Data augmentation for diffusions

Omiros Papaspiliopoulos * Gareth O. Roberts † Osnat Stramer ‡

February 12, 2013

Abstract

The problem of formal likelihood-based (either classical or Bayesian) inference for discretely observed multi-dimensional diffusions is particularly challenging. In principle this involves data-augmentation of the observation data to give representations of the entire diffusion trajectory. Most currently proposed methodology splits broadly into two classes: either through the discretisation of idealised approaches for the continuous-time diffusion setup; or through the use of standard finite-dimensional methodologies discretisation of the diffusion model. The connections between these approaches have not been well-studied. This paper will provide a unified framework bringing together these approaches, demonstrating connections, and in some cases surprising differences. As a result, we provide, for the first time, theoretical justification for the various methods of imputing missing data. The inference problems are particularly challenging for reducible diffusions, and our framework is correspondingly more complex in that case. Therefore we treat the reducible and irreducible cases differently within the paper. Supplementary materials for the article are avilable on line.

1 Overview of likelihood-based inference for diffusions

Diffusion processes have gained much popularity as statistical models for observed and latent processes. Among others, their appeal lies in their flexibility to deal with non-linearity, time-inhomogeneity and heteroscedasticity by specifying two interpretable functionals, their amenability to efficient computations due to their Markov property, and the rich existing mathematical theory about their properties. As a result, they are used as models throughout Science; some book references related with this approach to modeling include Section 5.3 of [1] for physical systems, Section 8.3.3 (in conjunction with Section 6.3) of [12] for systems biology and mass action stochastic kinetics, and Chapter 10 of [27] for interest rates.

A mathematically precise specification of a d-dimensional diffusion process V is as the solution of a stochastic differential equation (SDE) of the type:

$$\mathrm{d}V_s = b(s, V_s; \theta_1)\,\mathrm{d}s + \sigma(s, V_s; \theta_2)\,\mathrm{d}B_s, \quad s \in [0, T]\ ; \tag{1}$$

where B is an m-dimensional standard Brownian motion, $b(\cdot, \cdot\,; \cdot) : \mathbb{R}_+ \times \mathbb{R}^d \times \Theta_1 \to \mathbb{R}^d$ is the drift and $\sigma(\cdot, \cdot\,; \cdot) : \mathbb{R}_+ \times \mathbb{R}^d \times \Theta_2 \to \mathbb{R}^{d\times m}$ is the diffusion coefficient. These

*ICREA and Department of Economics, Universitat Pompeu Fabra, omiros.papaspiliopoulos@upf.edu
†Department of Statistics, University of Warwick
‡Department of Statistics and Actuarial Science, University of Iowa, Iowa City, Iowa

functionals are typically specified up to a vector of unknown parameters $\theta = (\theta_1, \theta_2) \in \Theta_1 \times \Theta_2 \subseteq \mathbb{R}^p \times \mathbb{R}^q$. Since it is most common that the drift and diffusion coefficient do not depend on common parameters, we have made this explicit in the notation to simplify the following discussion. Standard conditions on b, σ are needed for (1) to have a unique solution, see for example Theorem 11A of [9]. We define $\Gamma = \sigma\sigma^T$ and we will assume that Γ is invertible for all θ_2, v, s, unless otherwise stated; note that T is used both to index final time and matrix transpose but the distinction between the two should be clear from the context. For notational convenience we define $A = \Gamma^{-1}$ and we occasionally drop the arguments from drift and diffusion functionals, as we have done in this paragraph.

A perspective on diffusions that is less formal but considerably more popular in applications is afforded by a discrete-time approximation of (1). For a given finite collection of time points $0 = \tau_0 < \tau_1 < \cdots < \tau_M = T$, the Euler-Maruyama approximation of (1) defines a non-linear Markovian time-series model with dynamics

$$V_{\tau_{j+1}} = V_{\tau_j} + b(\tau_j, V_{\tau_j}; \theta_1)(\tau_{j+1} - \tau_j) + \sigma(\tau_j, V_{\tau_j}; \theta_2)\sqrt{\tau_{j+1} - \tau_j}\,\epsilon_j\,,\ \epsilon_j \sim N(0, I_d)\,. \quad (2)$$

Unless b and σ are constant functions of V_t, this model implies a different distribution for the variables $\{V_{\tau_j}\} := \{V_{\tau_0}, \ldots, V_{\tau_M}\}$ than the one implied by the solution of (1). However, the distance between these two distributions converges to 0 as $\delta := \sup_j(\tau_{j+1} - \tau_j) \to 0$, see Theorem 10.2.2 and Remark 10.2.3 in [14].

Likelihood-based inference for θ given high-frequency observations $\{V_{\tau_j}\}$ is classical and in principle straightforward using either the approximating model (2) or continuous-time arguments. For any δ, (2) defines a locally Gaussian Markov model and the likelihood is immediately available. However, (2) is meant to approximate (1) and this will only be true in the high-frequency limit $\delta \to 0$. [10] showed that when the data frequency δ is fixed and the number of observations increases, the maximum pseudo-likelihood estimators for (θ_1, θ_2) based on (2) with data generated from (1) can be inconsistent, see also [19] for a simple example using the Ornstein-Uhlenbeck process. In the high-frequency regime, basic stochastic calculus allows us to do statistical inference adopting the continuous-time perspective too. The quadratic variation identity [see 13]

$$\lim_{\delta \to 0} \sum_{\tau_j \leq t} (V_{\tau_{j+1}} - V_{\tau_j})(V_{\tau_{j+1}} - V_{\tau_j})^T = \int_0^t \Gamma(s, V_s; \theta_2)\mathrm{d}s \quad \textit{in probability} \quad (3)$$

can be used to estimate consistently (as $\delta \to 0$) the components of θ_2 from high frequency observations on $[0, T]$; note that the identity might have to be applied over different time sub-intervals of $[0, T]$ to provide enough estimating equations. The Cameron-Martin-Girsanov theorem can be used to obtain a likelihood for θ_1 given the estimated value of θ_2. [16] (Section 2.1) point out that this theorem is typically not formulated in a way which is useful for statistical inference. After simple manipulation, one obtains the following function as the continuous-time likelihood for θ_1:

$$G(0, T, V, b, A; \theta) := \exp\left\{ \int_0^T b(s, V_s; \theta_1)^T A(s, V_s; \theta_2)\mathrm{d}V_s - \frac{1}{2}\int_0^T b(s, V_s; \theta_1)^T A(s, V_s; \theta_2)\, b(s, V_s; \theta_1)\mathrm{d}s \right\}. \quad (4)$$

