\\ &\qquad \left. -\frac{1}{2}\int_0^t \left(\boldsymbol{\mu}_{\boldsymbol{\sigma}}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_{\boldsymbol{\sigma}}-\boldsymbol{\mu}_{\mathbf{0}}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_{\mathbf{0}}\right)(\mathbf{X}_t)\,dt\right\}\\ &= \exp\left\{\int_0^t (\boldsymbol{\mu}_{\boldsymbol{\sigma}}-\boldsymbol{\mu}_{\mathbf{0}})^T\boldsymbol{\Sigma}^{-1/2}(\mathbf{X}_t)\,d\mathbf{W}_t \right.\\ &\qquad \left. -\frac{1}{2}\int_0^t (\boldsymbol{\mu}_{\boldsymbol{\sigma}}-\boldsymbol{\mu}_{\mathbf{0}})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_{\boldsymbol{\sigma}}-\boldsymbol{\mu}_{\mathbf{0}})(\mathbf{X}_t)\,dt\right\},\end{aligned}$$

when $\boldsymbol{\mu_0}$ is the true parameter with $\mathbf{WF}_{\mathbf{0},\boldsymbol{\theta}}^{(mathbfx)}$-probability 1, where we define

$$\boldsymbol{\mu}_{\boldsymbol{\sigma}} := \begin{pmatrix} \frac{\theta_1}{2} + x_1\left(\sum_{j=1}^K \sigma_{1j}x_j - \sum_{k,l=1}^K \sigma_{kl}x_k x_l\right) \\ \frac{\theta_2}{2} + x_2\left(\sum_{j=1}^K \sigma_{2j}x_j - \sum_{k,l=1}^K \sigma_{kl}x_k x_l\right) \\ \vdots \\ \frac{\theta_K}{2} + x_K\left(\sum_{j=1}^K \sigma_{Kj}x_j - \sum_{k,l=1}^K \sigma_{kl}x_k x_l\right) \end{pmatrix}, \boldsymbol{\mu_0} := \frac{1}{2}\begin{pmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_K \end{pmatrix},$$

$$\boldsymbol{\Sigma} := \begin{pmatrix} x_1(1-x_1) & -x_1x_2 & \dots & \dots & -x_1x_K \\ -x_2x_1 & x_2(1-x_2) & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & x_{K-1}(1-x_{K-1}) & -x_{K-1}x_K \\ -x_Kx_1 & \dots & \dots & -x_{K-1}x_K & x_K(1-x_K) \end{pmatrix}.$$

and take $\boldsymbol{\mu}^T$ to mean the transpose of $\boldsymbol{\mu}$. Taking expectations with respect to $\mathbb{WF}_{\mathbf{0},\boldsymbol{\theta}}^{(t,\mathbf{x},\mathbf{y})}$ on both sides of (5.6), and re-arranging the resulting terms gives

$$p_{\boldsymbol{\sigma}}^{\boldsymbol{\theta}}(t,\mathbf{x},\mathbf{y}) = p_{\mathbf{0}}^{\boldsymbol{\theta}}(t,\mathbf{x},\mathbf{y})\mathbb{E}_{\mathbb{WF}_{\mathbf{0},\boldsymbol{\theta}}^{(t,\mathbf{x},\mathbf{y})}}\left[\frac{d\mathbb{WF}_{\boldsymbol{\sigma},\boldsymbol{\theta}}^{(\mathbf{x})}}{d\mathbb{WF}_{\mathbf{0},\boldsymbol{\theta}}^{(\mathbf{x})}}(\mathbf{X}^t)\right]. \tag{5.7}$$

As the $K-1$ dimensional simplex is a closed subset of the compact set $[0,1]^K$, the multidimensional process also lives in a compact state space, and thus it is natural to expect the concepts of $\boldsymbol{\vartheta}$-uniform ergodicity and local asymptotic normality introduced in Chapter 2 to hold. The catch is of course that now the calculations will be more involved and proofs cannot rely on either point recurrence nor on the notions of speed and scale. In this chapter, we set out to outline how one might go about proving uniform in the selection and mutation parameters ergodicity of the Wright–Fisher diffusion in the multidimensional case, which would subsequently be useful when proving uniform local asymptotic normality, as well as generalising the results obtained in Chapter 3 for this setting. To this end, recall that $\boldsymbol{\vartheta}$-uniform ergodicity in the context of the K-dimensional Wright–Fisher diffusion means showing

that $\forall \varepsilon > 0$,

$$\lim_{T\to\infty} \sup_{\boldsymbol{\vartheta}\in\mathcal{K}} \mathbb{P}^{(\boldsymbol{\vartheta})}_{\mathbf{X},\nu} \left[\left| \frac{1}{T} \int_0^T h(\mathbf{X}_t)dt - \mathbb{E}^{(\boldsymbol{\vartheta})}_{\mathbf{X}} \left[h(\boldsymbol{\xi}) \right] \right| > \varepsilon \right] = 0 \tag{5.8}$$

holds for any bounded measurable function $h : \Delta_K \mapsto \mathbb{R}$, and any compact $\mathcal{K} \subset \boldsymbol{\Theta}$, where $\mathbb{P}^{(\boldsymbol{\vartheta})}_{\mathbf{X},\nu}$ denotes the measure induced on the space of continuous functions mapping the interval $[0, T]$ to Δ_K, $C_T(\Delta_K)$, by the process $\mathbf{X}$ corresponding to the generator (5.3) when the selection and mutation parameters are set to $\boldsymbol{\vartheta} = (\boldsymbol{\sigma}, \boldsymbol{\theta})$ and the process is started from the initial measure ν. $\mathbb{E}^{(\boldsymbol{\vartheta})}_{\mathbf{X}}$ then denotes taking expectation with respect to the invariant density given in (5.4).

5.2 Regeneration times & speed of convergence

We now introduce the notation present in Section 3 of [LL13], suitably translated to the setting of the multidimensional Wright–Fisher diffusion. To this end, we set $\mathbf{X} := (\mathbf{X}_t)_{t\geq 0}$ to be the process associated to the generator (5.3), $(\mathcal{F}_t)_{t\geq 0}$ the filtration generated by $\mathbf{X}$, and $p^{\boldsymbol{\theta}}_{\boldsymbol{\sigma}}(t, \mathbf{x}, \mathbf{y})$ the associated transition density of going from point $\mathbf{x}$ to $\mathbf{y}$ in time t when the selection matrix is set to $\boldsymbol{\sigma}$ and mutation parameters are set to $\boldsymbol{\theta}$. Furthermore, let $\mathbb{P}^{(\boldsymbol{\vartheta})}_{\mathbf{X},\mathbf{x}}$ denote the law of $\mathbf{X}$ when started from $\mathbf{x}$. Define the resolvent kernel as

$$U^1(\mathbf{x}, d\mathbf{y}) := \int_0^\infty e^{-t} p^{\boldsymbol{\theta}}_{\boldsymbol{\sigma}}(t, \mathbf{x}, d\mathbf{y})dt,$$

(note that in our case Assumption 2.1 in [LL13] follows immediately using Lebesgue measure as the reference measure), and observe that $U^1(\mathbf{x}, d\mathbf{y})$ is the one step transition kernel associated with the resolvent chain of the original process given by $(\mathbf{X}_{T_n})_{n\in\mathbb{N}}$, where $(T_n - T_{n-1})_{n\in\mathbb{N}}$ is a collection of i.i.d. Exp(1) waiting times. Furthermore, $\mathbf{X}$ being ergodic ([EK98, Theorem 5.5]) with invariant density $f_{\boldsymbol{\vartheta}}$ implies (by Proposition 6.7 in [HL03]) that we can find a compact set $C \subset \Delta_K$ such that $f_{\boldsymbol{\vartheta}}(C) := \int_C f_{\boldsymbol{\vartheta}}(\mathbf{x})d\mathbf{x} > 0$ and

$$U^1(\mathbf{x}, d\mathbf{y}) \geq \alpha 1_C(\mathbf{x})\lambda(d\mathbf{y}), \tag{5.9}$$

for some $\alpha \in (0,1]$, and $\lambda(\cdot) = f_{\boldsymbol{\vartheta}}(\cdot \cap C)$. Note that by restricting our attention to exponential time increments as above, we are able to shift from continuous to discrete time, which coupled with the Doeblin type condition (5.9), allows us to apply a Nummelin splitting to the resolvent chain. For two points $\mathbf{x}, \mathbf{y} \in \Delta_K$, we set

$$u^1(\mathbf{x}, \mathbf{y}) := \int_0^\infty e^{-t} p^{\boldsymbol{\theta}}_{\boldsymbol{\sigma}}(t, \mathbf{x}, \mathbf{y}) dt,$$

and define the splitting kernel $Q((\mathbf{x}, u), d\mathbf{y})$ mapping $\Delta_K \times [0,1]$ to Δ_K as follows

$$Q((\mathbf{x}, u), d\mathbf{y}) = \begin{cases} \lambda(d\mathbf{y}) & \text{if } (\mathbf{x}, u) \in C \times [0, \alpha] \\ \frac{U^1(\mathbf{x}, d\mathbf{y}) - \alpha\lambda(d\mathbf{y})}{1-\alpha} & \text{if } (\mathbf{x}, u) \in C \times (\alpha, 1] \\ U^1(\mathbf{x}, d\mathbf{y}) & x \notin C. \end{cases} \tag{5.10}$$

Observe that integrating out u in $Q((\mathbf{x}, u), d\mathbf{y})$ recovers the resolvent kernel $U^1(\mathbf{x}, d\mathbf{y})$. We now make use of the above to construct the enriched stochastic process $\mathbf{Z} := (\mathbf{Z}_t)_{t \geq 0}$ taking values in $\Delta_K \times [0,1] \times \Delta_K$.

Starting from $\mathbf{Z}_0 = (\mathbf{X}_0, u_0, \mathbf{x}_1)$, where $\mathbf{X}_0 = \mathbf{x}_0$ is the initial value of $\mathbf{X}$, $u_0 \in [0,1]$ and $\mathbf{x}_1 \in \Delta_K$, set $T_0 = 0$, $n = 1$ and proceed as follows:

1. Simulate the new jump time

$$\tau_n \sim e^{-t} \frac{p^{\boldsymbol{\theta}}_{\boldsymbol{\sigma}}(t, \mathbf{x}_{n-1}, \mathbf{x}_n)}{u^1(\mathbf{x}_{n-1}, \mathbf{x}_n)} dt$$

   and set $T_n := T_{n-1} + \tau_n$.

2. Set $Z^2_{T_{n-1}+s} = u_{n-1}$, and $\mathbf{Z}^3_{T_{n-1}+s} = \mathbf{x}_n$, $\forall s \in [0, \tau_n)$.

3. Fill in the path $(\mathbf{Z}^1_v)_{v=T_{n-1}}^{T_n}$ via a non-neutral Wright–Fisher bridge going from $\mathbf{x}_{n-1}$ to $\mathbf{x}_n$ in time τ_n, i.e. $\forall s < \tau_n$ sample

$$\mathbf{Z}^1_{T_{n-1}+s} \sim \frac{p^{\boldsymbol{\theta}}_{\boldsymbol{\sigma}}(s, \mathbf{x}_{n-1}, \mathbf{y}) p^{\boldsymbol{\theta}}_{\boldsymbol{\sigma}}(t-s, \mathbf{y}, \mathbf{x}_n)}{p^{\boldsymbol{\theta}}_{\boldsymbol{\sigma}}(t, \mathbf{x}_{n-1}, \mathbf{x}_n)} d\mathbf{y}.$$

4. At T_n, take $\mathbf{Z}^1_{T_n} = \mathbf{Z}^3_{T_{n-1}} = \mathbf{x}_n$, $Z^2_{T_n} \sim \text{Unif}([0,1])$ independently

of $\mathbf{Z}_s$ with $s < T_n$, and choose $\mathbf{Z}^3_{T_n} \sim Q((\mathbf{x}_n, Z^2_{T_n}), d\mathbf{y})$.

5. Set $\mathbf{Z}_{T_n} = (\mathbf{Z}^1_{T_n}, Z^2_{T_n}, \mathbf{Z}^3_{T_n})$, increment n by 1 and go back to step 1.

Note that the above construction is such that the first co-ordinate of the process $(\mathbf{Z}_t)_{t\geq 0}$ obeys the dynamics of the original process via the use of the bridge measures in step 3. The second co-ordinate is an independent uniform random variate which tells us how to simulate the next value of the resolvent chain via (5.10). We store the latter in the third co-ordinate of $(\mathbf{Z}_t)_{t\geq 0}$. In particular, whenever we enter C and the coin flip returns heads, we regenerate the path by drawing the next bridge end point from the minorising measure λ.

We point out that in contrast to the general construction in [LL08], we have omitted the case $p^{\boldsymbol{\theta}}_{\boldsymbol{\sigma}}(t, \mathbf{x}, \mathbf{y}) = 0$, as the non-neutral transition density is always strictly positive in view of the process having invariant density (5.4), for which $f_{\boldsymbol{\vartheta}}(\mathbf{y}) > 0$ for $\mathbf{y} \in \Delta_K \setminus \partial\Delta_K$ (i.e. $\mathbf{y}$ in the interior of Δ_K). On the boundary $\partial\Delta_K$, the invariant density $f_{\boldsymbol{\vartheta}}$ is equal to 0 (if the corresponding mutation parameter is greater than or equal to 1, and in this case the boundary is inaccessible) or it tends to ∞ on approach of the boundary (if the mutation parameter is less than 1). The strict positivity of the transition density also follows directly from (5.7) - the only way the RHS can be zero is if $p^{\boldsymbol{\theta}}_{\mathbf{0}}(t, \mathbf{x}, \mathbf{y}) = 0$, which fails to hold for any $\mathbf{x} \in \Delta_K$ and $\mathbf{y} \in \Delta_K \setminus \partial\Delta_K$ in view of the expansion (5.5).

If we denote by $\mathbb{P}^{(\boldsymbol{\vartheta})}_{\mathbf{Z},\mathbf{x}}$ the law of $\mathbf{Z}$ when started from the initial measure $\delta_{\mathbf{x}} \otimes \mathrm{Unif}_{[0,1]}(du) \otimes Q((\mathbf{x}, u), d\mathbf{y})$, then the above construction ensures that $\mathscr{L}((\mathbf{Z}^1_t)_{t\geq 0} | \mathbf{Z}^1_0 = \mathbf{x}) = \mathscr{L}((\mathbf{X}_t)_{t\geq 0} | \mathbf{X}_0 = \mathbf{x})$ (see Proposition 2.8 (c) in [LL08]). We further denote by $\mathbb{E}^{(\boldsymbol{\vartheta})}_{\mathbf{Z},\mathbf{x}}$ taking expectation with respect to $\mathbb{P}^{(\boldsymbol{\vartheta})}_{\mathbf{Z},\mathbf{x}}$, and note that although $\mathbf{Z}$ retains the Markov property with respect to the filtration it generates (see Theorem 2.7 in [LL08]), it is not in general strong Markov, but by construction it is strong Markov for the times T_n. The additional variates $(Z^2_t, \mathbf{Z}^3_t)_{t\geq 0}$ allow for a sequence of regeneration times of $\mathbf{Z}$ to be introduced as

follows. Set $A := C \times [0, \alpha] \times \Delta_K$, $S_0 = R_0 = 0$, and define

$$\begin{aligned} S_{n+1} &:= \inf \{T_m > R_n : \mathbf{Z}_{T_m} \in A\} \\ R_{n+1} &:= \inf \{T_m : T_m > S_{n+1}\} \end{aligned} \tag{5.11}$$

for $n \in \mathbb{N}$. The sequence $(R_n)_{n\in\mathbb{N}}$ is the natural generalisation for the corresponding quantities we used in Theorem 2.2:

1. Under $\mathbb{P}^{(\boldsymbol{\vartheta})}_{\mathbf{Z},\mathbf{x}}$ the sequence of jump times $(T_n)_{n\in\mathbb{N}}$ is independent of $(\mathbf{Z}^1_t)_{t\geq 0}$, with $(T_n - T_{n-1})_{n\in\mathbb{N}\setminus\{0\}} \sim_{i.i.d.} \text{Exp}(1)$ for $n \in \mathbb{N}$.

2. At each R_n we have that $\mathbf{Z}_{R_n} \sim \lambda(d\mathbf{x}) \otimes \text{Unif}_{[0,1]}(du) \otimes Q((\mathbf{x}, u), d\mathbf{y})$ $\forall n \geq 1$, which ensures that $\mathbf{Z}$ is regenerated at these times and implies that $\mathbf{Z}_{R_n+\cdot}$ is independent of $\mathscr{F}_{S_n-}$ (which we recall is the filtration generated by the original process $\mathbf{X}$).

3. The sequence $(\mathbf{Z}_{R_n})_{n\geq 1}$ are i.i.d.

All three statements follow immediately from steps 1 and 4 in the construction of $\mathbf{Z}$, coupled with the definition of R_n and S_n (formal proofs can be found in Propositions 2.6 and 2.13 in [LL08]).