In practice, the integrals involved in this expression will be approximated with Riemann sums using the observations $\{V_{\tau_j}\}$ yielding the approximation

$$G(0, T, \{\tau_j\}, \{V_{\tau_j}\}, b, A; \theta) := \exp\left\{ \sum_{j=0}^{M-1} b(\tau_j, V_{\tau_j}; \theta_1)^T A(\tau_j, V_{\tau_j}; \theta_2) \left[V_{\tau_{j+1}} - V_{\tau_j} - \frac{1}{2} b(\tau_j, V_{\tau_j}; \theta_1)(\tau_{j+1} - \tau_j) \right] \right\}. \quad (5)$$

It is easy to study the connections between the discrete time approach based on the Euler approximation and the approximated continuous time approach in this high-frequency framework. The Riemann approximation to the integrals in (5) yields precisely the same likelihood (up to proportionality constants) for θ_1 for fixed θ_2 as the one obtained from (2). When Γ is constant in v and s, the Euler likelihood using the highest order terms in δ yields the same estimating equation for Γ as the quadratic variation identity. For more general diffusion matrices we do not have such an explicit relation between the two approaches. For example, this is the case when $\Gamma(s, v; \theta_2) = \theta_2 H(s, v)$ where (only in this place in the article) θ_2 is an unknown matrix and H a known matrix-valued function. However, note that in such case there are several possible quadratic variation identities (e.g. by first rescaling the data by H^{-1}). Computationally, inference in the high-frequency regime might pose challenges when the b and Γ are non-linear functions of the parameters.

In practice it is not easy to check whether δ is small enough for the approximations described above to be satisfactory, and it is unclear what to do if it is deemed that δ is not small enough. Furthermore, in many applications of interest V is only partially observed; its components might be observed asynchronously, certain components might be latent or there might be observation error. Therefore, when estimating diffusion processes we typically have a subset of the data which is needed to obtain the likelihood function in closed form or a satisfactory approximation thereof. The missing data are the unobserved interpolating paths in the continuous-time approach, and the values of the process at a desired frequency δ in the discrete-time approach.

Since the beginning of the 21st century Monte Carlo methods based on the principle of data augmentation have emerged for likelihood-based inference for partially observed diffusions. These methods involve an "imputation" step, where the existing dataset is augmented with auxiliary variables, and an "estimation" step, where inference for the parameters is done on the basis of the augmented dataset. In this article we will focus on Monte Carlo maximum likelihood and Markov chain Monte Carlo data augmentation methods.

However, data augmentation for diffusions cannot be done using off-the-shelf algorithms which have successfully been applied to random-effects type models. The following features, which are more clearly pronounced in the continuous-time approach, have dictated the agenda on data augmentation for diffusions. First, the missing data are infinite-dimensional in nature. Second, the imputation step requires the non-trivial simulation of conditioned stochastic processes. Third, due to (3) the missing data contain infinite information about the diffusion coefficient whereas there is finite information in the observed data. Fourth, the algorithms derived by the continuous-time approach typically require a finite dimensional approximation for computer implementation. Finally, different or seemingly different methods have been developed by taking the discrete or the continuous-time formulation of the model as the starting point, i.e., taking (2) for δ small enough, or (4) as the joint model for observed and missing data.

On the other hand, data augmentation methods appropriately designed can handle effectively different types of incomplete observations. The missing data problem that turns out to be central to this technology is that of discrete-time observations, i.e., inference for (1) on the basis of observations $\{V_{t_i}, i = 0, \ldots, n\}$ with $0 = t_0 < t_1 \cdots < t_n = T$. This is an important problem in its own right. Additionally, computational methods for other types of partial observations typically involve reduction to the discretely-observed case by imputation of unobserved values at the observation times, in order to exploit the Markov structure of the joint model (1) sampled at discrete times. For this purpose we focus on the discretely-observed case.

Aims of this paper

The main aim in this paper is to provide a generic and transparent framework for data augmentation for diffusions. We introduce a generic program which can be followed in order to identify appropriate auxiliary variables, to design Monte Carlo maximum likelihood and Markov chain Monte Carlo algorithms that are valid even in the limit where continuous paths are imputed, and to approximate these limiting algorithms.

We effectively pin down the methodology to deriving importance sampling representations for the probability distribution of a conditioned diffusion. The representations require relatively simple stochastic calculus and, as we show, they are all effectively based on the same decomposition of the diffusion measure, expression (8) in this paper. The representations in turn provide identities for the transition density of the diffusion. In particular, we formally prove an identity for the transition density of multivariate non-reducible diffusions that was heuristically derived for scalar processes in [16] (for definition of reducible diffusions, see Section 2.1). A main result in this paper, in Section 3, is to demonstrate that such identities are instrumental in the design of data augmentation algorithms. We describe a general methodology that includes as special cases the data augmentation algorithm for reducible diffusions introduced in [21] and the limit of the algorithm introduced in [12] within a discrete-time approach for non-reducible processes.

In this paper we also bridge data augmentation methodologies that have been developed by discrete and continuous-time approaches. We demonstrate that the state-of-the-art discrete-time importance sampling approximation to the likelihood function derived in [8] is almost identical to a finite-dimensional approximation of the continuous-time importance sampling approximation introduced in [16]. This connection is established by means of some intermediate results: the identity for the transition density mentioned above; a novel discrete-time approximation for simulating conditioned diffusions; a careful treatment of the effect of discretising the continuous-time expressions.

The ideas in this paper go much beyond the context of elliptic diffusions that we treat here. The generic program we propose suggests how to deal with hypoelliptic or jump diffusions, or observation schemes different from discrete-time observations. Effectively, the framework is a powerful extension of non-centring for stochastic processes, which was introduced in [17, 18], and we expect it to be relevant well beyond the context of stochastic differential equations.

In Section 2, we provide a precise introduction to the problem, couching it in modern computational statistics language and providing a description of existing importance sampling methodologies for diffusion, bridges and giving novel results on properties of discretisation approaches. Section 3 uses the framework in the previous section to construct and study MCMC methods for Bayesian inference, and the respective discretisations. The paper concludes with discussion in Section 4 with proofs and other technical material placed in appendices. In all sections, we treat the reducible and irreducible cases separately.

2 Importance sampling representations for the missing data and likelihood approximations

2.1 Preliminaries

We focus here on discretely observed diffusions, i.e., consider observations $\{V_{t_i}, i = 0, \ldots, n\}$ with $0 = t_0 < t_1 \cdots < t_n = T$ from (1). By the Markov property, the likelihood for this

set of observations is obtained as a product of transition density terms,

$$p_{s,t}(u,v;\theta) = \Pr\left[\, V_t \in \mathrm{d}v \mid V_s = u \,\right] / \,\mathrm{d}v, \quad t > s,\, u, v \in \mathbb{R}^d\,, \tag{6}$$

which, however, are unavailable for most diffusion processes. Due to the Markov property of either the continuous-time model (1) or its Euler approximation (2), the values of the process in between observation times are conditionally independent given the observations and the parameters. Hence, in the rest of this section and without loss of generality we will study importance sampling representations of the missing data only for one pair of observations, $V_0 = u, V_T = v$, and we will treat θ as known. Intermediate times at which we will be simulating the missing data will be denoted by $\{\tau_j\}$, where $0 = \tau_0 < \tau_1 < \cdots < \tau_M = T$ and without loss of generality we will assume that they are equally spaced, $\tau_{k+1} - \tau_k = \delta$ and $M\delta = T$. In the rest of the paper, when dealing with several observations we will assume that $t_i - t_{i-1}$ is a multiple of δ, for each i; this is to avoid unnecessary notation that introduces a different step δ for each pair of observations, which might of course be useful in practice.

Additionally, let $\mathbb{P}(0, T, u; \theta)$ be the law of (1) conditioned upon $V_0 = u$; thus (6) is the density of a marginal distribution of $\mathbb{P}(0, T, u; \theta)$. Correspondingly, let $\mathbb{P}(0, T, u, v; \theta)$ denote the law of (1) conditioned upon $V_0 = u$ and $V_T = v$. The diffusion process conditioned upon its endpoints is a stochastic process known as the diffusion bridge. Therefore, the missing data conditioned upon the observations are diffusion bridges and their distribution is $\mathbb{P}(0, T, u, v; \theta)$. A time subscript in the probability measures denotes their projection to the corresponding time interval, e.g. $\mathbb{P}(0, T, u, v; \theta)_t$ denotes the law of the diffusion bridge over paths on $[0, t]$, for $t \leq T$. We will use V to refer both to the solution of (1) and to the corresponding bridge; the distinction will be clear from the context, especially since we will be associating them with different measures. Finally, for a probability measure $\mathbb{P}$, $\mathbb{E}_{\mathbb{P}}$ will denote expectations with respect to it.

2.1.1 An important special case: reducible diffusions

We also introduce a special class of diffusion processes for which the theory is simpler and more efficient simulation methods can be devised. These are diffusions for which a transformation $\eta(\cdot, \cdot; \theta_2) : \mathbb{R}_+ \times \mathbb{R}^d \times \Theta_2 \to \mathbb{R}^m$ exists such that $X_s := \eta(s, V_s; \theta_2)$ solves an SDE with unit diffusion coefficient. Such a transformation described by η is called the *Lamperti* transformation of the diffusion process. Necessary conditions for η follow directly from the application of Itô's lemma. The following is a sufficient condition in the special case $d = m$, $\nabla\eta(s, v; \theta_2)\sigma(s, v; \theta_2) = I_d$, for all s, v, θ_2, where the differential operator applies to v, and I_d is the d-dimensional identity matrix. This is trivially satisfied when $d = m = 1$ (under mild smoothness conditions on σ) by $\eta(s, v; \theta_2) = \int^v 1/\sigma(s, u; \theta_2)\mathrm{d}u$, but it is much harder or impossible to find such transformation in high dimensions. To avoid unnecessary generality, reducible diffusions will refer to the smaller class of processes for which $d = m$, the sufficient condition is satisfied and additionally $\eta(s, \cdot; \theta_2)$ is twice differentiable in its arguments and admits an inverse η^{-1} for each s, θ_2. Note that under these conditions X solves the SDE

$$\mathrm{d}X_s = \alpha(s, X_s; \theta)\mathrm{d}s + \mathrm{d}B_s \tag{7}$$

where α is obtained by a direct application of Itô's lemma. Mild and standard conditions on α imply that the law of X is equivalent to the Wiener measure and their density is obtained by the Cameron-Martin-Girsanov theorem. Thus, after the Lamperti transformation we obtain a process which is equivalent, in terms of the corresponding measures, to a Gaussian process, although the original process is not. For this reason the Lamperti

transformation should always be used where possible, since it makes the model closer to a Gaussian for which theory and computations are typically easier. However, as pointed out above, for many important classes of multivariate diffusions the transformation does not exist.

In terms of notation, $\mathbb{W}$ and $\mathbb{P}^X$ will be used to denote probability measures for Brownian motion and the diffusion X in (7). The transition density of Brownian motion is Gaussian and we will denote the Gaussian density with mean 0 and covariance matrix Σ evaluated at y by $\mathcal{G}(y;\Sigma)$.