The process $\mathbf{Z}$ admits an invariant distribution which we denote by $\mathbb{P}^{(\boldsymbol{\vartheta})}_{\mathbf{Z}}$, and whose projection onto the first co-ordinate is denoted $\mathbb{P}^{(\boldsymbol{\vartheta})}_{\mathbf{Z}^1}$. The latter measure is such that for any measurable function h integrable with respect to $\mathbb{P}^{(\boldsymbol{\vartheta})}_{\mathbf{Z}^1}$, we have that

$$\begin{aligned} \mathbb{E}^{(\boldsymbol{\vartheta})}_{\mathbf{Z}^1}[h(\boldsymbol{\eta})] &= \ell_{\boldsymbol{\vartheta}}\, \mathbb{E}^{(\boldsymbol{\vartheta})}_{\mathbf{Z}^1}\left[\int_{R_1}^{R_2} h(\mathbf{Z}^1_s) ds\right], \\ \ell_{\boldsymbol{\vartheta}} &= \mathbb{E}^{(\boldsymbol{\vartheta})}[R_2 - R_1]^{-1} = \mathbb{E}^{(\boldsymbol{\vartheta})}_{\mathbf{Z},\nu}[R_1]^{-1}, \end{aligned} \tag{5.12}$$

where $\mathbb{E}^{(\boldsymbol{\vartheta})}_{\mathbf{Z}^1}$ denotes taking expectation with respect to the measure $\mathbb{P}^{(\boldsymbol{\vartheta})}_{\mathbf{Z}^1}$, and $\boldsymbol{\eta} \sim \mathbb{P}^{(\boldsymbol{\vartheta})}_{\mathbf{Z}^1}$ (see Proposition 3.4 in [LL13]). We point out that the superscript in the expectation defining $\ell_{\boldsymbol{\vartheta}}$ is included to make explicit the dependence of these terms on the underlying parameter $\boldsymbol{\vartheta}$. Furthermore, $\mathbb{P}^{(\boldsymbol{\vartheta})}_{\mathbf{Z}^1}$ coincides with the invariant measure of the original

process $\mathbf{X}$ (by construction), thus allowing us to deduce that

$$\mathbb{E}_{\mathbf{X}}^{(\boldsymbol{\vartheta})}\left[h(\boldsymbol{\xi})\right] = \mathbb{E}_{\mathbf{Z}^1}^{(\boldsymbol{\vartheta})}\left[h(\boldsymbol{\eta})\right] = \ell_{\boldsymbol{\vartheta}}\,\mathbb{E}_{\mathbf{Z}^1}^{(\boldsymbol{\vartheta})}\left[\int_{R_1}^{R_2} h(\mathbf{Z}_s^1)ds\right].$$

The above allows us to re-write

$$\begin{aligned}&\mathbb{P}_{\mathbf{X},\nu}^{(\boldsymbol{\vartheta})}\left[\left|\frac{1}{T}\int_0^T h(\mathbf{X}_s)ds - \mathbb{E}_{\mathbf{X}}^{(\boldsymbol{\vartheta})}\left[h(\boldsymbol{\xi})\right]\right| > \varepsilon\right]\\ &\quad= \mathbb{P}_{\mathbf{Z},\nu}^{(\boldsymbol{\vartheta})}\left[\left|\frac{1}{T}\int_0^T h(\mathbf{Z}_s^1)ds - \ell_{\boldsymbol{\vartheta}}\,\mathbb{E}_{\mathbf{Z}^1}^{(\boldsymbol{\vartheta})}\left[\int_{R_1}^{R_2} h(\mathbf{Z}_s^1)ds\right]\right| > \varepsilon\right], \quad (5.13)\end{aligned}$$

where $\mathbb{P}_{\mathbf{Z},\nu}^{(\boldsymbol{\vartheta})}$ here denotes the law of $\mathbf{Z}$ when the first co-ordinate $\mathbf{Z}^1$ is started from the distribution ν. The final quantity we shall require prior to tackling the RHS of (5.13), is the counting process $(N_t)_{t\geq 0}$ defined by

$$N_t := \sup\left\{n : R_n \leq t\right\}, N_0 := 0.$$

We can now deal with the RHS of (5.13) in the same way that the corresponding probability in Theorem 2.2 was bounded from above using Theorem 3.2 in [LLL11]. We emphasise that rather than apply Theorem 5.2 in [LL13] for $p \in \{2, 3, \dots\}$ directly, we use the arguments there to explicitly track the dependence of the resulting moment bounds in terms of the parameter of interest $\boldsymbol{\vartheta}$, and thus briefly summarise the arguments used there in what follows. Note further that in Theorem 5.2 in [LL13], the authors are interested in controlling the rate of convergence in terms of the initial starting point $\mathbf{x}$, and thus first bound the relevant probability in terms of the moments of the regeneration times, and subsequently use Theorem 4.1 therein to bound these moments by a suitable Lyapunov function. In our setting it might not necessarily be the case that Theorem 4.1 is applicable (in view of Assumption 2.2 in [LL13], which still needs to be shown to hold for the Wright–Fisher diffusion), and thus we stop one step earlier and bound all the relevant integrals in terms of the moments of these regeneration times.

Following the arguments in Theorem 5.2 in [LL13], we decompose the path into the following chunks

$$\begin{aligned}
\mathbb{P}^{(\boldsymbol{\vartheta})}_{\mathbf{Z},\nu}\left[\left|\frac{1}{T}\int_0^T h(\mathbf{Z}^1_s)ds - \mathbb{E}^{(\boldsymbol{\vartheta})}_{\mathbf{Z}^1}\left[h(\boldsymbol{\eta})\right]\right| > \varepsilon\right] & \\
&\leq \mathbb{P}^{(\boldsymbol{\vartheta})}_{\mathbf{Z},\nu}\left[\left|\int_0^{R_1} \bar{h}(\mathbf{Z}^1_s)ds\right| > \frac{T\varepsilon}{3}\right] \\
&+ \mathbb{P}^{(\boldsymbol{\vartheta})}_{\mathbf{Z},\nu}\left[\left|\int_{R_1}^{R_{N_T+1}} \bar{h}(\mathbf{Z}^1_s)ds\right| > \frac{T\varepsilon}{3}; \Omega_T\right] \\
&+ \mathbb{P}^{(\boldsymbol{\vartheta})}_{\mathbf{Z},\nu}\left[\left|\int_{T}^{R_{N_T+1}} \bar{h}(\mathbf{Z}^1_s)ds\right| > \frac{T\varepsilon}{3}; \Omega_T\right] \\
&+ \mathbb{P}^{(\boldsymbol{\vartheta})}_{\mathbf{Z},\nu}\left[\Omega^c_T\right] \\
&=: A_1 + A_2 + A_3 + A_4,
\end{aligned}$$

where we define

$$\Omega_T := \left\{\left|\frac{N_T}{T} - \ell_{\boldsymbol{\vartheta}}\right| \leq \ell_{\boldsymbol{\vartheta}}\delta\right\}, \quad \delta = \frac{\varepsilon}{\|h\|_\infty}, \quad \bar{h} = h - \mathbb{E}^{(\boldsymbol{\vartheta})}_{\mathbf{Z}^1}\left[h(\boldsymbol{\eta})\right].$$

As in the proof of Theorem 5.2 in [LL13], each term is tackled separately.

For A_1, a simple Markov inequality together with the fact that $\|\bar{h}\|_\infty \leq 2\|h\|_\infty$ leads to

$$A_1 \leq \left(\frac{6\|h\|_\infty}{t\varepsilon}\right)^p \mathbb{E}^{(\boldsymbol{\vartheta})}_{\mathbf{Z},\nu}\left[R_1^p\right].$$

For A_2, the Fuk-Nagaev inequality given in Theorem A.1 in [LL13], implies that

$$\begin{aligned}
A_2 \leq C(p)\frac{\left(\ell^{p-1}_{\boldsymbol{\vartheta}} \vee \ell_{\boldsymbol{\vartheta}}\right)}{t^{p-1}\delta^{2(p-1)}}\Big(&2^p\mathbb{E}^{(\boldsymbol{\vartheta})}\left[(R_2 - R_1)^p\right] \\
&+ 4^{p-1}\mathbb{E}^{(\boldsymbol{\vartheta})}\left[(R_2 - R_1)^2\right]^{p-1}\Big).
\end{aligned}$$

For A_3 the same argument as the one used to deal with term C in the proof of Theorem 2.3, allows for the bound

$$A_3 \leq \frac{6^p \|h\|_\infty^p 2\ell_{\boldsymbol{\vartheta}}}{t^{p-1}\varepsilon^p} \mathbb{E}^{(\boldsymbol{\vartheta})} \left[(R_2 - R_1)^p \right].$$

Finally for A_4, following the arguments in Theorem 5.1 in [LL13] for $p \in \{2, 3, \dots\}$, leads to

$$\begin{aligned} A_4 \leq & \frac{2^{2(p-1)}}{t^{p-1}\varepsilon^{p-1}} \left(\mathbb{E}_{\mathbf{Z},\nu}^{(\boldsymbol{\vartheta})} \left[R_1^p \right] + \ell_{\boldsymbol{\vartheta}}^{-p} \right) \\ & + C(p) \frac{\left(\ell_{\boldsymbol{\vartheta}}^{p-1} \vee \ell_{\boldsymbol{\vartheta}} \right)}{\varepsilon^{2(p-1)} t^{p-1}} \left(\mathbb{V}^{(\boldsymbol{\vartheta})} \left[\bar{\eta}_1 \right]^{p-1} + \mathbb{E}^{(\boldsymbol{\vartheta})} \left[|\bar{\eta}_1|^p \right] \right), \end{aligned}$$

where $\bar{\eta}_1 := -(R_2 - R_1 - \ell_{\boldsymbol{\vartheta}}^{-1})$, and $\mathbb{V}^{(\boldsymbol{\vartheta})}$ denotes the variance.

In view of the following relations,

$$\begin{aligned} \ell_{\boldsymbol{\vartheta}} &= \mathbb{E}^{(\boldsymbol{\vartheta})} \left[R_2 - R_1 \right]^{-1}, \\ \mathbb{E}^{(\boldsymbol{\vartheta})} \left[|\bar{\eta}_1|^p \right] &\leq 2^{p-1} \left(\mathbb{E}^{(\boldsymbol{\vartheta})} \left[(R_2 - R_1)^p \right] + \mathbb{E}^{(\boldsymbol{\vartheta})} \left[R_2 - R_1 \right]^{-p} \right), \\ \mathbb{V}^{(\boldsymbol{\vartheta})} \left[\bar{\eta}_1 \right] &= \mathbb{E}^{(\boldsymbol{\vartheta})} \left[\bar{\eta}_1^2 \right], \end{aligned}$$

we get that the only terms we need to bound from above are

$$\mathbb{E}_{\mathbf{Z},\nu}^{(\boldsymbol{\vartheta})} \left[R_1^p \right], \qquad \mathbb{E}^{(\boldsymbol{\vartheta})} \left[(R_2 - R_1)^p \right],$$

for $p \in \{2, 3, \dots\}$, whilst we need to bound $\mathbb{E}^{(\boldsymbol{\vartheta})}[R_2 - R_1]$ from below to deal with $\ell_{\boldsymbol{\vartheta}}$. It now remains to show how to deal with these quantities in a way that allows us to track their dependence on the parameter $\boldsymbol{\vartheta}$. In the next section we illustrate how this is achieved in Theorem 4.1 in [LL13] by resorting to their Assumption 2.2, explaining the difficulty involved in verifying this assumption in our case, and briefly showing how the moments of hitting times of sets can be controlled via standard multidimensional diffusion theory and Lyapunov functions for the Wright–Fisher case. We point out that in [LL13] the authors are solely interested in controlling the resulting quantities in terms of the initial starting point $\mathbf{x}$, and thus they treat $\ell_{\boldsymbol{\vartheta}}$ as a constant. In contrast, in our setting we need to be able to bound this quantity from above in

terms of $\boldsymbol{\vartheta}$, which will not readily follow from Theorem 4.1 nor any of the techniques present in [LL13] as these consider upper bounds of the moments as opposed to the lower bounds we require here, in virtue of (5.12).

5.3 Controlling the moments of regeneration times

We end this chapter by discussing the unsolved problem of controlling the moments of the regeneration times which we have used to bound the LHS of (5.8). Once suitable control over these components in the parameter $\boldsymbol{\vartheta}$ is obtained, uniform in the selection and mutation parameter ergodicity for the multidimensional Wright–Fisher diffusion follows, and thus one might start looking into extending Theorem 2.8 from Chapter 2 for this case, to establish uniform local asymptotic normality. We start by giving a brief overview of how Theorem 4.1 in [LL13] grants control over the moments of the regeneration times via Lyapunov functions, whilst along the way illustrating precisely where this method hinges upon their Assumption 2.2. We then end this section by showing how standard multidimensional diffusion theory coupled with the right choice of Lyapunov function enables one to control the moments of the hitting times of the K-dimensional Wright–Fisher diffusion (for a specific set of parameter configurations).

5.3.1 Theorem 4.1

The crucial quantity necessary to establish control over the moments of the regeneration times defined in Section 5.2 is the drift condition (5.1) of Assumption 5.1.

The calculations in Section 5.2 lead to bounds that depend on the quantities $\mathbb{E}^{(\boldsymbol{\vartheta})}[(R_2 - R_1)^p]$ and $\mathbb{E}^{(\boldsymbol{\vartheta})}_{\mathbf{Z},\nu}[R_1^p]$. To control these, the authors in [LL13] concentrate on expressions of the form $\mathbb{E}^{(\boldsymbol{\vartheta})}_{\mathbf{Z},\mathbf{x}}[\int_0^{R_1} r(s)ds]$ in Theorem 4.1, where $r(\cdot)$ is some rate function which allows access to the desired moment. The proof of Theorem 4.1 is split into 3 parts;

the first establishes that

$$\mathbb{E}^{(\boldsymbol{\vartheta})}_{\mathbf{Z},\mathbf{x}}\left[\int_0^{\tilde{S}_1} r(s)ds\right] \leq aV(\mathbf{x}) + b, \tag{5.14}$$

for a, b positive constants (in the remainder of this section we shall reserve a, b to mean some positive constants whose exact values do not need to be tracked), and the stopping time in the limit of integration is the resolvent chain's hitting time of the set C used to construct the regeneration times, i.e.

$$\tilde{S}_1 := \inf\{T_m : \mathbf{Z}^1_{T_m} \in C\}, \qquad \tilde{S}_n := \inf\{T_m > \tilde{S}_{n-1} : \mathbf{Z}^1_{T_m} \in C\},$$

for $n \geq 2$. The second part of the proof uses and extends (5.14) to the case when the upper limit of integration is S_1 as defined in (5.11), and similarly the third part to the case when the upper limit is R_1 as given in (5.11). Proving (5.14) relies on results for the resolvent kernel found in [DFMS04] and [DFG09], whilst the extensions leading to similar bounds for S_1 and R_1 rely on (5.14) and exploiting the structure introduced by the Nummelin splitting scheme used to construct the process $\mathbf{Z}$. Thus in what follows we focus on how to prove (5.14).

To this end, observe first that Assumption 5.1 above is a drift condition for the continuous time process, however as shown in Theorem 4.9 in [DFG09], this same assumption induces a similar drift condition on the resolvent chain, with a different petite set and functions $\bar{\Phi}, \bar{V}$ (which turn out to be very similar to the ones for the continuous times process), but with the same rate function r. As given in the aforementioned theorem, we have that $\bar{\Phi}(t(1 + \Phi'(1))) \sim \Phi(t)$ as $t \to \infty$ (where $\sim$ in this context means that the two quantities are asymptotically equivalent), and $\|\bar{V} - (1 + \Phi'(1))V\|_\infty < \infty$, such that $\bar{V}(\mathbf{x}) \leq aV(\mathbf{x}) + b$.

The above, coupled with Proposition 2.2 in [DFMS04], implies that for

measurable $A \in \Delta_K$ with $f_{\boldsymbol{\vartheta}}(A) > 0$, and $\bar{T}_A := \inf\{n \geq 1 : \mathbf{X}_{T_n} \in A\}$

$$\mathbb{E}^{(\boldsymbol{\vartheta})}_{\mathbf{Z},\mathbf{x}} \left[\sum_{k=0}^{\bar{T}_A - 1} r(k) \right] \leq a\bar{V}(\mathbf{x}) + b \leq aV(\mathbf{x}) + b, \tag{5.15}$$

which allows bounding the moments of the regeneration times from above by a suitable Lyapunov function (and we emphasize here that in (5.15) we have viewed $\mathbf{X}_{T_n}$ as the first co-ordinate of $\mathbf{Z}_{T_n}$). Note that finding the correct set B and functions Φ, V such that the drift condition (5.1) holds for a given diffusion is rather hard to verify in practice, however if the function V belongs to the domain of the generator $\mathscr{G}$ of $\mathbf{X}$, then Theorem 3.4 in [DFG09] gives that (5.2) implies Assumption 5.1, which is an easier bound to get to.