2.2 ABC for diffusions

One of the earliest simulation methods for diffusion bridges is proposed in [20], although here we present the idea from a different perspective than the original article. Cast in modern terms, this method can be interpreted as an Approximate Bayesian Computation approach to simulating diffusion bridges. Simulation of bridges is a type of inverse problem. One way to solve this problem is by simulating a diffusion path forwards according to $\mathbb{P}(0,T,u;\theta)$ and accept it if $V_T = v$. This would yield exact draws from $\mathbb{P}(0,T,u,v;\theta)$ but the probability that $V_T = v$, is zero under $\mathbb{P}(0,T,u;\theta)$. However, using Bayes theorem and the Markov property, we have for $\epsilon > 0$ such that $T - \epsilon > 0$,

$$\begin{aligned}\mathbb{P}(0,T,u,v;\theta)_{T-\epsilon} &= \mathbb{P}(0,T-\epsilon,u,z;\theta) \otimes \frac{p_{0,T-\epsilon}(u,z;\theta)p_{T-\epsilon,T}(z,v;\theta)}{p_{0,T}(u,v;\theta)}\,\mathrm{d}z \\ &= \frac{p_{T-\epsilon,T}(V_{T-\epsilon},v;\theta)}{p_{0,T}(u,v;\theta)}\,\mathbb{P}(0,T-\epsilon,u;\theta) \propto p_{T-\epsilon,T}(V_{T-\epsilon},v;\theta)\,\mathbb{P}(0,T-\epsilon,u;\theta)\end{aligned} \tag{8}$$

This identity immediately suggests an importance sampling algorithm for diffusion bridges on $[0, T-\epsilon]$; simulate paths forwards according to $\mathbb{P}(0,T-\epsilon,u;\theta)$ and weight them by $p_{T-\epsilon,T}(V_{T-\epsilon},v;\theta)$. The weight is actually intractable due to the intractability of the transition density. Nevertheless, for ϵ small enough it can be approximated by a Gaussian density according to (2); hence the ABC algorithm. Note that this importance sampling representation yields an identity for the transition density by integrating both sides over paths:

$$p_{0,T}(u,v;\theta) = \int p_{T-\epsilon,T}(V_{T-\epsilon},v;\theta)\,\mathrm{d}\mathbb{P}(0,T-\epsilon,u;\theta) = \int p_{T-\epsilon,T}(z,v;\theta)\,p_{0,T-\epsilon}(u,z;\theta)\mathrm{d}z \tag{9}$$

which is the Chapman-Kolmogorov equation. Albeit simple, this importance sampling representation is not efficient. As $\epsilon \to 0$, i.e., as the bias in the weights due to the Euler approximation decreases, the weights converge to zero, $\mathbb{P}(0,T,u;\theta)$-a.s. In practice, given a set of simulated paths and for ϵ small enough, a single path will have massively larger weight than all others.

2.3 Efficient importance sampling of diffusion bridges

Efficient importance sampling schemes can be derived by exploiting known probabilistic properties of diffusion bridges. First, note that the diffusion bridge is a Markov process. Additionally, it is known that the bridge of (1) solves the following SDE:

$$\mathrm{d}V_s = (\, b(s,V_s;\theta_1) \,+\, \Gamma(s,V_s;\theta_2)\,\nabla_{V_s}\log p_{s,T}(V_s,v;\theta)\,)\,\mathrm{d}s + \sigma(s,V_s;\theta_2)\,\mathrm{d}B_s \tag{10}$$

on $s \in [0,T]$ with $V_0 = u$. This result can be proved using the theory of h-transforms, see for example Chapter IV.39 of [23], which is also based on a decomposition of measures similar to (8). This representation is not helpful for direct simulation of the bridge because the drift of (10) involves the transition density. Nevertheless, it is useful for deriving efficient importance sampling algorithms, as we now describe. First, note that the SDEs for the conditioned and unconditioned diffusion have the same diffusion coefficient. Second, given that the solution of (10) hits $V_T = v$ $\mathbb{P}(0,T,u,v;\theta) - a.s.$ at time T, we approximate the drift and the diffusion coefficient of (1) at times s near T by the constants $b(T,v;\theta_1)$ and $\sigma(T,v;\theta_2)$ respectively. Then the unconditional process is a scaled Brownian motion with drift, for which the bridge process is tractable and solves an SDE whose drift is given by $(v - V_s)/(T-s)$. These two observations suggest that we can improve on the ABC method by proposing paths according to a stochastic process which hits v at time T almost surely, thus incorporating part of the weight into the proposal. [6] proposed the following stochastic process for this purpose,

$$\mathrm{d}V_s \;=\; \frac{v - V_s}{T-s}\,\mathrm{d}s + \sigma(s,V_s;\theta_2)\,\mathrm{d}B_s, \quad s \in [0,T]\,, V_0 = u\,. \tag{11}$$

Let $\mathbb{Q}(0,T,u;\theta_2,v)$ denote the law generated by the solution of this SDE; the notation reflects that this is not the law of a conditioned process, instead of one whose drift depends on v. [6] showed that $\mathbb{Q}(0,T,u;\theta_2,v)$ is equivalent to $\mathbb{P}(0,T,u,v;\theta)$ under certain conditions on b and σ (see Assumption 4.2 of their paper) and obtained the density between the two measures up to proportionality. Here, we present a slightly different argument directly based on (8) and we obtain an identity for the transition density.

We have that

$$\begin{aligned}
\frac{\mathrm{d}\mathbb{P}(0,T,u,v;\theta)_{T-\epsilon}}{\mathrm{d}\mathbb{Q}(0,T,u;\theta_2,v)_{T-\epsilon}} &= \frac{\mathrm{d}\mathbb{P}(0,T,u,v;\theta)_{T-\epsilon}}{\mathrm{d}\mathbb{P}(0,T-\epsilon,u;\theta)} \times \frac{\mathrm{d}\mathbb{P}0,(T-\epsilon,u;\theta)}{\mathrm{d}\mathbb{Q}(0,T,u;\theta_2,v)_{T-\epsilon}} \\
&= \frac{p_{T-\epsilon,T}(V_{T-\epsilon},v;\theta)}{p_{0,T}(u,v;\theta)} \frac{\mathrm{d}\mathbb{P}(0,T-\epsilon,u;\theta)}{\mathrm{d}\mathbb{Q}(0,T,u;\theta_2,v)_{T-\epsilon}} \\
&= \frac{1}{p_{0,T}(u,v;\theta)}\, G(0,T-\epsilon,V,b,A;\theta)\, p_{T-\epsilon,T}(V_{T-\epsilon},v;\theta) \\
&\times \exp\left\{ -\int_0^{T-\epsilon} \frac{(v-V_s)^T}{T-s} A(s,V_s;\theta_2)\mathrm{d}V_s \right. \\
&\left. +\frac{1}{2}\int_0^{T-\epsilon} \frac{1}{(T-s)^2}(v-V_s)^T A(s,V_s;\theta_2)(v-V_s)\mathrm{d}s \right\}
\end{aligned}$$