In [LL13], the authors show how standard norm-like Lyapunov functions (for instance $V(\mathbf{x}) = \|\mathbf{x}\|^m$ for suitably chosen m) verify (5.2) for general multidimensional diffusions outside a ball of specific radius. Considering functions which display an analogous behaviour at the boundary $\partial\Delta_K$ for the Wright–Fisher diffusion (such as $V(\mathbf{x}) = -\log x_1$, [Buz19, Section 4.4.3], or $V(\mathbf{x}) = x_1^{-1}(1 - x_1)$) does not lead to bounds similar to (5.2), and thus Theorem 4.1 cannot be directly applied. It is also not a priori clear whether such a bound is indeed feasible for this case and thus remains an open problem. In a first attempt to gain a better understanding of what a bound of the type (5.2) implies for the Wright–Fisher diffusion, in the next section we tackle the slightly easier problem of dealing with moments of the hitting times of a specific set by making use of standard multidimensional diffusion theory. Note however that whilst this might provide some intuition on how these moments of the regeneration times can be tackled via Lyapunov functions, it does not offer much in terms of gaining control over the term $\ell_{\boldsymbol{\vartheta}}$, for which we shall require a separate approach altogether.

We summarise the above described proof of Theorem 5.2 in [LL13] via the schematic diagram in Figure 5.3.1.

5.3.2 Controlling the moments of hitting times

Suppose now that we are interested in controlling the moments of the hitting times of the multidimensional Wright–Fisher diffusion $\mathbf{X}$ in terms of the parameter $\boldsymbol{\vartheta}$. In the one-dimensional case, we exploited the notions of speed and scale to derive relatively simple recursive second order ODEs, whose solutions we could bound from above in $\boldsymbol{\vartheta}$ relatively easily, leading to the desired bounds as seen in Theorem 2.2. Directly extending this method to the multidimensional setting is not feasible as these notions do not extend to higher dimensions, so instead the approach here will be to appeal to Lyapunov functions.

Below we quote the second part of a more general result (Theorem 3.9 in [Kha12]) for multidimensional diffusions where Lyapunov functions are used to ensure that the underlying process is recurrent, as well as provide simple upper bounds on functions of the hitting times of specific sets. We have suitably translated the result for the K-dimensional Wright–Fisher diffusion with generator $\mathscr{G}$ given in (5.3), and omitted the first part of said theorem (as we already have that the multidimensional Wright–Fisher diffusion with PIM is recurrent).

Theorem 5.2 (Theorem 3.9 in [Kha12] for the Wright–Fisher diffusion)**.** *Let* $\mathbf{X}$ *be the* K*-dimensional Wright–Fisher diffusion with generator* $\mathscr{G}$*, and denote by* $G := \frac{\partial}{\partial t} + \mathscr{G}$*. If there exists a non-negative function* $V(t, \mathbf{x})$ *which is twice continuously differentiable with respect to* $\mathbf{x}$ *and continuously differentiable with respect to* t*, such that*

$$GV(t, \mathbf{x}) \leq -\alpha(t) \tag{5.16}$$

holds $\forall t \geq 0$ *and* $\mathbf{x} \in D \subset \Delta_K$*, where* $\alpha(t) \geq 0$ *is a function such that*

$$\beta(t) = \int_0^t \alpha(s)ds \to \infty, \qquad \text{as } t \to \infty,$$

then $\mathbb{E}^{(\boldsymbol{\vartheta})}_{\mathbf{X},\mathbf{x}}[\beta(T_{D^c})]$ *exists and satisfies the inequality*

$$\mathbb{E}^{(\boldsymbol{\vartheta})}_{\mathbf{X},\mathbf{x}}\left[\beta(T_{D^c})\right] \leq \beta(0) + V(0,\mathbf{x}),$$

where $D^c = \Delta_K \setminus D$ *and* $T_{D^c} := \inf\{t \geq 0 : \mathbf{X}_t \in D^c\}$.

We point out here that the above theorem is phrased in [Kha12] in terms of the random variable τ_U which is defined as the first time for which the process exits the set U, as opposed to the way in which we phrased the above theorem in terms of first hitting time of a set. This theorem gives us a precise control on the desired functions of the hitting times provided we can find a suitable candidate function $V(t,\mathbf{x})$, a suitable function $\alpha(t)$, and a suitable set D. The choice of function $\alpha(\cdot)$ is dictated by what function $\beta(\cdot)$ of the hitting time we wish to control, so if we want to deal with moments, a natural candidate would be $\alpha(s) = s^{p-1}$ where p is the desired moment (however we opt for a slightly different function as detailed below). The choice of V and D is somewhat more involved and depends on the process $\mathbf{X}$; below we give the details pertinent to a special case to show the technicalities involved in choosing V and D, as well as the fact that such a technique is practicable for the Wright–Fisher diffusion, however we emphasise that this special case is rather restrictive and further work is necessary for this approach to be applicable in the general case.

In what follows we consider a general mutation structure, i.e. we do not assume PIM and instead deal with the infinitesimal mutation matrix given by $(\gamma_{ij})_{i,j=1}^K$. Set $V(t,\mathbf{x}) = \frac{(t+1)^p}{p}(-\log(x_1))$ and $\alpha(t) = (t+1)^{p-1}$ (the reason behind this choice over t^{p-1} becomes clearer in the calculations below), then we have that

$$\begin{aligned}
\frac{\partial V}{\partial t}(t,\mathbf{x}) &= (t+1)^{p-1}(-\log x_1), & \frac{\partial V}{\partial x_1}(t,\mathbf{x}) &= \frac{(t+1)^p}{p}\left(-\frac{1}{x_1}\right),\\
\frac{\partial V}{\partial x_j}(t,\mathbf{x}) &= 0, & \frac{\partial^2 V}{\partial x_1^2}(t,\mathbf{x}) &= \frac{(t+1)^p}{p}\left(\frac{1}{x_1^2}\right),\\
\frac{\partial^2 V}{\partial x_j \partial x_1}(t,\mathbf{x}) &= 0, & \frac{\partial^2 V}{\partial x_j^2}(t,\mathbf{x}) &= 0,
\end{aligned}$$

where $j \neq 1$. Substituting the above into the expression for the generator as given in (5.3), we get that

$$\begin{aligned}
GV(t,x) &= \frac{\partial V}{\partial t}(t,\mathbf{x}) + \frac{1}{2}\sum_{i,j=1}^{K} a_{ij}(t,\mathbf{x})\frac{\partial^2 V}{\partial x_i \partial x_j}(t,\mathbf{x}) \\
&\quad + \sum_{i=1}^{K} b_i(t,\mathbf{x})\frac{\partial V}{\partial x_i}(t,\mathbf{x}) \\
&= (t+1)^{p-1}(-\log x_1) + \frac{1}{2}\sum_{i,j=1}^{K} x_i\left(\delta_{ij} - x_j\right)\frac{\partial^2 V}{\partial x_i \partial x_j}(t,\mathbf{x}) \\
&\quad + \sum_{i=1}^{K}\left(\sum_{j=1}^{K}\gamma_{ji}x_j + x_i\left(\sum_{j=1}^{K}\sigma_{ij}x_j - \sum_{k,l=1}^{K}\sigma_{kl}x_k x_l\right)\right)\frac{\partial V}{\partial x_i}(t,\mathbf{x}) \\
&= (t+1)^{p-1}(-\log x_1) + \frac{(t+1)^p}{p}\left[\frac{1}{2}x_1\left(1-x_1\right)\frac{1}{x_1^2}\right. \\
&\quad \left. + \left(\sum_{j=1}^{K}\gamma_{j1}x_j + x_1\left(\sum_{j=1}^{K}\sigma_{1j}x_j - \sum_{k,l=1}^{K}\sigma_{kl}x_k x_l\right)\right)\left(-\frac{1}{x_1}\right)\right] \\
&= (t+1)^{p-1}\left\{ -\log x_1 + \frac{t+1}{p}\left[\frac{1}{2}\frac{(1-x_1)}{x_1}\right.\right. \\
&\quad \left.\left. - \left(\sum_{j=1}^{K}\gamma_{j1}\frac{x_j}{x_1} + \left(\sum_{j=1}^{K}\sigma_{1j}x_j - \sum_{k,l=1}^{K}\sigma_{kl}x_k x_l\right)\right)\right]\right\}.
\end{aligned}$$

Comparing this last quantity with the RHS of (5.16), we deduce that it suffices to show that

$$\begin{aligned}
&-\log x_1 + \frac{t+1}{p}\left[\frac{1}{2}\frac{(1-x_1)}{x_1}\right. \\
&\quad \left. - \left(\sum_{j=1}^{K}\gamma_{j1}\frac{x_j}{x_1} + \left(\sum_{j=1}^{K}\sigma_{1j}x_j - \sum_{k,l=1}^{K}\sigma_{kl}x_k x_l\right)\right)\right] \leq -1,
\end{aligned}$$

holds on some subset $D \subset \Delta_K$, $\forall t \geq 0$. Note that had we set $\alpha(t) = t^{p-1}$, then we would not be able to find a set $D \subset \Delta_K$ for which the above would hold for any $t \geq 0$. Taking $\alpha(t) = (t+1)^{p-1}$ ensures that such D can be found (indeed we could have chosen $\alpha(t) = (t+\varepsilon)^{p-1}$ for

any $\varepsilon > 0$). If we let $\gamma_1^- := \min_{j\in\{1,\dots,K\}} \gamma_{j1}$, $\sigma^+ := \max_{k,l\in\{1,\dots,K\}} \sigma_{kl}$, and $\sigma_1^- := \min_{j\in\{1,\dots,K\}} \sigma_{1j}$, then

$$\begin{aligned}
&-\log x_1 + \frac{1}{2}\frac{(1-x_1)}{x_1}\frac{t+1}{p} \\
&\quad + \left(-\sum_{j=1}^{K}\gamma_{j1}\frac{x_j}{x_1} - \left(\sum_{j=1}^{K}\sigma_{1j}x_j - \sum_{k,l=1}^{K}\sigma_{kl}x_k x_l\right)\right)\frac{t+1}{p} \\
&\leq -\log x_1 + \frac{1}{2}\frac{(1-x_1)}{x_1}\frac{t+1}{p} \\
&\quad + \left(\gamma_1^-\left(-1-\frac{1-x_1}{x_1}\right) + \sigma^+ - \sigma_1^-\right)\frac{t+1}{p} \\
&= -\log x_1 + \frac{t+1}{p}\left(\frac{1}{2}-\gamma_1^-\right)\frac{(1-x_1)}{x_1} + \left(\tilde{\sigma}-\gamma_1^-\right)\frac{t+1}{p},
\end{aligned}$$

where we set $\tilde{\sigma} = \sigma^+ - \sigma_1^-$, and have used the identities $\sum_{j=1}^{K} x_j = 1 = \sum_{k,l=1}^{K} x_k x_l$ as we are working on the $K-1$ dimensional simplex. Using the above, we are left with finding a set $D \subset \Delta_K$ such that

$$-\log x_1 + \frac{t+1}{p}\left(\frac{1}{2}-\gamma_1^-\right)\frac{(1-x_1)}{x_1} + \left(\tilde{\sigma}-\gamma_1^-\right)\frac{t+1}{p} \leq -1$$

holds $\forall \mathbf{x} \in D$ and $\forall t \geq 0$. If $2\gamma_1^- > 1$, then the term $(1-x_1)x_1^{-1}$ goes to $-\infty$ as x_1 goes to 0, and indeed it does so fast enough to offset the fact that the log term goes off to ∞. Note in particular that we cannot accommodate the case when $2\gamma_1^- \leq 1$ with the current choice of Lyapunov function, for then the LHS above will always remain positive. Whilst this means that the result we derive here is only applicable to a subset of the cases of interest (namely those with a high enough mutation parameter), finding a Lyapunov function for which a bound of the type (5.16) holds for all possible values of the mutation parameters is a non-trivial task in view of the change in the diffusion's boundary behaviour at the first co-ordinate when $\gamma_1^- < 1/2$ and when $\gamma_1^- \geq 1/2$.

With the above restriction on γ_1^-, we now characterise the set D explicitly by making some further assumptions on $\tilde{\sigma}$. To this end,

define

$$f(t,x) := -\log x + \frac{t+1}{p}\left(\frac{1}{2} - \gamma_1^-\right)\frac{(1-x)}{x} + \left(\tilde{\sigma} - \gamma_1^-\right)\frac{t+1}{p}$$

and observe that for $f(t,x)$ to be non-increasing in t we need

$$\frac{\partial}{\partial t} f(t,x) \leq 0 \iff x \leq \frac{\gamma_1^- - \frac{1}{2}}{\tilde{\sigma} - \frac{1}{2}},$$

and since $x \in [0,1]$, this will only be possible if $\tilde{\sigma} > 1/2$. So for $\tilde{\sigma}, \gamma_1^- > 1/2$, if we choose x smaller than $\frac{\gamma_1^- - 1/2}{\tilde{\sigma} - 1/2}$, $f(t,x) \leq f(0,x)$ and now it suffices to find a set on which $f(0,x)$ is bounded from above by -1.

Observe that $f(0,x)$ has a maximum at $x_{\max} := -p^{-1}(1/2 - \gamma_1^-)$, $f(0,x) \to -\infty$ as $x \to 0$, and $\frac{\partial}{\partial x} f(0,x) > 0$ for all $x \in (0, x_{\max})$. If $f(0, x_{\max}) \leq -1$, then we can set

$$D := \left\{ \mathbf{x} \in \Delta_K : x_1 \leq \frac{\gamma_1^- - 1/2}{\tilde{\sigma} - 1/2} \right\}, \tag{5.17}$$

and then (5.16) holds on $(0,\infty) \times D$ as required. Otherwise, we are guaranteed that $\exists \hat{x} \in (0, x_{\max})$ for which $f(0,x) \leq -1 \ \forall x \in (0, \hat{x})$ in light of the fact that $f(0,x)$ is non-decreasing over $(0, x_{\max})$. Thus if we set

$$D := \left\{ \mathbf{x} \in \Delta_K : x_1 \in \left(0, \left(\hat{x} \wedge \frac{\gamma_1^- - \frac{1}{2}}{\tilde{\sigma} - \frac{1}{2}}\right)\right) \right\}, \tag{5.18}$$

then we have that (5.16) holds on $(0,\infty) \times D$ as required.

Thus by setting D to be (5.17) or (5.18) (depending on p, γ_1^- and $\tilde{\sigma}$, but regardless observe that D will always be non-empty and open), taking $V(t,\mathbf{x}) = \frac{(t+1)^p}{p}(-\log x_1)$, and $\alpha(t) = (t+1)^{p-1}$, by Theorem 5.2 we get that

$$\mathbb{E}_{\mathbf{x}}\left[T_{D^c}^p\right] \leq \mathbb{E}_{\mathbf{x}}\left[(1 + T_{D^c})^p\right] \leq p\left(\beta(0) + V(0,\mathbf{x})\right) = 1 - \log x_1.$$

In addition, we get that for any initial distribution ν on Δ_K for which $\int_{\Delta_K} -\log x_1 \nu(d\mathbf{x}) < \infty$,

$$\begin{aligned}\mathbb{E}_\nu\left[T_{D^c}^p\right] \leq \mathbb{E}_\nu\left[(T_{D^c}+1)^p\right] &= \int_{\Delta_K} \mathbb{E}_{\mathbf{x}}\left[(1+T_{D^c})^p\right] \nu(d\mathbf{x}) \\ &\leq 1 + \int_{\Delta_K} -\log x_1 \nu(d\mathbf{x})\end{aligned}$$

is finite.

Despite the fact that we can control these moments in terms of $\boldsymbol{\vartheta}$, there is no straightforward way in which we can relate them to the regeneration times introduced in Section 5.2. In particular, the regeneration times are defined at the resolvent chain's times $\{T_n\}_{n\in\mathbb{N}}$ whose underlying dynamics are driven by an exponential random variable, and are "blind" to what happens to the process in between the exponential time increments. Thus obtaining control over these regeneration times by appealing to the more granular hitting times used above does not seem to be feasible.

Nonetheless the above calculations offer some hope that Lyapunov functions might be a useful tool in obtaining upper bounds over the regeneration times for some parameter configurations, besides serving as a suitable testbed through which certain properties of the generator (5.3) can be brought to light. In particular, the above imposed restriction on the mutation parameters suggest that different Lyapunov functions will be necessary to reflect the change in boundary behaviour as the mutation parameters cross the threshold value of 1/2. However, it might equally be the case that these Lyapunov type arguments are not the right tool to use for the low mutation regime, as there is no guarantee that a suitable Lyapunov function exists.

To illustrate this heuristically, assume PIM and that $\theta_i < 1$ for all $i = 1, \ldots, K$. The invariant density $f_{\boldsymbol{\vartheta}}$ is strictly increasing on approach of the boundary $\partial\Delta_K$, and is trough shaped as can be seen in Figure 5.3.2 (where we have plotted the invariant density

for a neutral three-dimensional Wright–Fisher diffusion, which is different to the one corresponding to the non-neutral process, but the general shape of $f_{\boldsymbol{\vartheta}}$ at the boundary $\partial\Delta_K$ is the same in both cases).

In comparison, the bound (5.16) can be interpreted as finding a set D whose probability of being visited is strictly decreasing in time, which would be impossible to find given that no such set exists for the invariant density. Of course this is far from being a valid formal statement, but it offers a clear visual representation of the potential limitations of Lyapunov type arguments. As mentioned earlier, independently of the above one would need to devise a way to control $\ell_{\boldsymbol{\vartheta}}$ from above, which coupled with the necessary bounds for the above moments, enables uniform in the parameter $\boldsymbol{\vartheta}$ ergodicity to be established for the multidimensional Wright–Fisher diffusion.

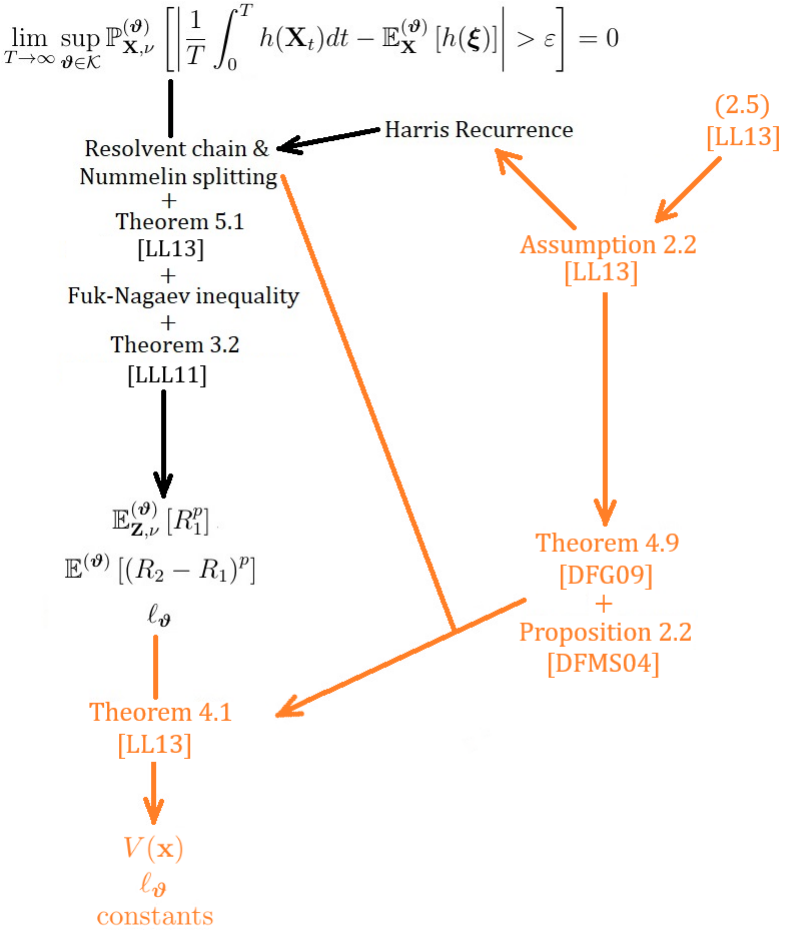


Figure 5.3.1: Flow chart illustrating the proof structure for Theorem 5.2 in [LL13]. Arrows and statements in black represent quantities and arguments that can be entertained for the K-dimensional Wright–Fisher diffusion, whilst quantities in orange represent statements or results that are used in Theorem 5.2 in [LL13] but have not been proved to hold or apply in the Wright–Fisher case

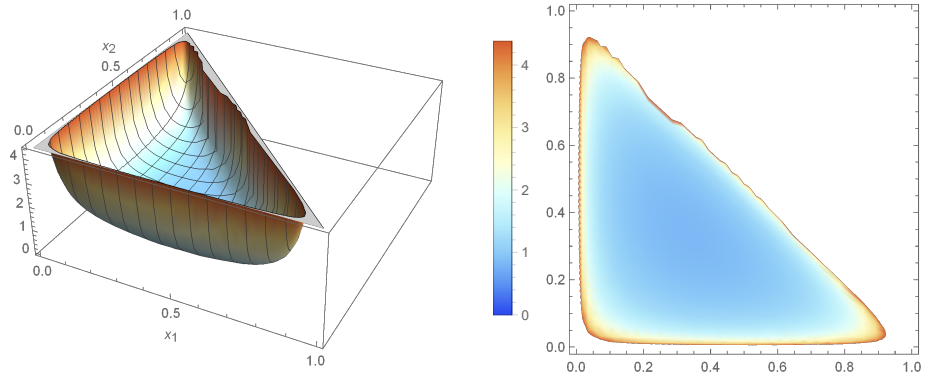


Figure 5.3.2: Plot of the invariant density for a neutral three-dimensional Wright–Fisher diffusion with mutation parameter $\boldsymbol{\theta} = (0.5, 0.5, 0.5)$. The figure on the left is a three-dimensional rendition of the resulting invariant Dirichlet density $f_{\boldsymbol{\vartheta}}$, whilst the plot on the right is the resulting contour plot

Chapter 6

Discussion

In this thesis we have derived readily verifiable criteria to establish $\boldsymbol{\vartheta}$-uniform ergodicity for bounded scalar diffusions with either entrance or regular boundaries, and extended this result for a specific class of unbounded functions in the case of a diffusion with solely entrance boundaries in Chapter 2. These conditions were subsequently used to deduce the uniform in the selection and mutation parameters ergodicity of the scalar Wright–Fisher diffusion, as well as the uniform local asymptotic normality of the family of laws induced by the solutions to the corresponding SDEs. In Chapter 3 we then utilised these two properties of the Wright–Fisher diffusion to show that the ML and Bayesian estimators for selection have a desirable set of properties in the continuous observation regime, where the results hinged on the classical Ibragimov-Has'minskii conditions.

Apart from the theoretical guarantees proved for this limiting observational regime in the one dimensional case, a practicable MCMC scheme to conduct exact inference on the allele age and selection coefficient was proposed in Chapter 4. The method was applied to simulated data and the output obtained suggests that it performs reasonably well even when a limited amount of data is available and priors are mis-specified. We concluded this thesis by considering the challenges involved in extending the results present in Chapter 2 to the multidimensional case, providing a brief sketch of how this might be achieved via regeneration arguments, and highlighting

the remaining open problems.

Whilst the results in Chapter 2 imposed no undue restrictions upon the mutation parameters (other than requiring them to be greater than or equal to 1 to allow for the Radon–Nikodym derivative to be defined in Corollary 2.6 and Theorem 2.8), Chapter 3 considers inference for the selection parameter σ when the mutation rates are known. Although this assumption is a limitation on the study, we emphasise that in the continuous observation regime considered in Chapter 3 these can be inferred directly once the diffusion gets arbitrarily close to either boundary (see Remark 3.1 for the corresponding details). Nonetheless, extending this work to include mutation parameters greater than or equal to 1 would be of great interest, and proves to be rather challenging using the setup of Chapter 3, as illustrated in Section 3.4. The likelihood ratio function in this case would involve expressions featuring the unbounded functions $(1-x)x^{-1}$ and $x(1-x)^{-1}$ (as witnessed in Theorem 2.8), which require much more delicate arguments (if not an entirely different approach) to proving Propositions 3.5 and 3.6, in order to establish the same conclusions as in Theorem 3.2. These observations suggest that perhaps a better way of tackling inference for the mutation parameters is to directly analyse the boundary behaviour of the diffusion rather than extend the approach adopted in Chapter 3. By understanding this boundary behaviour, characterising it precisely, and relating it explicitly to the mutation parameters, one might be able to devise a suitable statistical framework within which to phrase and prove results similar to those present in Chapter 3 for the mutation parameters.

On a similar note to the above, extending the framework developed in Chapter 4 to accommodate for the mutation parameters is rather tricky. The main issue here is finding a suitable dominating measure which is independent of the parameters of interest and with respect to which all the quantities involved admit a tractable density. As is evident throughout this chapter, the mutation parameters appear in all the dominating measures via the laws of the underlying neutral

Wright–Fisher diffusion, and thus this dependence needs to be tackled for a Gibbs sampler to be viable. One way around this would be to fix a specific value for the mutation parameters and use this as the reference measure, however as pointed out in Section 4.5, this would only be feasible for mutation parameters greater than or equal to 1, for otherwise the Radon–Nikodym derivative between the laws of the Wright–Fisher diffusions with differing mutation rates need not be defined. In addition, the inclusion of the mutation parameters translates into the presence of unbounded functions in the exponent of the resulting Radon–Nikodym derivatives, and thus the Poisson point process construction utilised in Chapter 4 breaks down. It is evident that further work is required before one can entertain the idea of conducting joint inference of allele age, selection and mutation. We point out however that other extensions such as accounting for historical demographic changes, and allowing for soft sweeps are more straightforward.

The conclusions obtained in Chapter 3 offer a certain degree of confidence that standard inferential schemes for the selection parameter will return reliable results, at least in the limiting regime of continuous observations. Whilst this is a rather favourable starting point, it does not guarantee that these same schemes will return similar results in the discrete observation setting. The empirical results present in Section 3.3 do suggest that this might be the case when the observation regime converges to densely sampled data, however this is yet to be formally proved. Moreover, the techniques used here do not provide much in terms of insight on how the corresponding results could be addressed since the likelihood ratio would now be given by a product of transition densities, and as illustrated in Chapter 4, dealing with the latter is rather delicate in view of the intractable terms involved.

The arguments illustrated in Chapter 5 shed light on the increased difficulty in proving similar results to those obtained in Chapter 2 for the multidimensional diffusion setting. Figure 5.3.1 provides a brief recap of the techniques used in [LL13] to obtain bounds on the rate

of convergence of time averages to state space averages for a class of multidimensional diffusions, whilst also illustrating which parts of the proof hold for the Wright–Fisher case. The main problem here was verifying Assumption 5.1 or bound (5.2) via a suitable choice of Lyapunov function for the Wright–Fisher case. In particular, as seen in the calculations in Section 5.3.2, the results might only apply to a subset of the cases of interest (i.e. when the mutation parameters are sufficiently large), whilst the remaining cases (i.e. smaller mutation parameters) might require a different technique all together as the Lyapunov function theory seems to fall short.

Bibliography

[AR09] Christophe Andrieu and Gareth O. Roberts. The pseudo-marginal approach for efficient Monte Carlo computations. *Ann. Statist.*, 37(2):697 – 725, 2009.

[aR11] Krzysztof Łatuszyński and G. Roberts. Clts and asymptotic variance of time-sampled Markov chains. *Methodol. Comput. Appl. Probab.*, 15:237–247, 2011.

[Bar65] A. A. Barker. Monte Carlo calculations of the radial distribution functions for a proton-electron plasma. *Aust. J. Phys.*, 18(2):119–134, 1965.

[Bea03] Mark A. Beaumont. Estimation of population growth or decline in genetically monitored populations. *Genetics*, 164(3):1139–1160, 2003.

[BEG00] A. D. Barbour, S. N. Ethier, and R. C. Griffiths. A transition function expansion for a diffusion model with selection. *Ann. Appl. Probab.*, 10(1):123 – 162, 2000.

[BPR06] Alexandros Beskos, Omiros Papaspiliopoulos, and Gareth O. Roberts. Retrospective exact simulation of diffusion sample paths with applications. *Bernoulli*, 12(6):1077–1098, 2006.

[BPR08] A. Beskos, O. Papaspiliopoulos, and G. Roberts. A factorisation of diffusion measure and finite sample path

constructions. *Methodol. Comput. Appl. Probab.*, 10:85–104, 2008.

[BPRF06] Alexandros Beskos, Omiros Papaspiliopoulos, Gareth O. Roberts, and Paul Fearnhead. Exact and computationally efficient likelihood-based estimation for discretely observed diffusion processes. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 68(3):333–382, 2006.

[BR05] Alexandros Beskos and Gareth O. Roberts. Exact simulation of diffusions. *Ann. Appl. Probab.*, 15(4):2422 – 2444, 2005.

[Buz19] Eugenio Buzzoni. *Properties of the Wright–Fisher diffusion with seed banks and multiple islands.* Doctoral thesis, Technische Universität Berlin, Berlin, 2019.

[BYN] Jonathan P. Bollback, Thomas L. York, and Rasmus Nielsen. Estimation of 2n.

[Can74] C. Cannings. The latent roots of certain Markov chains arising in genetics: a new approach. I. Haploid models. *Advances in Appl. Probability*, 6:260–290, 1974.

[D+20] Marianne Dehasque et al. Inference of natural selection from ancient DNA. *Evolution Letters*, 4(2):94–108, 2020.

[Dev86] Luc Devroye. *Non-Uniform Random Variate Generation(originally published with.* Springer-Verlag, 1986.

[DFG09] Randal Douc, Gersende Fort, and Arnaud Guillin. Subgeometric rates of convergence of f-ergodic strong Markov processes. *Stochastic Process. Appl.*, 119(3):897–923, 2009.

[DFMS04] Randal Douc, Gersende Fort, Eric Moulines, and Philippe Soulier. Practical drift conditions for subgeometric rates of convergence. *Ann. Appl. Probab.*, 14(3):1353 – 1377, 2004.

[DMS93] D. A. Dawson, B. Maisonneuve, and J. Spencer. *École d'Été de Probabilités de Saint-Flour XXI—1991*, volume 1541 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 1993. Papers from the school held in Saint-Flour, August 18–September 4, 1991, Edited by P. L. Hennequin.

[Dur08] Richard Durrett. *Probability models for DNA sequence evolution*. Probability and its Applications (New York). Springer, New York, second edition, 2008.

[EG93] S. N. Ethier and R. C. Griffiths. The Transition Function of a Fleming–Viot Process. *The Annals of Probability*, 21(3):1571 – 1590, 1993.

[EK86] Stewart N. Ethier and Thomas G. Kurtz. *Markov processes*. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. John Wiley & Sons, Inc., New York, 1986. Characterization and convergence.

[EK98] S. N. Ethier and Thomas G. Kurtz. Coupling and ergodic theorems for Fleming–Viot processes. *Ann. Probab.*, 26(2):533 – 561, 1998.

[EM10] Charles L Epstein and Rafe Mazzeo. Wright–Fisher diffusion in one dimension. *SIAM J. Math. Anal.*, 42(2):568–608, 2010.

[F+19] A. Fages et al. Tracking five millennia of horse management with extensive ancient genome time series. *Cell*, 177:1419 – 1435.e31, 2019.

[FALJW16] A. Ferrer-Admetlla, Christoph Leuenberger, J. Jensen, and Daniel Wegmann. An approximate Markov model for the Wright-–Fisher diffusion and its application to time series data. *Genetics*, 203:831 – 846, 2016.

[Fel51] William Feller. Diffusion processes in genetics. In *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability, 1950*, pages 227–246. University of California Press, Berkeley and Los Angeles, Calif., 1951.

[Fis99] R. A. Fisher. *The genetical theory of natural selection.* Oxford University Press, Oxford, variorum edition, 1999. Revised reprint of the 1930 original, Edited, with a foreword and notes, by J. H. Bennett.

[GaR17] Flávio B. Gonçalves, Krzysztof Łatuszyński, and Gareth O. Roberts. Barker's algorithm for Bayesian inference with intractable likelihoods. *Braz. J. Probab. Stat.*, 31(4):732–745, 2017.

[GJS18] Robert C Griffiths, Paul A Jenkins, and Dario Spano. Wright–Fisher diffusion bridges. *Theor. Popul. Biol.*, 122:67–77, 2018.

[GL83] Robert C. Griffiths and Wen-Hsiung Li. Simulating allele frequencies in a population and the genetic differentiation of populations under mutation pressure. *Theor. Popul. Biol.*, 23(1):19–33, 1983.

[Gri79] R. C. Griffiths. A transition density expansion for a multi-allele diffusion model. *Adv. in Appl. Probab.*, 11(2):310–325, 1979.

[Gri80] Robert C Griffiths. Lines of descent in the diffusion approximation of neutral Wright–Fisher models. *Theor. Popul. Biol.*, 17(1):37–50, 1980.

[Gri84] R. C. Griffiths. Asymptotic line-of-descent distributions. *J. Math. Biol.*, 21(1):67–75, 1984.

[GS09] R. Griffiths and Dario Spanò. Diffusion processes and coalescent trees. *arXiv preprint arXiv:1003.4650*, 2009.

[GS12] Shota Gugushvili and Peter Spreij. Parametric inference for stochastic differential equations: a smooth and match approach. *ALEA Lat. Am. J. Probab. Math. Stat.*, 9(2):609–635, 2012.

[HDBY20a] Zhangyi He, Xiaoyang Dai, Mark Beaumont, and Feng Yu. Detecting and quantifying natural selection at two linked loci from time series data of allele frequencies with forward-in-time simulations. *Genetics*, 216(2):521–541, 2020.

[HDBY20b] Zhangyi He, Xiaoyang Dai, Mark Beaumont, and Feng Yu. Estimation of Natural Selection and Allele Age from Time Series Allele Frequency Data Using a Novel Likelihood-Based Approach. *Genetics*, 216(2):463–480, 10 2020.