where we have used (8) in the second equation and Girsanov's theorem for the processes in (1) and (11) applied over the time-interval $[0,T-\epsilon]$, for the third equation followed by a rearrangement of terms. We now need to study the stability of the last expression as $\epsilon \to 0$. The first term is constant in ϵ and the second has the well defined limit $G(0,T,V,b,A;\theta)$. For small ϵ, we will approximate $p_{T-\epsilon,T}(V_{T-\epsilon},v;\theta)$ by a zeroth-order Euler approximation,

$$(2\pi)^{-d/2}\,|A(T-\epsilon,V_{T-\epsilon};\theta_2)|^{1/2}\,\epsilon^{-d/2}\exp\left\{-\frac{1}{2\epsilon}(v-V_{T-\epsilon})^T A(T-\epsilon,V_{T-\epsilon};\theta_2)(v-V_{T-\epsilon})\right\}.$$

The ratio between this approximation and the transition density converges to 1 almost surely under $\mathbb{Q}(0,T,u;\theta_2,v)$ and the assumed conditions. To determine the limit of the remaining terms we carefully apply Itô's formula on the exponent of the Euler approximation above, which brings up terms that cancel others in the exponent of the likelihood ratio obtained above; see Appendix A for details. Thus we obtain a well-defined limit as

$\epsilon \to 0$, which is

$$\frac{\mathrm{d}\mathbb{P}(0,T,u,v;\theta)}{\mathrm{d}\mathbb{Q}(0,T,u;\theta_2,v)} = \frac{1}{p_{0,T}(u,v;\theta)} R^0(\theta) \tag{12}$$

where

$$\begin{aligned} R^0(\theta) = (2\pi T)^{-d/2} \left|A(T,v;\theta_2)\right|^{1/2} \exp\left\{-\frac{1}{2T}(v-u)^T A(0,u;\theta_2)(v-u)\right\} \\ \times G(0,T,V,b,A;\theta)\zeta(0,T,V,A;\theta_2)\,. \end{aligned}$$

In the above expression, we define

$$\zeta(0,T,V,A;\theta_2) = \exp\left\{-\frac{1}{2}\int_0^T \frac{1}{T-s}(v-V_s)^T(\diamond\,\mathrm{d}A)(v-V_s)\right\}\,.$$

The $\diamond$-stochastic integral is obtained as the limit of sums where the integrand is computed at the right limit of each time interval as opposed to the left limit used in the definition of the Itô integral, or the middle point used in the Stratonovich integral. On the technical side, to establish this limit as $\epsilon \to 0$, we need to ensure that ζ is well defined. Note the division by $T-s$ in the integrand could be the cause of instability. A rough way to ensure that this stochastic integral is well defined is to assume that A and σ are bounded. The argument is sketched in Appendix A.

Notice now that we have an identity for the transition density of (1). Integrating both sides of the likelihood ratio above with respect to $\mathbb{Q}(0,T,u;\theta_2,v)$ we obtain that

$$p_{0,T}(u,v;\theta) = \mathbb{E}_{\mathbb{Q}(0,T,u;\theta_2,v)}\left[R^0(\theta)\right]\,. \tag{13}$$

When $d = m$ and $\sigma = I_d$, (11) is the Brownian bridge, i.e., the Brownian motion conditioned to take the values u, v at times $0, T$ respectively, and $\zeta = 1$. In general, though, (11) will not correspond to a conditioned diffusion, it is instead a stochastic process controlled to hit the value v at time T.

2.3.1 The reducible case

An alternative class of proposals is available for reducible diffusions (recall the definition and notation in Section 2.1.1), obtained as $V_s = \eta^{-1}(s, X_s; \theta_2)$, for X a Brownian bridge with endpoints $\eta(0,u;\theta_2)$ and $\eta(T,v;\theta_2)$ at times 0 and T respectively. This family of stochastic processes are diffusion bridges and do not solve an SDE of the type (11) unless $\sigma = I_d$, in which case we obtain again the Brownian bridge proposal. Thus, according to this scheme, paths X are proposed according to $\mathbb{W}(0,T,\eta(0,u;\theta_2),\eta(T,v;\theta_2))$ and then are deterministically transformed to yield a proposed diffusion bridge. The same argument we used above can be employed to derive the weights that should be associated to bridges proposed in this manner to represent the diffusion bridge (10), the only difference being that we work with $\mathbb{P}^X$ and $\mathbb{W}$. Working as above we obtain that for any $u, v \in \mathbb{R}^d$,

$$\begin{aligned} \frac{\mathrm{d}\mathbb{P}^X(T,\eta(0,u;\theta_2),\eta(T,v;\theta_2);\theta)}{\mathrm{d}\mathbb{W}(T,\eta(0,u;\theta_2),\eta(T,v;\theta_2))} &= \frac{1}{p_{0,T}(u,v;\theta)} R^0_\eta(\theta) \\ R^0_\eta(\theta) &= \frac{\mathcal{G}(\eta(T,v;\theta_2)-\eta(0,u;\theta_2);T\,I_d)}{J(T,v;\theta_2)} G(0,T,X,\alpha,I_d;\theta)\,. \end{aligned} \tag{14}$$

The product $p_{0,T}(u,v;\theta)J(T,v;\theta_2)$ is the transition density of (7), which is obtained from the one of (1) by the change of variables $V_T \to X_T = \eta(T,v;\theta_2)$ and induces the Jacobian term $J(T,v;\theta_2)$; X is the proposed Brownian bridge. Since the proposed bridge V is a transformation of X, the likelihood ratio above is the corresponding importance sampling weight for sampling from (10). In this case, we obtain the transition density identity,

$$p_{0,T}(u,v;\theta) = \mathbb{E}_{\mathbb{W}(0,T,\eta(0,u;\theta_2),\eta(T,v;\theta_2))}\left[R^0_\eta(\theta)\right]\,. \tag{15}$$

2.4 Finite dimensional approximation of the proposals

The proposals and the weights involved in the importance sampling algorithms for diffusion bridges will typically have to be approximated for computer implementation. Under additional assumptions on the coefficients of (1), which when $d > 1$ are stronger than assuming that V is reducible, such approximations might be avoidable - this is the exact simulation and inference paradigm of [3]. In this section we deal with the approximation of the proposals, and address the approximation of the weights in Section 2.5.