[HL03] R. Höpfner and E. Löcherbach. Limit theorems for null recurrent Markov processes. *Mem. Amer. Math. Soc.*, 161(768):vi+92, 2003.

[HP05] Joachim Hermisson and Pleuni S Pennings. Soft sweeps: molecular population genetics of adaptation from standing genetic variation. *Genetics*, 169(4):2335–2352, 2005.

[HSK+05] S Hummel, D Schmidt, B Kremeyer, B Herrmann, and M Oppermann. Detection of the CCR5-Δ32 HIV resistance gene in bronze age skeletons. *Genes Immun.*, 6(4):371–374, 2005.

[IH81] I. A. Ibragimov and R. Z. Has'minskii. *Statistical estimation: Asymptotic theory*. Springer-Verlag, 1981.

[IW89] Nobuyuki Ikeda and Shinzo Watanabe. *Stochastic differential equations and diffusion processes*, volume 24 of *North-Holland Mathematical Library*. North-Holland Publishing Co., Amsterdam; Kodansha, Ltd., Tokyo, second edition, 1989.

[JS17] Paul A. Jenkins and Dario Spanò. Exact simulation of the Wright—Fisher diffusion. *Ann. Appl. Probab.*, 27(3):1478 – 1509, 2017.

[Kha12] Rafail Khasminskii. *Stochastic stability of differential equations*, volume 66 of *Stochastic Modelling and Applied Probability*. Springer, Heidelberg, second edition, 2012. With contributions by G. N. Milstein and M. B. Nevelson.

[Kim64] Motoo Kimura. Diffusion models in population genetics. *J. Appl. Probability*, 1:177–232, 1964.

[KN97] Stephen M Krone and Claudia Neuhauser. Ancestral processes with selection. *Theor. Popul. Biol.*, 51(3):210–237, 1997.

[KPR21] Guillaume Kon Kam King, Omiros Papaspiliopoulos, and Matteo Ruggiero. Exact inference for a class of hidden Markov models on general state spaces. *Electron. J. Stat.*, 15(1):2832 – 2875, 2021.

[KT81] Samuel Karlin and Howard M. Taylor. *A second course in stochastic processes*. Academic Press, Inc. [Harcourt Brace Jovanovich, Publishers], New York-London, 1981.

[Kut04] Yury A. Kutoyants. *Statistical inference for ergodic diffusion processes*. Springer Series in Statistics. Springer-Verlag London, Ltd., London, 2004.

[LC60] Lucien Le Cam. Locally asymptotically normal families of distributions. Certain approximations to families of distributions and their use in the theory of estimation and testing hypotheses. *Univ. California Publ. Statist.*, 3:37–98, 1960.

[LL08] E. Löcherbach and D. Loukianova. On Nummelin splitting for continuous time Harris recurrent Markov processes and application to kernel estimation for

multi-dimensional diffusions. *Stochastic Process. Appl.*, 118:1301–1321, 2008.

[LL13] Eva Löcherbach and Dasha Loukianova. Polynomial deviation bounds for recurrent Harris processes having general state space. *ESAIM Probab. Stat.*, 17:195–218, 2013.

[LLL11] Eva Löcherbach, Dasha Loukianova, and Oleg Loukianov. Polynomial bounds in the ergodic theorem for one-dimensional diffusions and integrability of hitting times. *Ann. Inst. Henri Poincaré Probab. Stat.*, 47(2):425–449, 2011.

[LPR+09] Arne Ludwig, Melanie Pruvost, Monika Reissmann, Norbert Benecke, Gudrun A. Brockmann, Pedro Castaños, Michael Cieslak, Sebastian Lippold, Laura Llorente, Anna-Sapfo Malaspinas, Montgomery Slatkin, and Michael Hofreiter. Coat color variation at the beginning of horse domestication. *Science*, 324(5926):485–485, 2009.

[Mat20] Iain Mathieson. Estimating time-varying selection coefficients from time series data of allele frequencies. *bioRxiv*, 2020.

[MM13] Iain Mathieson and Gil McVean. Estimating selection coefficients in spatially structured populations from time series data of allele frequencies. *Genetics*, 193(3):973–984, 2013.

[MMES12] Anna-Sapfo Malaspinas, Orestis Malaspinas, Steven N. Evans, and Montgomery Slatkin. Estimating allele age and selection coefficient from time-serial data. *Genetics*, 192(2):599–607, 2012.

[NK97] Claudia Neuhauser and Stephen M Krone. The genealogy of samples in models with selection. *Genetics*, 145(2):519–534, 1997.

[NR20] Richard Nickl and Kolyan Ray. Nonparametric statistical inference for drift vector fields of multi-dimensional diffusions. *Ann. Statist.*, 48(3):1383–1408, 2020.

[NS17] Richard Nickl and Jakob Söhl. Nonparametric Bayesian posterior contraction rates for discretely observed scalar diffusions. *Ann. Statist.*, 45(4):1664–1693, 2017.

[PH06a] Pleuni S Pennings and Joachim Hermisson. Soft sweeps iii: The signature of positive selection from recurrent mutation. *PLOS Genetics*, 2(12):1–15, 12 2006.

[PH06b] Pleuni S. Pennings and Joachim Hermisson. Soft Sweeps II—Molecular Population Genetics of Adaptation from Recurrent Mutation or Migration. *Mol. Biol. Evol*, 23(5):1076–1084, 03 2006.

[PR08] Stefano Peluchetti and Gareth O Roberts. An empirical study of the efficiency of EA for diffusion simulation. 2008.

[PSB19] Cyriel Paris, Bertrand Servin, and Simon Boitard. Inference of selection from genetic time series using various parametric approximations to the Wright–Fisher model. *G3-GENES GENOM GENET*, 9(12):4073–4086, 2019.

[PvZ09] Laura Panzar and Harry van Zanten. Nonparametric Bayesian inference for ergodic diffusions. *J. Statist. Plann. Inference*, 139(12):4193–4199, 2009.

[PY81] J. Pitman and M. Yor. Bessel processes and infinitely divisible laws. In D. Williams, editor, *Stochastic Integrals*, volume 851 of *Lecture Notes in Mathematics*, pages 285–370. Springer-Verlag, 1981.

[SBS14] Matthias Steinrücken, Anand Bhaskar, and Yun S. Song. A novel spectral method for inferring general diploid selection from time series genetic data. *Ann. Appl. Stat.*, 8(4):2203–2222, 2014.

[SCWGSL17] Edson Sandoval-Castellanos, Saskia Wutke, Constantino Gonzalez-Salazar, and Arne Ludwig. Coat colour adaptation of post-glacial horses to increasing forest vegetation. *Nat. Ecol. Evol.*, 1(12):1816–1819, 2017.

[SES16] Joshua G. Schraiber, Steven N. Evans, and Montgomery Slatkin. Bayesian inference of natural selection from allele frequency time series. *Genetics*, 203(1):493–511, 2016.

[Shi76] N. Shimakura. Équations différentielles provenant de la génétique des populations. *Tohoku Math. J. (2)*, 29:287–318, 1976.

[SJKS20] Jaromir Sant, Paul A. Jenkins, Jere Koskela, and Dario Spanò. Convergence of likelihood ratios and estimators for selection in non-neutral Wright–Fisher diffusions. *Scand. J. Stat.*, 2020. *To appear.*

[SPR+13] Giorgos Sermaidis, Omiros Papaspiliopoulos, Gareth O. Roberts, Alexandros Beskos, and Paul Fearnhead. Markov Chain Monte Carlo for exact inference for diffusions. *Scand. J. Stat.*, 40(2):294–321, 2013.

[SS12] Yun S Song and Matthias Steinrücken. A simple method for finding explicit analytic transition densities of diffusion processes with general diploid selection. *Genetics*, 190(3):1117–1129, 2012.

[Tav84] S. Tavaré. Lines-of-descent and genealogical processes, and their applications in population genetics models. *Advances in Applied Probability*, 16:27–27, 1984.

[vdMvZ13] Frank van der Meulen and Harry van Zanten. Consistent nonparametric Bayesian inference for discretely observed scalar diffusions. *Bernoulli*, 19(1):44–63, 2013.

[vZ01] J. H. van Zanten. A note on consistent estimation of multivariate parameters in ergodic diffusion models. *Scand. J. Statist.*, 28(4):617–623, 2001.

[W+16] Saskia Wutke et al. Spotted phenotypes in horses lost attractiveness in the middle ages. *Scientific Reports*, 6, 2016.

[Wat79] G. A. Watterson. Estimating and testing selection: the two-alleles, genetic selection diffusion model. *Adv. in Appl. Probab.*, 11(1):14–30, 1979.

[Wri31] Sewall Wright. Evolution in Mendelian populations. *Genetics*, 16(2):97–159, 1931.

[Wri49] S. Wright. Adaptation and selection. In *Genetics, Paleontology and Evolution (G. L. Jepson, E. Mayr, and G. G. Simpson, eds.)*, pages 365 – 389. Princeton Univ. Press, 1949.

[WS99] Ellen G. Williamson and Montgomery Slatkin. Using maximum likelihood to estimate population size from temporal changes in allele frequencies. *Genetics*, 152(2):755–761, 1999.

[WY08] Huiqing Wang and Chuancun Yin. Moments of the first passage time of one-dimensional diffusion with two-sided barriers. *Statist. Probab. Lett.*, 78(18):3373–3380, 2008.

[ŽSSS15] Daniel Živković, Matthias Steinrücken, Yun S. Song, and Wolfgang Stephan. Transition densities and sample frequency spectra of diffusion processes with selection and variable population size. *Genetics*, 200(2):601–617, 2015.

Appendix A

Proposal updates for when $t_c = t_2, t_3$

Here we illustrate the update procedures for when $t_c = t_2$ or $t_c = t_3$. In these cases, the update procedure is split into separate cases depending on where the proposed allele age falls in relation to the first couple of observation times. We point out that the update mechanisms illustrated below account only for updating the latent path at the observation times and skeleton points once a proposal for t_0 is drawn as in Step 1 in subsection 4.3.3. In an actual initial path update we would need to first propose an allele age and then apply the correct procedure depending on the proposed value for t_0.

Note that proposed latent path values at the observation times and skeleton points will be denoted with a tilde, whilst (in slight abuse of notation) the same quantity without a tilde will denote the current value of the chain. So for instance $\tilde{X}_{t_1}$ will denote a proposed value for the latent diffusion at time t_1, whilst X_{t_1} will mean the value the latent diffusion at time t_1 admits in the current iteration of the chain.

A.1 $t_c = t_2$

Assume now that $X_{t_3} = x_3$ is fixed, and observe that in this case the time interval we need to consider is $[t_0, t_3]$, and that we need to consider separately the cases when $t_0 \in (-\infty, t_1)$ and when $t_0 \in (t_1, t_2)$.

A.1.1 Proposal mechanism and likelihood contribution when $t_0 \in (-\infty, t_1)$

In this case, using the same calculations as those used to derive (4.14), we have that the joint density of the data $\mathbf{Y}_{t_1:t_2}$, the allele age t_0, the latent path values at the observation times $\mathbf{X}_{t_1:t_2}$ and corresponding skeleton points $\{\Phi_k\}_{k=1}^3$ is given by

$$\begin{aligned} & p_2(t_0)(1-X_{t_1})^{n_{t_1}} \binom{n_{t_2}}{Y_{t_2}} X_{t_2}^{Y_{t_2}} (1-X_{t_2})^{n_{t_2}-Y_{t_2}} \\ & \quad \times e^{\frac{\sigma}{2}x_3 - \varphi_\sigma^-(t_3-t_0)} \\ & \quad \times \prod_{i=1}^{3} \prod_{\substack{\{j:\xi_{i,j}\le\lambda_\sigma,\\ \psi_{1,j}<t_1-t_0\}}} \frac{\varphi_\sigma^+ - \varphi_\sigma(\omega_{i,\psi_{i,j}})}{\varphi_\sigma^+ - \varphi_\sigma^-} \end{aligned} \tag{A.1}$$

with respect to the dominating measure

$$\mathbb{PP} \otimes \mathbb{PP}^{(t_3-t_1)} \otimes \mathbb{WF}_{0,\boldsymbol{\theta}}^{(0)} \otimes \mathrm{Leb}((-\infty, t_c) \times [0,1]^2) \otimes \Sigma(\otimes_{i=1}^2 n_{t_i})$$

where as before $\mathbb{PP}$ is the law of a unit rate Poisson point process on $(0,\infty)^2$, $\mathbb{PP}^{(t_3-t_1)}$ is the law of a unit rate Poisson point process on $(t_1, t_3) \times (0,\infty)$, $\mathbb{WF}_{0,\boldsymbol{\theta}}^{(0)}$ is the law of a neutral Wright–Fisher diffusion started from 0, and $\Sigma(\otimes_{i=1}^2 n_{t_i})$ is the counting measure on $\{0,\ldots,n_{t_1}\} \otimes \{0,\ldots,n_{t_2}\}$.

Given the above likelihood contribution, we employ the following procedure to generate a proposal:

1. Conditional on $t_0 = t$, draw $U_1 \sim \mathrm{Unif}([0,1])$ and

$$\tilde{X}_{t_1} \sim \mathbb{WF}_{\sigma,\boldsymbol{\theta}}^{(0,t_1-t)},$$

(i.e. draw $\tilde{X}_{t_1}$ from the law of a non-neutral Wright–Fisher diffusion started at 0, sampled at time $t_1 - t$) and check whether $U_1 < (1-\tilde{X}_{t_1})^{n_{t_1}}$. If this is true, proceed to 2, otherwise redraw.

2. Conditional on $\tilde{X}_{t_1} = x$, draw $U_2 \sim \text{Unif}([0,1])$

$$\tilde{X}_{t_2} \sim \mathbb{WF}_{\sigma,\boldsymbol{\theta}}^{(t_3-t_1,x,x_3,t_2-t_1)},$$

(i.e. draw $\tilde{X}_{t_2}$ from the law of a non-neutral Wright–Fisher diffusion bridge starting at x and ending at x_3 in time $t_3 - t_1$, sampled at time $t_2 - t_1$) and check if $U_2 < \frac{1}{M_2}\tilde{X}_{t_2}^{Y_{t_2}}(1-\tilde{X}_{t_2})^{n_{t_2}-Y_{t_2}}$ with $M_2 := \sup_{z\in[0,1]} z^{Y_{t_2}}(1-z)^{n_{t_2}-Y_{t_2}}$. If this is true, proceed to 3, otherwise redraw.

3. Conditional on $t_0 = t$, $\tilde{X}_{t_1} = x$

$$\begin{aligned}
\kappa_1 &\sim \text{Pois}(\lambda_{\max}(t_1 - t)),\\
\{\tilde{\psi}_{1,j}\}_{j=1}^{\kappa_1} &\sim_{iid} \text{Unif}((t, t_1)),\\
\{\tilde{\xi}_{1,j}\}_{j=1}^{\kappa_1} &\sim_{iid} \text{Unif}((0, \lambda_{\max})),\\
\{\tilde{\gamma}_{1,j}\}_{j=1}^{\kappa_1} &\sim_{iid} \text{Unif}([0,1]),\\
\tilde{\omega}_1 &\sim \mathbb{WF}_{0,\boldsymbol{\theta}}^{(t_1-t,0,x)}
\end{aligned}$$

4. Conditional on $\tilde{X}_{t_1} = x, \tilde{X}_{t_2} = y$, draw

$$\begin{aligned}
\kappa_2 &\sim \text{Pois}(\lambda_{\max}(t_2 - t_1)),\\
\{\tilde{\psi}_{2,j}\}_{j=1}^{\kappa_2} &\sim_{iid} \text{Unif}((t_1, t_2)),\\
\{\tilde{\xi}_{2,j}\}_{j=1}^{\kappa_2} &\sim_{iid} \text{Unif}((0, \lambda_{\max})),\\
\{\tilde{\gamma}_{2,j}\}_{j=1}^{\kappa_2} &\sim_{iid} \text{Unif}([0,1]),\\
\tilde{\omega}_2 &\sim \mathbb{WF}_{0,\boldsymbol{\theta}}^{(t_2-t_1,x,y)}
\end{aligned}$$

5. Conditional on $\tilde{X}_{t_2} = y$, draw

$$\begin{aligned}
\kappa_3 &\sim \text{Pois}(\lambda_{\max}(t_3 - t_2)),\\
\{\tilde{\psi}_{3,j}\}_{j=1}^{\kappa_3} &\sim_{iid} \text{Unif}((t_2, t_3)),\\
\{\tilde{\xi}_{3,j}\}_{j=1}^{\kappa_3} &\sim_{iid} \text{Unif}((0, \lambda_{\max})),\\
\{\tilde{\gamma}_{3,j}\}_{j=1}^{\kappa_3} &\sim_{iid} \text{Unif}([0,1]),\\
\tilde{\omega}_3 &\sim \mathbb{WF}_{0,\boldsymbol{\theta}}^{(t_3-t_2,y,x_3)}
\end{aligned}$$

6. If

$$\prod_{i=1}^{3} \prod_{\substack{\{j:\tilde{\xi}_{i,j} \le \lambda_\sigma, \\ \tilde{\psi}_{1,j} < t_1 - t\}}} 1\left\{ \frac{\varphi_\sigma(\tilde{\omega}_{i,\tilde{\psi}_{i,j}}) - \varphi_\sigma^-}{\varphi_\sigma^+ - \varphi_\sigma^-} < \tilde{\gamma}_{i,j} \right\} = 1,$$

set $\tilde{\Psi}_k = \{\tilde{\psi}_{k,j}\}$, $\tilde{\Xi}_k = \{\tilde{\xi}_{k,j}\}$, $\tilde{\omega}_k^\Psi = \{\tilde{\omega}_{k,\psi_{1,j}}\}$, $\tilde{\Phi}_k = (\tilde{\Psi}_k, \tilde{\Xi}_k, \tilde{\omega}_k^\Psi)$ for $k = 1, 2, 3$ and proceed, else go back to 3.