2.4.1 The reducible case

In the case of reducible diffusions the proposal process is a diffusion bridge itself. It is in fact a simple Brownian bridge, the SDE of which we can solve exactly in order to generate skeletons $\{X^\delta_{\tau_j}\}_{j=0}^M$ according to the following scheme, where recall that $\delta = \tau_{k+1} - \tau_k$,

$$X^\delta_{\tau_{k+1}} = X^\delta_{\tau_k} + \delta \frac{\eta(T, v; \theta_2) - X^\delta_{\tau_k}}{T - \tau_k} + \sqrt{\frac{T - \tau_{k+1}}{T - \tau_k}} \left(B_{\tau_k} - B_{\tau_{k-1}}\right) \quad X_{\tau_0} = \eta(0, u; \theta_2)\,, \quad (16)$$

where $\{B_{\tau_j}\} \sim \mathbb{W}^\delta(0, T, 0)$ is a skeleton of Brownian motion on $[0, T]$. The law of $\{X_{\tau_j}\}$ will be denoted by $\mathbb{W}^\delta(0, T, \eta(0, u; \theta_2), \eta(T, v; \theta_2))$. Note that both $\mathbb{W}^\delta(0, T, \eta(0, u; \theta_2), \eta(T, v; \theta_2))$ and $\mathbb{W}^\delta(0, T, 0)$ are exact projections of $\mathbb{W}(0, T, \eta(0, u; \theta_2), \eta(T, v; \theta_2))$ and $\mathbb{W}(0, T, 0)$ respectively on the lattice $\{\tau_j\}$.

2.4.2 The irreducible case

It is impossible to simulate exact skeletons of (11) when σ is not constant in the state variable, but we can consider two different approximate discretisations:

1. One possibility is to use the Euler-Maruyama scheme to discretise (11). It is convenient here to define the scheme as a continuous-time one, i.e., to define the approximating diffusion U^δ_s as follows

$$\mathrm{d}U^\delta_s = \tilde{b}(s, U^\delta_s)\,\mathrm{d}s + \tilde{\sigma}(s, Y^\delta_s; \theta_2)\,\mathrm{d}B_s$$

   where $\tilde{\sigma}(s, Y^\delta_s; \theta_2) = \sigma(\tau_k, Y^\delta_{\tau_k}; \theta_2)$ and $\tilde{b}(s, U^\delta_s) = (v - U^\delta_{\tau_k})/(T - \tau_k)$, for all $\tau_k \le s < \tau_{k+1}$ and $U^\delta_0 = V_0 = u$; the generated discrete-time chain will be denoted by $\{U^\delta_{\tau_j}\}$, and its transition density $g^\delta_{\tau_k, \tau_{k+1}}(w, z; \theta_2)$.

2. Another possibility is to discretise based on a local linearisation of (11), according to the ideas proposed in a series of articles by Ozaki and Shoji [see, for example, 26, 25]. We define the approximating diffusion Y^δ_s as follows:

$$\mathrm{d}Y^\delta_s = \frac{v - Y^\delta_s}{T - s}\,\mathrm{d}s + \tilde{\sigma}(s, Y^\delta_s; \theta_2)\,\mathrm{d}B_s \quad (17)$$

   with $\tilde{\sigma}$ defined as above and $Y^\delta_0 = V_0 = u$. The point is that (17) is a piecewise-linear SDE which can be easily solved and leads to the following discrete-time chain:

$$Y^\delta_{\tau_{k+1}} = Y^\delta_{\tau_k} + \delta \frac{v - Y^\delta_{\tau_k}}{T - \tau_k} + \sigma(\tau_k, Y^\delta_{\tau_k}; \theta_2) \sqrt{\frac{T - \tau_{k+1}}{T - \tau_k}} \left(B_{\tau_k} - B_{\tau_{k-1}}\right), \quad Y^\delta_{\tau_0} = u\,, \quad (18)$$

   where as in (16) $\{B_{\tau_j}\} \sim \mathbb{W}^\delta(0, T, 0)$ is a Brownian skeleton. In the rest of the paper, the law of $\{Y^\delta_{\tau_j}\}$ will be denoted by $\mathbb{Q}^\delta(0, T, u; \theta_2, v)$, and its transition density by $q^\delta_{\tau_k, \tau_{k+1}}(w, z; \theta_2)$.

The explosion of the drift of (11) near time T means that the usual Lipschitz and growth conditions that are typically used in the literature to study the order of discretisation schemes (see for example [14, 15] for extensive treatment of various such results) do hold here. Theorem 1 below shows that under usual assumptions on σ, which are detailed in Appendix B, both the Euler-Maruyama and the local linearisation approaches provide strong approximation of order 1/2 to (11), i.e., they retain the order they have when the drift of the approximated SDE is more regular. In order to avoid excessive notation, the result is proved for $d = 1$ in Appendix B.

Theorem 1. *Let $d = 1$ and consider the technical conditions on σ stated in Appendix B. Then, for any step size $0 < \delta < \min\{1, T\}$ there exist constants $L > 0$ and $E > 0$ independent of δ such that for all $0 < t < T$*

$$\mathbb{E}[(Y_t^\delta - V_t)^2] \leq L\delta\,, \quad \mathbb{E}[(U_t^\delta - V_t)^2] \leq E\delta$$

where the expectation is taken with respect to the Brownian motion which drives both (11) *and the approximations.*

The theorem shows that the two schemes are equivalent in terms of their order, as it happens under regular conditions on the drift of the approximated diffusion, see for example [25]. However, their efficiency in practical situations can differ considerably. Our simulation studies (not included here) mainly for square-root diffusion coefficients, show that the mean square error when using the Euler approximation to (11) is bigger than when using (18) for all $0 < t < T$. The advantage of using (18) is more apparent for times near the end point.

Quantitive results on relative efficiency of discretisation schemes are scarce in the literature. We can provide a concrete such result for the two schemes under consideration in this section for a simple, although still interesting, process. The following proposition is proved in Appendix B.

Proposition 1. *Let V be the solution to the standard Brownian bridge SDE, i.e., equation* (17) *with $d = 1, \sigma = 1$, $u = v = 0\,, T = 1$. Let $0 < \delta < \min\{1, T\}$ be any step size and U_t^δ the corresponding Euler-Maruyama approximation; let $\tau_k = k\delta$, for $k = 1, \ldots, \lfloor T/\delta \rfloor$. Then,*

$$\begin{aligned}
\mathbb{E}[(V_{\tau_k} - U_{\tau_k}^\delta)^2] &\geq \frac{\delta^2}{9}(1-\tau_k)^2\left[\frac{1}{(1-\tau_k+\delta)^3} - 1\right] + \frac{\delta^3}{3}(1-\tau_k)^2 \\
\mathbb{E}[(V_{\tau_k} - U_{\tau_k}^\delta)^2] &\leq \frac{\delta^2}{9}(1-\tau_k)^2\left[\frac{1}{(1-\tau_k)^3} - 1\right] + \frac{\delta^3}{3}(1-\tau_k)^2\left[\frac{1}{(1-\tau_k)^4} - 1\right]
\end{aligned}$$

where the expectation is taken with respect to the Brownian motion that drives both the exact solution and the approximation.

First, note that the local linearisation scheme is exact for this SDE, i.e., $\mathbb{E}[(V_t - Y_t^\delta)^2] = 0$ for all t. Second, the tight bounds allow us to refine our understanding of the theoretical properties of the Euler-Maruyama scheme for processes which are controlled to hit an endpoint. The result implies that for any $t < 1$ the Euler scheme provides a strong approximation of order 1, not just 1/2 that it is ensured in Theorem 1. This is unsurprising since the Brownian bridge SDE satisfies standard Lipschitz and growth conditions on $[0, t]$ (e.g Theorem 10.2.2 in [14]), and the noise is additive since $\sigma = 1$. On the other hand, $\mathbb{E}[(V_{1-\delta} - U_{1-\delta}^\delta)^2]$ is only of $\mathcal{O}(\delta)$, i.e., we recover the lower order of Theorem 1 for times arbitrarily close to the endpoint.

2.5 Discretisation of the weights

The identities derived in (13), (15) can be used for the Monte Carlo estimation of the transition density of the process for given parameters θ. Such estimates can be embedded to Monte Carlo maximum likelihood schemes, e.g. as in [8, 2], or other type of stochastic optimisation or approximation methods, e.g. Monte Carlo EM algorithms as in [3, 6], in order to compute the maximum likelihood estimator of θ. Section 3 shows that these identities are the main building block of MCMC algorithms for Bayesian inference for θ.

For practical implementation, the continuous-time identities typically require a finite dimensional approximation on the basis of a skeleton of the proposal process produced as described in Section 2.4. In particular, $G(0,T,\{\tau_j\},\{Y^\delta_{\tau_j}\},b,A;\theta)$ is the obvious Riemann-sum approximation based on a skeleton produced by (18), and

$$\zeta(0,T,\{\tau_j\},\{Y^\delta_{\tau_j}\},A;\theta_2) := \exp\left\{-\frac{1}{2}\sum_{j=0}^{M-2}\frac{1}{T-\tau_{j+1}}(v-Y^\delta_{\tau_{j+1}})^T[A(\tau_{j+1},Y^\delta_{\tau_{j+1}};\theta_2)-A(\tau_j,Y^\delta_{\tau_j};\theta_2)](v-Y^\delta_{\tau_{j+1}})\right\} \quad (19)$$

is the approximation of ζ, where, as we pointed out earlier, the integrand is evaluated at the right limit of each time interval of length δ. Note that the sum above runs up to $M-2$ and not $M-1$, which is the case in (5); this is due to the fact that $(v-V_s)/(T-s)$ yields $0/0$ at $s=T$, almost surely under $\mathbb{Q}(0,T,u;\theta_2,v)$, thus effectively we set this ratio equal to 0. We can then approximate $p_{0,T}(u,v;\theta)$ by $\mathbb{E}_{\mathbb{Q}^\delta(0,T,u;\theta_2,v)}\left[R^\delta(\theta)\right]$, where

$$\begin{aligned} R^\delta(\theta) &= (2\pi T)^{-d/2}\,|A(T,v;\theta_2)|^{1/2}\exp\left\{-\frac{1}{2T}(v-u)^T A(0,u;\theta_2)(v-u)\right\} \\ &\quad\times G(0,T,\{\tau_j\},\{Y^\delta_{\tau_j}\},b,A;\theta)\,\zeta(0,T,\{\tau_j\},\{Y^\delta_{\tau_j}\},A;\theta_2)\,. \end{aligned} \quad (20)$$

In the case of reducible diffusions, $G(0,T,\{\tau_j\},\{X_{\tau_j}\},\alpha,I_d;\theta)$ is the approximation of the Girsanov formula of the transformed process based on a Brownian bridge skeleton $\{X_{\tau_j}\}$, and $p_{0,T}(u,v;\theta)$ is approximated by $\mathbb{E}_{\mathbb{W}^\delta(0,T,\eta(0,u;\theta_2),\eta(T,v;\theta_2))}\left[R^\delta_\eta(\theta)\right]$ where

$$R^\delta_\eta(\theta) = \frac{\mathcal{G}(\eta(T,v;\theta_2)-\eta(0,u;\theta_2);T)}{J(T,v;\theta_2)}\,G(0,T,\{\tau_j\},\{X_{\tau_j}\},\alpha,I_d;\theta)\,.$$

We comment on the convergence of $R^\delta(\theta)$ and $R^\delta_\eta(\theta)$ to $R^0(\theta)$ and $R^0_\eta(\theta)$ respectively, as $\delta\to 0$ in Section 3.4.