7. Compute α as in (A.5), and run a Metropolis–Hastings accept-reject step. If we accept, set $\Psi_k = \tilde{\Psi}_k$, $\Xi_k = \tilde{\Xi}_k$, $\omega_k^\Psi = \tilde{\omega}_k^\Psi$, $\Phi_k = \tilde{\Phi}_k$ for $k = 1, 2, 3$ and $X_{t_1} = \tilde{X}_{t_1}, X_{t_2} = \tilde{X}_{t_2}$, or else retain the old values.

A proposal generated via the above mechanism has density given by

$$\begin{aligned} q_2(t)(1 - \tilde{X}_{t_1})^{n_{t_1}} \tilde{X}_{t_2}^{Y_{t_2}} (1 - \tilde{X}_{t_2})^{n_{t_2} - Y_{t_2}} \frac{e^{\frac{\sigma}{2} x_3 - \varphi_\sigma^-(t_3 - t)}}{p_\sigma^{\boldsymbol{\theta}}(t_3 - t_1, \tilde{X}_{t_1}, x_3)} \\ \times \prod_{i=1}^{3} \prod_{\substack{\{j:\tilde{\xi}_{i,j} \le \lambda_\sigma, \\ \tilde{\psi}_{1,j} < t_1 - t_0\}}} \frac{\varphi_\sigma^+ - \varphi_\sigma(\tilde{\omega}_{i,\tilde{\psi}_{i,j}})}{\varphi_\sigma^+ - \varphi_\sigma^-} \end{aligned} \tag{A.2}$$

with respect to the dominating measure

$$\mathbb{PP} \otimes \mathbb{PP}^{(t_3 - t_1)} \otimes \mathbb{WF}_{0,\boldsymbol{\theta}}^{(0)} \otimes \mathrm{Leb}((-\infty, t_c) \times [0,1]^2).$$

A.1.2 Proposal mechanism and likelihood contribution when $t_0 \in (t_1, t_2)$

We now derive the proposal mechanism for when the allele age falls in between the first and second observation time. So assume that now $t_0 \in (t_1, t_2)$, which implies $X_{t_1} = 0, \Psi_1 = \Xi_1 = \omega_1^\Psi = \emptyset$, $\Phi_1 = (\Psi_1, \Xi_1, \omega_1^\Psi)$, and leads to the following expression for the joint density of the data Y_{t_2}, the allele age t_0, the latent path value at the second observation

time X_{t_2} and corresponding skeleton points $\{\Phi_k\}_{k=2}^3$

$$p_2(t_0)\binom{n_{t_2}}{Y_{t_2}} X_{t_2}^{Y_{t_2}}(1-X_{t_2})^{n_{t_2}-Y_{t_2}} e^{\frac{\sigma}{2}x_3-\varphi_\sigma^-(t_3-t_0)} \times \prod_{i=2}^{3} \prod_{\substack{\{j:\xi_{i,j}\le\lambda_\sigma,\\ \psi_{2,j}<t_2-t_0\}}} \frac{\varphi_\sigma^+-\varphi_\sigma(\omega_{i,\psi_{i,j}})}{\varphi_\sigma^+-\varphi_\sigma^-} \tag{A.3}$$

with respect to

$$\mathbb{PP}\otimes\mathbb{PP}^{(t_3-t_2)}\otimes\mathbb{WF}_{0,\boldsymbol{\theta}}^{(0)}\otimes\mathrm{Leb}((-\infty,t_c)\times[0,1])\otimes\Sigma(n_{t_2}).$$

In this case, we generate a candidate path via the following procedure

1. Conditional on $t_0=t$, draw $U\sim\mathrm{Unif}([0,1])$ and

   $$\tilde{X}_{t_2}\sim\mathbb{WF}_{0,\boldsymbol{\theta}}^{(t_3-t,0,x_3,t_2-t)}$$

   and check $U<\frac{1}{M}\tilde{X}_{t_2}^{Y_{t_2}}(1-\tilde{X}_{t_2})^{n_{t_2}-Y_{t_2}}$ with $M:=\sup_{z\in[0,1]}z^{Y_{t_2}}(1-z)^{n_{t_2}-Y_{t_2}}$. If this last condition is true, proceed, else redraw.

2. Conditional on $t_0=t$, $\tilde{X}_{t_2}=x$, draw

   $$\begin{aligned}\kappa_2&\sim\mathrm{Pois}(\lambda_{\max}(t_2-t)),\\ \{\tilde{\psi}_{2,j}\}_{j=1}^{\kappa_2}&\sim_{iid}\mathrm{Unif}((t,t_2)),\\ \{\tilde{\xi}_{2,j}\}_{j=1}^{\kappa_2}&\sim_{iid}\mathrm{Unif}((0,\lambda_{\max})),\\ \{\tilde{\gamma}_{2,j}\}_{j=1}^{\kappa_2}&\sim_{iid}\mathrm{Unif}([0,1]),\\ \tilde{\omega}_2&\sim\mathbb{WF}_{0,\boldsymbol{\theta}}^{(t_2-t,0,x)}\end{aligned}$$

   and

   $$\begin{aligned}\kappa_3&\sim\mathrm{Pois}(\lambda_{\max}(t_3-t_2)),\\ \{\tilde{\psi}_{3,j}\}_{j=1}^{\kappa_3}&\sim_{iid}\mathrm{Unif}((t_2,t_3)),\\ \{\tilde{\xi}_{3,j}\}_{j=1}^{\kappa_3}&\sim_{iid}\mathrm{Unif}((0,\lambda_{\max})),\\ \{\tilde{\gamma}_{3,j}\}_{j=1}^{\kappa_3}&\sim_{iid}\mathrm{Unif}([0,1]),\\ \tilde{\omega}_3&\sim\mathbb{WF}_{0,\boldsymbol{\theta}}^{(t_3-t_2,x,x_3)}\end{aligned}$$

3. If

$$\prod_{i=2}^{3} \prod_{\substack{\{j:\tilde{\xi}_{i,j}\le\lambda_\sigma,\\ \tilde{\psi}_{2,j}<t_2-t_0\}}} 1_{\left\{\frac{\varphi_\sigma(\tilde{\omega}_{i,\tilde{\psi}_{i,j}})-\varphi_\sigma^-}{\varphi_\sigma^+-\varphi_\sigma^-}<\tilde{\gamma}_{i,j}\right\}} = 1$$

then set $\tilde{\Psi}_k = \{\tilde{\psi}_{k,j}\}$, $\tilde{\Xi}_k = \{\tilde{\xi}_{k,j}\}$, $\tilde{\omega}_k^\Psi = \{\tilde{\omega}_{k,\tilde{\psi}_{k,j}}\}$, $\tilde{\Phi}_k = (\tilde{\Psi}_k, \tilde{\Xi}_k, \tilde{\omega}_k^\Psi)$ for $k = 2, 3$ and proceed, otherwise return to 2.

4. Compute α as in (A.5), and run a Metropolis–Hastings accept-reject step. If we accept, set $\Psi_k = \tilde{\Psi}_k, \Xi_k = \tilde{\Xi}_k, \omega_k^\Psi = \tilde{\omega}_k^\Psi, \Phi_k = \tilde{\Phi}_k$ for $k = 2, 3$, and $X_{t_2} = \tilde{X}_{t_2}$, otherwise retain the old values.

The above generated proposal has density

$$\begin{aligned} q_2(t)\tilde{X}_{t_2}^{Y_{t_2}}(1-\tilde{X}_{t_2})^{n_{t_2}-Y_{t_2}} \frac{e^{\frac{\sigma}{2}x_3-\varphi_\sigma^-(t_3-t)}}{p_\sigma^{\boldsymbol{\theta}}(t_3-t,0,x_3)} \\ \times \prod_{i=2}^{3} \prod_{\substack{\{j:\tilde{\xi}_{i,j}\le\lambda_\sigma,\\ \tilde{\psi}_{2,j}<t_2-t_0\}}} \frac{\varphi_\sigma^+-\varphi_\sigma(\tilde{\omega}_{i,\psi_{i,j}})}{\varphi_\sigma^+-\varphi_\sigma^-} \end{aligned} \tag{A.4}$$

with respect to

$$\mathbb{PP} \otimes \mathbb{PP}^{(t_3-t_2)} \otimes \mathbb{WF}_{0,\boldsymbol{\theta}}^{(0)} \otimes \mathrm{Leb}((-\infty, t_c) \times [0,1]).$$

A.1.3 Acceptance probabilities

Putting (A.1), (A.2), (A.3) and (A.4) together gives the following acceptance probabilities depending on the current and proposed values

of the allele age

$$\alpha = \begin{cases} \min\left\{1, \frac{p_2(\tilde{t}_0)}{p_2(t_0)} \frac{q_2(t_0|\tilde{t}_0)}{q_2(\tilde{t}_0|t_0)} e^{\frac{\sigma}{2}\left(X_{t_1} - \tilde{X}_{t_1}\right)} \right. \\ \qquad \left. \times \frac{a(t_3 - t_1, \tilde{X}_{t_1}, x_3, \sigma)}{a(t_3 - t_1, X_{t_1}, x_3, \sigma)} \frac{p_0^{\boldsymbol{\theta}}(t_3 - t_1, \tilde{X}_{t_1}, x_3)}{p_0^{\boldsymbol{\theta}}(t_3 - t_1, X_{t_1}, x_3)} \right\} & \text{if } t_0, \tilde{t}_0 < t_1 \\ \min\left\{1, \frac{p_2(\tilde{t}_0)}{p_2(t_0)} \frac{q_2(t_0|\tilde{t}_0)}{q_2(\tilde{t}_0|t_0)} e^{\frac{\sigma}{2} X_{t_1} - \varphi_\sigma^-\left(t_1 - \tilde{t}_0\right)} \right. \\ \qquad \left. \times \frac{a(t_3 - \tilde{t}_0, 0, x_3, \sigma)}{a(t_3 - t_1, X_{t_1}, x_3, \sigma)} \frac{p_0^{\boldsymbol{\theta}}(t_3 - \tilde{t}_0, 0, x_3)}{p_0^{\boldsymbol{\theta}}(t_3 - t_1, X_{t_1}, x_3)} \right\} & \text{if } t_0 < t_1, \\ & \text{and } \tilde{t}_0 \in (t_1, t_2) \\ \min\left\{1, \frac{p_2(\tilde{t}_0)}{p_2(t_0)} \frac{q_2(t_0|\tilde{t}_0)}{q_2(\tilde{t}_0|t_0)} e^{-\frac{\sigma}{2} \tilde{X}_{t_1} - \varphi_\sigma^-(t_0 - t_1)} \right. \\ \qquad \left. \times \frac{a(t_3 - t_1, \tilde{X}_{t_1}, x_3, \sigma)}{a(t_3 - t_0, 0, x_3, \sigma)} \frac{p_0^{\boldsymbol{\theta}}(t_3 - t_1, \tilde{X}_{t_1}, x_3)}{p_0^{\boldsymbol{\theta}}(t_3 - t_0, 0, x_3)} \right\} & \text{if } t_0 \in (t_1, t_3), \\ & \text{and } \tilde{t}_0 < t_1 \\ \min\left\{1, \frac{p_2(\tilde{t}_0)}{p_2(t_0)} \frac{q_2(t_0|\tilde{t}_0)}{q_2(\tilde{t}_0|t_0)} e^{-\varphi_\sigma^-\left(t_0 - \tilde{t}_0\right)} \right. \\ \qquad \left. \times \frac{a(t_3 - \tilde{t}_0, 0, x_3, \sigma)}{a(t_3 - t_0, 0, x_3, \sigma)} \frac{p_0^{\boldsymbol{\theta}}(t_3 - \tilde{t}_0, 0, x_3)}{p_0^{\boldsymbol{\theta}}(t_3 - t_0, 0, x_3)} \right\} & \text{if } t_0, \tilde{t}_0 \in (t_1, t_2) \end{cases} \tag{A.5}$$

As in subsection 4.3.3, the problematic terms are the intractable terms of the form $a(t, x, y, \sigma)$, as the neutral transition densities can be dealt with using a refinement scheme. Using a pseudo-marginal implementation, we need only estimate these quantities unbiasedly via the Poisson estimator (4.21) for a new proposal, whilst the old estimates are stored and recycled as detailed in the discussion following (4.21). In this case we draw

$$\kappa \sim \begin{cases} \text{Pois}(\lambda_\sigma(t_3 - t_1)) & \text{if } \tilde{t}_0 < t_1 \\ \text{Pois}(\lambda_\sigma(t_3 - \tilde{t}_0)) & \text{if } \tilde{t}_0 \in (t_1, t_2) \end{cases}$$

$$\tau \sim \begin{cases} \text{Unif}((t_1, t_3)) & \text{if } \tilde{t}_0 < t_1 \\ \text{Unif}((\tilde{t}_0, t_3)) & \text{if } \tilde{t}_0 \in (t_1, t_2) \end{cases}$$

$$\zeta \sim \begin{cases} \mathbb{WF}_{0,\boldsymbol{\theta}}^{(t_3 - t_1, \tilde{X}_{t_1}, x_3)} & \text{if } \tilde{t}_0 < t_1 \\ \mathbb{WF}_{0,\boldsymbol{\theta}}^{(t_3 - \tilde{t}_0, 0, x_3)} & \text{if } \tilde{t}_0 \in (t_1, t_2) \end{cases}$$

and then compute

$$\begin{array}{ll} \tilde{a}(t_3 - t_1, \tilde{X}_{t_1}, x_3, \sigma, \tau, \zeta_\tau) & \text{if } \tilde{t}_0 < t_1 \\ \tilde{a}(t_3 - \tilde{t}_0, 0, x_3, \sigma, \tau, \zeta_\tau) & \text{if } \tilde{t}_0 \in (t_1, t_2). \end{array} \tag{A.6}$$

We replace the terms of the form $a(t, x, y, \sigma)$ in (A.5) by the corresponding $\tilde{a}(t, x, y, \sigma, \tau, \zeta_\tau)$ term from (A.6), and accordingly store

$$(\tau^{(k+1)}, \zeta_\tau^{(k+1)}) = \begin{cases} (\tau, \zeta_\tau) & \text{if we accept the corresponding proposal} \\ (\tau^{(k)}, \zeta_\tau^{(k)}) & \text{otherwise.} \end{cases}$$

A.2 Procedure when $t_c = t_3$

Finally we fix $X_{t_4} = x_4$, and set $[t_0, t_4]$ as the initial path segment observation time interval. We now have three separate cases we need to consider: when $t_0 \in (-\infty, t_1)$, $t_0 \in (t_1, t_2)$, $t_0 \in (t_2, t_3)$.

A.2.1 Proposal mechanism and likelihood contribution when $t_0 \in (-\infty, t_1)$

For $t_0 < t_1$, we have that the joint density of the data $\mathbf{Y}_{t_1:t_3}$, the allele age t_0, the latent path at the observation times $\mathbf{X}_{t_1:t_3}$ and corresponding skeleton points $\{\Phi_k\}_{k=1}^4$ is given by

$$\begin{aligned} & p_2(t_0) \prod_{i=1}^{3} \left[\binom{n_{t_i}}{Y_{t_i}} X_{t_i}^{Y_{t_i}} (1 - X_{t_i})^{n_{t_i} - Y_{t_i}} \right] e^{\frac{\sigma}{2} x_4 - \varphi_\sigma^-(t_4 - t_0)} \\ & \times \prod_{i=1}^{4} \prod_{\substack{\{j : \xi_{i,j} \le \lambda_\sigma, \\ \psi_{1,j} < t_1 - t_0\}}} \frac{\varphi_\sigma^+ - \varphi_\sigma(\omega_{i,\psi_{i,j}})}{\varphi_\sigma^+ - \varphi_\sigma^-} \end{aligned} \tag{A.7}$$

with respect to

$$\mathbb{PP} \otimes \mathbb{PP}^{(t_4-t_1)} \otimes \mathbb{WF}^{(0)}_{0,\boldsymbol{\theta}} \otimes \mathrm{Leb}((-\infty, t_c) \times [0,1]^3) \otimes \Sigma(\otimes_{i=1}^3 n_{t_i}),$$

where we have that $\Sigma(\otimes_{i=1}^3 n_{t_i})$ denotes the counting measure over $\otimes_{i=1}^3 \{0, \ldots, n_{t_i}\}$.

The proposal mechanism for this case is as follows

1. Conditional on $t_0 = t$, draw $U_1 \sim \mathrm{Unif}([0,1])$ and

   $$\tilde{X}_{t_1} \sim \mathbb{WF}^{(0,t_1-t)}_{\sigma,\boldsymbol{\theta}}$$

   (i.e. draw $\tilde{X}_{t_1}$ from the law of a non-neutral Wright–Fisher diffusion started at 0, sampled at time $t_1 - t$), and check whether $U_1 < (1 - \tilde{X}_{t_1})^{n_{t_1}}$. If this is true we continue, else we redraw.

2. Conditional on $\tilde{X}_{t_1} = x$, draw $U_2 \sim \mathrm{Unif}([0,1])$ and

   $$\tilde{X}_{t_2} \sim \mathbb{WF}^{(x,t_2-t_1)}_{\sigma,\boldsymbol{\theta}}$$

   (i.e. draw $\tilde{X}_{t_2}$ from the law of a non-neutral Wright–Fisher diffusion started at x, sampled at time $t_2 - t_1$), and check whether $U_2 < (1 - \tilde{X}_{t_2})^{n_{t_2}}$. If this is true we continue, else we redraw.

3. Conditional on $\tilde{X}_{t_2} = y$, draw $U_3 \sim \mathrm{Unif}([0,1])$ and

   $$\tilde{X}_{t_3} \sim \mathbb{WF}^{(t_4-t_2,y,x_4,t_3-t_2)}_{\sigma,\boldsymbol{\theta}}$$

   (i.e. draw $\tilde{X}_{t_3}$ from the law of a non-neutral Wright–Fisher diffusion bridge started at y and ending at x_4 in time $t_4 - t_2$, sampled at time $t_3 - t_2$), and check whether $U_3 < \frac{1}{M_3} \tilde{X}_{t_3}^{Y_{t_3}} (1 - \tilde{X}_{t_3})^{n_{t_3} - Y_{t_3}}$ with $M_3 := \sup_{z \in [0,1]} z^{Y_{t_3}} (1-z)^{n_{t_3} - Y_{t_3}}$. If this is true we continue, else we redraw.

4. Conditional on $t_0 = t$, $\tilde{X}_{t_1} = x$, draw

$$\begin{aligned}
\kappa_1 &\sim \text{Pois}(\lambda_{\max}(t_1 - t)), \\
\{\tilde{\psi}_{1,j}\}_{j=1}^{\kappa_1} &\sim_{iid} \text{Unif}((t, t_1)), \\
\{\tilde{\xi}_{1,j}\}_{j=1}^{\kappa_1} &\sim_{iid} \text{Unif}((0, \lambda_{\max})), \\
\{\tilde{\gamma}_{1,j}\}_{j=1}^{\kappa_1} &\sim_{iid} \text{Unif}([0,1]), \\
\tilde{\omega}_1 &\sim \mathbb{WF}_{0,\boldsymbol{\theta}}^{(t_1 - t, 0, x)}
\end{aligned}$$

5. Conditional on $\tilde{X}_{t_1} = x, \tilde{X}_{t_2} = y$, draw

$$\begin{aligned}
\kappa_2 &\sim \text{Pois}(\lambda_{\max}(t_2 - t_1)), \\
\{\tilde{\psi}_{2,j}\}_{j=1}^{\kappa_2} &\sim_{iid} \text{Unif}((t_1, t_2)), \\
\{\tilde{\xi}_{2,j}\}_{j=1}^{\kappa_2} &\sim_{iid} \text{Unif}((0, \lambda_{\max})), \\
\{\tilde{\gamma}_{2,j}\}_{j=1}^{\kappa_2} &\sim_{iid} \text{Unif}([0,1]), \\
\tilde{\omega}_2 &\sim \mathbb{WF}_{0,\boldsymbol{\theta}}^{(t_2 - t_1, x, y)}
\end{aligned}$$

6. Conditional on $\tilde{X}_{t_2} = y, \tilde{X}_{t_3} = z$, draw

$$\begin{aligned}
\kappa_3 &\sim \text{Pois}(\lambda_{\max}(t_3 - t_2)), \\
\{\tilde{\psi}_{3,j}\}_{j=1}^{\kappa_3} &\sim_{iid} \text{Unif}((t_2, t_3)), \\
\{\tilde{\xi}_{3,j}\}_{j=1}^{\kappa_3} &\sim_{iid} \text{Unif}((0, \lambda_{\max})), \\
\{\tilde{\gamma}_{3,j}\}_{j=1}^{\kappa_3} &\sim_{iid} \text{Unif}([0,1]), \\
\tilde{\omega}_3 &\sim \mathbb{WF}_{0,\boldsymbol{\theta}}^{(t_3 - t_2, y, z)}
\end{aligned}$$

7. Conditional on $\tilde{X}_{t_3} = z$, draw

$$\begin{aligned}
\kappa_4 &\sim \text{Pois}(\lambda_{\max}(t_4 - t_3)), \\
\{\tilde{\psi}_{4,j}\}_{j=1}^{\kappa_4} &\sim_{iid} \text{Unif}((t_3, t_4)), \\
\{\tilde{\xi}_{4,j}\}_{j=1}^{\kappa_4} &\sim_{iid} \text{Unif}((0, \lambda_{\max})), \\
\{\tilde{\gamma}_{4,j}\}_{j=1}^{\kappa_4} &\sim_{iid} \text{Unif}([0,1]), \\
\tilde{\omega}_4 &\sim \mathbb{WF}_{0,\boldsymbol{\theta}}^{(t_4 - t_3, z, x_4)}
\end{aligned}$$

8. If

$$\prod_{i=1}^{4} \prod_{\substack{\{j:\tilde{\xi}_{i,j} \le \lambda_\sigma, \\ \tilde{\psi}_{1,j} < t_1 - t_0\}}} 1\left\{ \frac{\varphi_\sigma(\tilde{\omega}_{i,\tilde{\psi}_{i,j}}) - \varphi_\sigma^-}{\varphi_\sigma^+ - \varphi_\sigma^-} < \tilde{\gamma}_{i,j} \right\} = 1$$

set $\tilde{\Psi}_k = \{\tilde{\psi}_{k,j}\}$, $\tilde{\Xi}_k = \{\tilde{\xi}_{k,j}\}$, $\tilde{\omega}_k^\Psi = \{\tilde{\omega}_{k,\tilde{\psi}_{k,j}}\}$, $\tilde{\Phi}_k = (\tilde{\Psi}_k, \tilde{\Xi}_k, \tilde{\omega}_k^\Psi)$ for $k = 1, 2, 3, 4$ and proceed, otherwise go back to 4.

9. Compute α as in (A.14), and run a Metropolis–Hastings accept-reject step. If we accept, set $\Psi_k = \tilde{\Psi}_k$, $\Xi_k = \tilde{\Xi}_k$, $\omega_k^\Psi = \tilde{\omega}_k^\Psi$, $\Phi_k = \tilde{\Phi}_k$, for $k = 1, 2, 3, 4$, and $X_{t_i} = \tilde{X}_{t_i}$ for $i = 1, 2, 3$, otherwise retain the old values.

The resulting proposal has density given by

$$\begin{aligned} & q_2(t)(1 - \tilde{X}_{t_1})^{n_{t_1}} (1 - \tilde{X}_{t_2})^{n_{t_2}} \tilde{X}_{t_3}^{Y_{t_3}} (1 - \tilde{X}_{t_3})^{n_{t_3} - Y_{t_3}} \\ & \quad \times \frac{e^{\frac{\sigma}{2} x_4 - \varphi_\sigma^-(t_4 - t)}}{p_\sigma^{\boldsymbol{\theta}}(t_4 - t_2, \tilde{X}_{t_2}, x_4)} \\ & \quad \times \prod_{i=1}^{4} \prod_{\substack{\{j:\tilde{\xi}_{i,j} \le \lambda_\sigma, \\ \tilde{\psi}_{1,j} < t_1 - t_0\}}} \frac{\varphi_\sigma^+ - \varphi_\sigma(\tilde{\omega}_{i,\tilde{\psi}_{i,j}})}{\varphi_\sigma^+ - \varphi_\sigma^-} \end{aligned} \tag{A.8}$$

with respect to

$$\mathbb{PP} \otimes \mathbb{PP}^{(t_4 - t_1)} \otimes \mathbb{WF}_{0,\boldsymbol{\theta}}^{(0)} \otimes \mathrm{Leb}((-\infty, t_c)) \times [0, 1]^3).$$

A.2.2 Proposal mechanism and likelihood contribution when $t_0 \in (t_1, t_2)$

In the case when $t_0 \in (t_1, t_2)$, we set $X_{t_1} = 0, \Psi_1 = \Xi_1 = \omega_1^\Psi = \emptyset$, $\Phi_1 = (\Psi_1, \Xi_1, \omega_1^\Psi)$, and observe that the joint density of the data $\mathbf{Y}_{t_1:t_3}$, the allele age t_0, the latent path at the observation times $\mathbf{X}_{t_2:t_3}$ and

corresponding skeleton points $\{\Phi_k\}_{k=1}^4$ is given by

$$p_2(t_0)(1-X_{t_2})^{n_{t_2}}\binom{n_{t_3}}{Y_{t_3}}X_{t_3}^{Y_{t_3}}(1-X_{t_3})^{n_{t_3}-Y_{t_3}}e^{\frac{\sigma}{2}x_4-\varphi_\sigma^-(t_4-t_0)}$$
$$\times\prod_{i=2}^{4}\prod_{\substack{\{j:\xi_{i,j}\le\lambda_\sigma,\\ \psi_{2,j}<t_2-t_0\}}}\frac{\varphi_\sigma^+-\varphi_\sigma(\omega_{i,\psi_{i,j}})}{\varphi_\sigma^+-\varphi_\sigma^-} \tag{A.9}$$

with respect to

$$\mathbb{PP}\otimes\mathbb{PP}^{(t_4-t_2)}\otimes\mathbb{WF}_{0,\boldsymbol{\theta}}^{(0)}\otimes\mathrm{Leb}((-\infty,t_c)\times[0,1]^2)\otimes\Sigma(\otimes_{i=1}^3 n_{t_i}).$$

To generate a candidate path, we set $X_{t_1}=0, \tilde{\Psi}_1=\tilde{\Xi}_1=\tilde{\omega}_1^\Psi=\emptyset, \tilde{\Phi}_1=(\tilde{\Psi}_1,\tilde{\Xi}_1,\tilde{\omega}_1^\Psi)$, and then

1. Conditional on $t_0=t$, draw $U_2\sim\mathrm{Unif}([0,1])$ and

$$\tilde{X}_{t_2}\sim\mathbb{WF}_{\sigma,\boldsymbol{\theta}}^{(0,t_2-t)}$$

(i.e. draw $\tilde{X}_{t_2}$ from the law of a non-neutral Wright–Fisher diffusion started at 0, sampled at time t_2-t), and check whether $U_2<(1-\tilde{X}_{t_2})^{n_{t_2}}$. If this is true, we proceed, else we redraw.

2. Conditional on $t_0=t, \tilde{X}_{t_2}=x$, draw $U_3\sim\mathrm{Unif}([0,1])$ and

$$\tilde{X}_{t_3}\sim\mathbb{WF}_{\sigma,\boldsymbol{\theta}}^{(t_4-t_2,x,x_4,t_3-t_2)}$$

(i.e. draw $\tilde{X}_{t_3}$ from the law of a non-neutral Wright–Fisher diffusion started at x and ending at x_4 in time t_4-t_2, sampled at t_3-t_2), and again check that $U_3<\frac{1}{M_3}\tilde{X}_{t_3}^{Y_{t_3}}(1-\tilde{X}_{t_3})^{n_{t_3}-Y_{t_3}}$, with $M_3:=\sup_{z\in[0,1]}z^{Y_{t_3}}(1-z)^{n_{t_3}-Y_{t_3}}$. Proceed if true, else redraw.

3. Conditional on $t_0 = t, \tilde{X}_{t_2} = x$, draw

$$\begin{aligned}
\kappa_2 &\sim \text{Pois}(\lambda_{\max}(t_2 - t)), \\
\{\tilde{\psi}_{2,j}\}_{j=1}^{\kappa_2} &\sim_{iid} \text{Unif}((t, t_2)), \\
\{\tilde{\xi}_{2,j}\}_{j=1}^{\kappa_2} &\sim_{iid} \text{Unif}((0, \lambda_{\max})), \\
\{\tilde{\gamma}_{2,j}\}_{j=1}^{\kappa_2} &\sim_{iid} \text{Unif}([0,1]), \\
\tilde{\omega}_2 &\sim \mathbb{WF}_{0,\boldsymbol{\theta}}^{(t_2 - t, 0, x)}
\end{aligned}$$

4. Conditional on $\tilde{X}_{t_2} = x, \tilde{X}_{t_3} = y$, draw

$$\begin{aligned}
\kappa_3 &\sim \text{Pois}(\lambda_{\max}(t_3 - t_2)), \\
\{\tilde{\psi}_{3,j}\}_{j=1}^{\kappa_3} &\sim_{iid} \text{Unif}((t_2, t_3)), \\
\{\tilde{\xi}_{3,j}\}_{j=1}^{\kappa_3} &\sim_{iid} \text{Unif}((0, \lambda_{\max})), \\
\{\tilde{\gamma}_{3,j}\}_{j=1}^{\kappa_3} &\sim_{iid} \text{Unif}([0,1]), \\
\tilde{\omega}_3 &\sim \mathbb{WF}_{0,\boldsymbol{\theta}}^{(t_3 - t_2, x, y)}
\end{aligned}$$

5. Conditional on $\tilde{X}_{t_3} = y$, draw

$$\begin{aligned}
\kappa_4 &\sim \text{Pois}(\lambda_{\max}(t_4 - t_3)), \\
\{\tilde{\psi}_{4,j}\}_{j=1}^{\kappa_4} &\sim_{iid} \text{Unif}((t_3, t_4)), \\
\{\tilde{\xi}_{4,j}\}_{j=1}^{\kappa_4} &\sim_{iid} \text{Unif}((0, \lambda_{\max})), \\
\{\tilde{\gamma}_{4,j}\}_{j=1}^{\kappa_4} &\sim_{iid} \text{Unif}([0,1]), \\
\tilde{\omega}_4 &\sim \mathbb{WF}_{0,\boldsymbol{\theta}}^{(t_4 - t_3, y, x_4)}
\end{aligned}$$

6. If

$$\prod_{i=2}^{4} \prod_{\substack{\{j:\tilde{\xi}_{i,j} \le \lambda_\sigma, \\ \tilde{\psi}_{2,j} < t_2 - t_0\}}} 1\left\{ \frac{\varphi_\sigma(\tilde{\omega}_{i,\tilde{\psi}_{i,j}}) - \varphi_\sigma^-}{\varphi_\sigma^+ - \varphi_\sigma^-} < \tilde{\gamma}_{i,j} \right\} = 1$$

set $\tilde{\Psi}_k = \{\tilde{\psi}_{k,j}\}$, $\tilde{\Xi}_k = \{\tilde{\xi}_{k,j}\}$, $\tilde{\omega}_k^\Psi = \{\tilde{\omega}_{k,\psi_{k,j}}\}$, $\tilde{\Phi}_k = (\tilde{\Psi}_k, \tilde{\Xi}_k, \tilde{\omega}_k^\Psi)$ for $k = 2, 3, 4$, and proceed. Otherwise go back to 3.

7. Compute α as in (A.14), and run a Metropolis–Hastings accept-

reject step. If we accept, set $\Psi_k = \tilde{\Psi}_k$, $\Xi_k = \tilde{\Xi}_k$, $\omega_k^\Psi = \tilde{\omega}_k^\Psi$, $\Phi_k = \tilde{\Phi}_k$ for $k = 2, 3, 4$, and $X_{t_i} = \tilde{X}_{t_i}$ for $i = 2, 3$, otherwise retain the old values.

A proposal generated according to the above leads to the following density

$$q_2(t)(1-\tilde{X}_{t_2})^{n_{t_2}}\tilde{X}_{t_3}^{Y_{t_3}}(1-\tilde{X}_{t_3})^{n_{t_3}-Y_{t_3}}\frac{e^{\frac{\sigma}{2}x_4-\varphi_\sigma^-(t_4-t_0)}}{p_\sigma^{\boldsymbol{\theta}}(t_4-t_2,X_{t_2},x_4)} \times\prod_{i=2}^{4}\prod_{\substack{\{j:\tilde{\xi}_{i,j}\le\lambda_\sigma,\\ \tilde{\psi}_{2,j}<t_2-t_0\}}}\frac{\varphi_\sigma^+-\varphi_\sigma(\tilde{\omega}_{i,\tilde{\psi}_{i,j}})}{\varphi_\sigma^+-\varphi_\sigma^-} \tag{A.10}$$

with respect to

$$\mathbb{PP}\otimes\mathbb{PP}^{(t_4-t_2)}\otimes\mathbb{WF}_{0,\boldsymbol{\theta}}^{(0)}\otimes\mathrm{Leb}((-\infty,t_c)\times[0,1]^2).$$

A.2.3 Proposal mechanism and likelihood contribution when $t_0 \in (t_2, t_3)$

The final case to consider is when $t_0 \in (t_2, t_3)$. We set $X_{t_k} = 0, \Psi_k = \Xi_k = \omega_k^\Psi = \emptyset, \Phi_k = (\Psi_k, \Xi_k, \omega_k^\Psi)$ for $k = 1, 2$. The joint density of the data $\mathbf{Y}_{t_1:t_3}$, the allele age t_0, the latent path at the observation time X_{t_3} and corresponding skeleton points $\{\Phi_k\}_{k=1}^4$ in this case is now

$$p_2(t_0)\binom{n_{t_3}}{Y_{t_3}}X_{t_3}^{Y_{t_3}}(1-X_{t_3})^{n_{t_3}-Y_{t_3}}e^{\frac{\sigma}{2}x_4-\varphi_\sigma^-(t_4-t_0)} \times\prod_{i=3}^{4}\prod_{\substack{\{j:\xi_{i,j}\le\lambda_\sigma,\\ \psi_{3,j}<t_3-t_0\}}}\frac{\varphi_\sigma^+-\varphi_\sigma(\omega_{i,\psi_{i,j}})}{\varphi_\sigma^+-\varphi_\sigma^-} \tag{A.11}$$

with respect to

$$\mathbb{PP}\otimes\mathbb{PP}^{(t_4-t_3)}\otimes\mathbb{WF}_{0,\boldsymbol{\theta}}^{(0)}\otimes\mathrm{Leb}((-\infty,t_c)\times[0,1])\otimes\Sigma(\otimes_{i=1}^3 n_{t_i}).$$

We generate a proposal by setting $\tilde{X}_{t_k} = 0, \tilde{\Psi}_k = \tilde{\Xi}_k = \tilde{\omega}_k^\Psi = \emptyset$ for $k = 1, 2$, and

1. Conditional on $t_0 = t$, draw $U \sim \mathrm{Unif}([0,1])$ and

$$\tilde{X}_{t_3} \sim \mathbb{WF}_{\sigma,\boldsymbol{\theta}}^{(t_4-t,0,x_4,t_3-t)}$$

(i.e. draw $\tilde{X}_{t_3}$ from the law of a non-neutral Wright–Fisher diffusion started at 0 and ending at x_4 in time $t_4 - t$, sampled at $t_3 - t$), and check that $U < \frac{1}{M} X_{t_3}^{Y_{t_3}} (1 - X_{t_3})^{n_{t_3} - Y_{t_3}}$ with $M := \sup_{z \in [0,1]} z^{Y_{t_3}} (1-z)^{n_{t_3} - Y_{t_3}}$. If this condition is met, proceed, else redraw.

2. Conditional on $t_0 = t$ and $\tilde{X}_{t_3} = x$, draw

$$\begin{aligned}
\kappa_3 &\sim \mathrm{Pois}(\lambda_{\max}(t_3 - t)), \\
\{\tilde{\psi}_{3,j}\}_{j=1}^{\kappa_3} &\sim_{iid} \mathrm{Unif}((t, t_3)), \\
\{\tilde{\xi}_{3,j}\}_{j=1}^{\kappa_3} &\sim_{iid} \mathrm{Unif}((0, \lambda_{\max})), \\
\{\tilde{\gamma}_{3,j}\}_{j=1}^{\kappa_3} &\sim_{iid} \mathrm{Unif}([0,1]), \\
\tilde{\omega}_3 &\sim \mathbb{WF}_{0,\boldsymbol{\theta}}^{(t_3-t,0,x)}
\end{aligned}$$

3. Conditional on $\tilde{X}_{t_3} = x$, draw

$$\begin{aligned}
\kappa_4 &\sim \mathrm{Pois}(\lambda_{\max}(t_4 - t_3)), \\
\{\tilde{\psi}_{4,j}\}_{j=1}^{\kappa_4} &\sim_{iid} \mathrm{Unif}((t_3, t_4)), \\
\{\tilde{\xi}_{4,j}\}_{j=1}^{\kappa_4} &\sim_{iid} \mathrm{Unif}((0, \lambda_{\max})), \\
\{\tilde{\gamma}_{4,j}\}_{j=1}^{\kappa_4} &\sim_{iid} \mathrm{Unif}([0,1]), \\
\tilde{\omega}_4 &\sim \mathbb{WF}_{0,\boldsymbol{\theta}}^{(t_4-t_3,x,x_4)}
\end{aligned}$$

4. If

$$\prod_{i=3}^{4} \prod_{\substack{\{j : \tilde{\xi}_{i,j} \le \lambda_\sigma, \\ \tilde{\psi}_{3,j} < t_3 - t_0\}}} 1\left\{ \frac{\varphi_\sigma(\tilde{\omega}_{i,\tilde{\psi}_{i,j}}) - \varphi_\sigma^-}{\varphi_\sigma^+ - \varphi_\sigma^-} < \tilde{\gamma}_{i,j} \right\} = 1$$

then set $\tilde{\Psi}_k = \{\tilde{\psi}_{k,j}\}$, $\tilde{\Xi}_k = \{\tilde{\xi}_{k,j}\}$, $\tilde{\omega}_k^\Psi = \{\tilde{\omega}_{k,\tilde{\psi}_{k,j}}\}$, $\tilde{\Phi}_k = (\tilde{\Psi}_k, \tilde{\Xi}_k, \tilde{\omega}_k^\Psi)$ and proceed, else go back to 2.

5. Compute α as in (A.14), and run a Metropolis–Hastings accept-

reject step. If we accept, set $\Psi_k = \tilde{\Psi}_k, \Xi_k = \tilde{\Xi}_k, \omega_k^{\Psi} = \tilde{\omega}_k^{\Psi}$ for k=3,4 and $X_{t_3} = \tilde{X}_{t_3}$, otherwise retain the old values.

A proposal generated according to the above has density

$$q_2(t)\tilde{X}_{t_3}^{Y_{t_3}}(1-\tilde{X}_{t_3})^{n_{t_3}-Y_{t_3}}\frac{e^{\frac{\sigma}{2}x_4-\varphi_\sigma^-(t_4-t_0)}}{p_\sigma^{\boldsymbol{\theta}}(t_4-t,0,x_4)} \times\prod_{i=3}^{4}\prod_{\substack{\{j:\tilde{\xi}_{i,j}\le\lambda_\sigma,\\ \tilde{\psi}_{3,j}<t_3-t_0\}}}\frac{\varphi_\sigma^+-\varphi_\sigma(\tilde{\omega}_{i,\tilde{\psi}_{i,j}})}{\varphi_\sigma^+-\varphi_\sigma^-} \tag{A.12}$$

with respect to

$$\mathbb{PP}\otimes\mathbb{PP}^{(t_4-t_3)}\otimes\mathbb{WF}_{0,\boldsymbol{\theta}}^{(0)}\otimes\mathrm{Leb}((-\infty,t_c)\times[0,1]).$$

A.2.4 Acceptance Probabilities

Putting (A.7), (A.8), (A.9), (A.10), (A.11) and (A.12) together we get that the corresponding acceptance probabilities for this initial path update, which can be found on the next page. Again we resort to the Poisson estimator which we compute by drawing

$$\kappa\sim\begin{cases}\mathrm{Pois}(\lambda_\sigma(t_4-t_2)) & \text{if } \tilde{t}_0<t_1 \text{ or } \tilde{t}_0\in(t_1,t_2)\\ \mathrm{Pois}(\lambda_\sigma(t_4-\tilde{t}_0)) & \text{if } \tilde{t}_0\in(t_2,t_3)\end{cases}$$

$$\tau\sim\begin{cases}\mathrm{Unif}((t_2,t_4)) & \text{if } \tilde{t}_0<t_1 \text{ or } \tilde{t}_0\in(t_1,t_2)\\ \mathrm{Unif}((\tilde{t}_0,t_4)) & \text{if } \tilde{t}_0\in(t_2,t_3)\end{cases}$$

$$\zeta\sim\begin{cases}\mathbb{WF}_{0,\boldsymbol{\theta}}^{(t_4-t_2,\tilde{X}_{t_2},x_4)} & \text{if } \tilde{t}_0<t_1 \text{ or } \tilde{t}_0\in(t_1,t_2)\\ \mathbb{WF}_{0,\boldsymbol{\theta}}^{(t_4-\tilde{t}_0,0,x_4)} & \text{if } \tilde{t}_0\in(t_2,t_3)\end{cases}$$

and then compute

$$\begin{matrix}\tilde{a}(t_4-t_2,\tilde{X}_{t_2},x_4,\sigma,\tau,\zeta_\tau) & \text{if } \tilde{t}_0<t_1 \text{ or } \tilde{t}_0\in(t_1,t_2)\\ \tilde{a}(t_4-\tilde{t}_0,0,x_4,\sigma,\tau,\zeta_\tau) & \text{if } \tilde{t}_0\in(t_2,t_3).\end{matrix} \tag{A.13}$$

We replace the terms of the form $a(t, x, y, \sigma)$ in (A.14) by the corresponding $\tilde{a}(t, x, y, \sigma, \tau, \zeta_\tau)$ term from (A.13), and accordingly store

$$(\tau^{(k+1)}, \zeta_\tau^{(k+1)}) = \begin{cases} (\tau, \zeta_\tau) & \text{if we accept the corresponding proposal} \\ (\tau^{(k)}, \zeta_\tau^{(k)}) & \text{otherwise.} \end{cases}$$

$$\alpha = \begin{cases} \frac{q_2(t_0|\tilde{t}_0)}{q_2(\tilde{t}_0|t_0)} \frac{p_2(\tilde{t}_0)}{p_2(t_0)} e^{\frac{\sigma}{2}\left(X_{t_2} - \tilde{X}_{t_2}\right)} \\ \quad \times \frac{a(t_4 - t_2, \tilde{X}_{t_2}, x_4, \sigma)}{a(t_4 - t_2, X_{t_2}, x_4, \sigma)} \frac{p_0^{\boldsymbol{\theta}}(t_4 - t_2, \tilde{X}_{t_2}, x_4)}{p_0^{\boldsymbol{\theta}}(t_4 - t_2, X_{t_2}, x_4)} & \text{if } t_0, \tilde{t}_0 < t_1 \\ \frac{q_2(t_0|\tilde{t}_0)}{q_2(\tilde{t}_0|t_0)} \frac{p_2(\tilde{t}_0)}{p_2(t_0)} e^{\frac{\sigma}{2}\left(X_{t_2} - \tilde{X}_{t_2}\right)} \\ \quad \times \frac{a(t_4 - t_2, \tilde{X}_{t_2}, x_4, \sigma)}{a(t_4 - t_2, X_{t_2}, x_4, \sigma)} \frac{p_0^{\boldsymbol{\theta}}(t_4 - t_2, \tilde{X}_{t_2}, x_4)}{p_0^{\boldsymbol{\theta}}(t_4 - t_2, X_{t_2}, x_4)} & \text{if } t_0 < t_1, \\ & \tilde{t}_0 \in (t_1, t_2) \\ \frac{q_2(t_0|\tilde{t}_0)}{q_2(\tilde{t}_0|t_0)} \frac{p_2(\tilde{t}_0)}{p_2(t_0)} e^{\frac{\sigma}{2}\left(X_{t_2}\right) - \varphi_\sigma^-\left(t_2 - \tilde{t}_0\right)} \\ \quad \times \frac{a(t_4 - \tilde{t}_0, 0, x_4, \sigma)}{a(t_4 - t_2, X_{t_2}, x_4, \sigma)} \frac{p_0^{\boldsymbol{\theta}}(t_4 - \tilde{t}_0, 0, x_4)}{p_0^{\boldsymbol{\theta}}(t_4 - t_2, X_{t_2}, x_4)} & \text{if } t_0 < t_1, \\ & \tilde{t}_0 \in (t_2, t_3) \\ \frac{q_2(t_0|\tilde{t}_0)}{q_2(\tilde{t}_0|t_0)} \frac{p_2(\tilde{t}_0)}{p_2(t_0)} e^{\frac{\sigma}{2}\left(X_{t_2} - \tilde{X}_{t_2}\right)} \\ \quad \times \frac{a(t_4 - t_2, \tilde{X}_{t_2}, x_4, \sigma)}{a(t_4 - t_2, X_{t_2}, x_4, \sigma)} \frac{p_0^{\boldsymbol{\theta}}(t_4 - t_2, \tilde{X}_{t_2}, x_4)}{p_0^{\boldsymbol{\theta}}(t_4 - t_2, X_{t_2}, x_4)} & \text{if } t_0 \in (t_1, t_2), \\ & \tilde{t}_0 < t_1 \\ \frac{q_2(t_0|\tilde{t}_0)}{q_2(\tilde{t}_0|t_0)} \frac{p_2(\tilde{t}_0)}{p_2(t_0)} e^{\frac{\sigma}{2}\left(X_{t_2} - \tilde{X}_{t_2}\right)} \\ \quad \times \frac{a(t_4 - t_2, \tilde{X}_{t_2}, x_4, \sigma)}{a(t_4 - t_2, X_{t_2}, x_4, \sigma)} \frac{p_0^{\boldsymbol{\theta}}(t_4 - t_2, \tilde{X}_{t_2}, x_4)}{p_0^{\boldsymbol{\theta}}(t_4 - t_2, X_{t_2}, x_4)} & \text{if } t_0, \tilde{t}_0 \in (t_1, t_2) \\ \frac{q_2(t_0|\tilde{t}_0)}{q_2(\tilde{t}_0|t_0)} \frac{p_2(\tilde{t}_0)}{p_2(t_0)} e^{\frac{\sigma}{2} X_{t_2} - \varphi_\sigma^-\left(t_2 - \tilde{t}_0\right)} \\ \quad \times \frac{a(t_4 - \tilde{t}_0, 0, x_4, \sigma)}{a(t_4 - t_2, X_{t_2}, x_4, \sigma)} \frac{p_0^{\boldsymbol{\theta}}(t_4 - \tilde{t}_0, 0, x_4)}{p_0^{\boldsymbol{\theta}}(t_4 - t_2, X_{t_2}, x_4)} & \text{if } t_0 \in (t_1, t_2), \\ & \tilde{t}_0 \in (t_2, t_3) \\ \frac{q_2(t_0|\tilde{t}_0)}{q_2(\tilde{t}_0|t_0)} \frac{p_2(\tilde{t}_0)}{p_2(t_0)} e^{-\frac{\sigma}{2}\tilde{X}_{t_2} - \varphi_\sigma^-\left(t_0^{(k)} - t_2\right)} \\ \quad \times \frac{a(t_4 - t_2, \tilde{X}_{t_2}, x_4, \sigma)}{a(t_4 - t_0, 0, x_4, \sigma)} \frac{p_0^{\boldsymbol{\theta}}(t_4 - t_2, \tilde{X}_{t_2}, x_4)}{p_0^{\boldsymbol{\theta}}(t_4 - t_0, 0, x_4)} & \text{if } t_0 \in (t_2, t_3), \\ & \tilde{t}_0 < t_1 \\ \frac{q_2(t_0|\tilde{t}_0)}{q_2(\tilde{t}_0|t_0)} \frac{p_2(\tilde{t}_0)}{p_2(t_0)} e^{-\frac{\sigma}{2}\tilde{X}_{t_2} - \varphi_\sigma^-\left(t_0^{(k)} - t_2\right)} \\ \quad \times \frac{a(t_4 - t_2, \tilde{X}_{t_2}, x_4, \sigma)}{a(t_4 - t_0, 0, x_4, \sigma)} \frac{p_0^{\boldsymbol{\theta}}(t_4 - t_2, \tilde{X}_{t_2}, x_4)}{p_0^{\boldsymbol{\theta}}(t_4 - t_0, 0, x_4)} & \text{if } t_0 \in (t_2, t_3), \\ & \tilde{t}_0 \in (t_1, t_2) \\ \frac{q_2(t_0|\tilde{t}_0)}{q_2(\tilde{t}_0|t_0)} \frac{p_2(\tilde{t}_0)}{p_2(t_0)} e^{-\varphi_\sigma^-\left(t_0^{(k)} - \tilde{t}_0\right)} \\ \quad \times \frac{a(t_4 - \tilde{t}_0, 0, x_4, \sigma)}{a(t_4 - t_0, 0, x_4)} \frac{p_0^{\boldsymbol{\theta}}(t_4 - \tilde{t}_0, 0, x_4)}{p_0^{\boldsymbol{\theta}}(t_4 - t_0, 0, x_4)} & \text{if } t_0, \tilde{t}_0 \in (t_2, t_3). \end{cases} \tag{A.14}$$