2.6 Bridging continuous and discrete-time approaches

We can study the efficient simulation of the missing data within the discrete-time approach. In this framework, for a pair of observations $V_{\tau_0}=u$ and $V_{\tau_M}=v$, the missing data are the variables $\{V_{\tau_j}\}_{j=1}^{M-1}$. The joint model of observed and missing data is the Euler-Maruyama approximation of (1), i.e., a Markov chain with Gaussian transition density

$$p^E_{\tau_j,\tau_{j+1}}(w,z;\theta) = \mathcal{G}(z-w-\delta\,b(\tau_j,w;\theta_1);\delta\,\Gamma(\tau_j,w;\theta_2))\,,\quad w,z\in\mathbb{R}^d$$

where the superscript indicates that this is the Euler approximation to (6). Note that the k-steps ahead transition density for $k>1$, which we will denote by $p^E_{\tau_j,\tau_j+k\delta}(u,v;\theta)$, is the k'th convolution of Euler transition density given above and it will be typically intractable unless $b(s,u;\theta_1)$ is linear in u, and $\Gamma(s,u;\theta_2)$ is constant in u. Thus, it is important to note that the Euler approximation provides tractable dynamics only one-step ahead. Recent results on the approximation error $|p^E_{0,T}(u,v;\theta)-p_{0,T}(u,v;\theta)|$ can be found in [11].

It is easy to see that the discrete-time process conditioned on the endpoints $V_{\tau_0} = u, V_{\tau_M} = v$ is also Markov with transition density

$$p^E_{\tau_j,\tau_{j+1}}(w,z;\theta)\frac{p^E_{\tau_{j+1},\tau_M}(z,v;\theta)}{p^E_{\tau_j,\tau_M}(w,v;\theta)} \propto p^E_{\tau_j,\tau_{j+1}}(w,z;\theta)p^E_{\tau_{j+1},\tau_M}(z,v;\theta)\,, \quad j=0,\ldots,M-2\,. \tag{21}$$

The transition density of the conditioned process is intractable due to the intractability of the convoluted Euler density. Hence direct simulation of the missing data is not feasible even in the discrete-time approach.

On the other hand, importance sampling is a feasible alternative. [8] replace the convoluted Euler densities in (21) by a nonstandard Euler approximation $\mathcal{G}(v-z-\delta(M-j-1)\,b(\tau_j,w;\theta_1);\delta(M-j-1)\,\Gamma(\tau_j,w;\theta_2))$. The latter approximation is used so that the product in (21) is a Gaussian density that can be used to generate proposals, which can then be weighted to yield samples from the distribution of the missing data. [8] call the stochastic process on the lattice $\{\tau_j\}$ generated in this manner, the modified Brownian bridge. The following Proposition, which can be immediately verified by simple algebra, provides insights into the properties of the modified Brownian bridge.

Proposition 2. *The transition density of the so called modified bridge in [8] is precisely the transition density of the stochastic process with the dynamics defined in* (18)*, i.e. the local linearisation of* (11)*.*

The relation of (18) to the proposal process (11) was missing in the literature. [8] noticed that (18) is similar to the Euler approximation of (11) except for the factor $(T-\tau_{k+1})/(T-\tau_k)$ in the variance. This is the reason they referred to (18) as the modified bridge. The better performance of the modified bridge over the Euler approximation to (11) was left in [8] as a surprising result. We have shown in Theorem 1 that this is not a modified approximation but in fact the local linearisation scheme.

As in [8], we next use (18) to derive an approximation for $p^E_{0,T}(u,v;\theta)$. It directly follows from the transition density of the missing data (by multiplying terms) that the weight associated to each proposed process $\{Y^\delta_{\tau_j}\}_{j=1}^{M-1} \sim \mathbb{Q}^\delta(0,T,u;\theta_2,v)$ should be

$$\frac{1}{p^E_{0,T}(u,v;\theta)}R^\delta_E\,, \quad R^\delta_E = \frac{\prod_{j=0}^{M-1} p^E_{\tau_j,\tau_{j+1}}(Y^\delta_{\tau_j},Y^\delta_{\tau_{j+1}};\theta)}{\prod_{j=0}^{M-2} q^\delta_{\tau_j,\tau_{j+1}}(Y^\delta_{\tau_j},Y^\delta_{\tau_{j+1}};\theta_2)}\,,$$

which yields the following identity

$$p^E_{0,T}(u,v;\theta) = \mathbb{E}_{\mathbb{Q}^\delta(0,T,u;\theta_2,v)}\left[R^\delta_E\right]\,. \tag{22}$$

Note that R^δ_E is an unbiased estimator of $p^E_{0,T}(u,v;\theta)$, which however is different from the true transition density.

Proposition 2 establishes that a discrete-time proposal process, the so-called modified Brownian bridge, coincides with the the discretisation using local linearisation of the continuous-time proposal process (17). Thus, we can directly compare the two corresponding estimators of the transition density R^δ_E and R^δ. The following result, which is proved in Appendix C, establishes a remarkable correspondence between the two. In order to avoid excessive notation, the result is proved (hence also stated) for $d=1$.

Proposition 3. *Let $d=1$. Then, we can construct $R^\delta(\theta)$ and $R^\delta_E(\theta)$ on the same probability space and then,*

$$\frac{R^\delta(\theta)}{R^\delta_E(\theta)} = \frac{|A(T,v;\theta_2)|^{1/2}}{|A(\tau_{M-1},Y^\delta_{\tau_{M-1}};\theta_2)|^{1/2}} \tag{23}$$

This correspondence appears even more striking when we repeat this calculation with other natural discretisations of (17). In particular, consider the Euler discretisation of (17) discussed in Section 2.4, in terms of a discrete-time skeleton $\{U^\delta_{\tau_j}\}$ and transition density $g^\delta_{\tau_k,\tau_{k+1}}(w,z;\theta_2)$. The approximation of the continuous-time estimator of the transition density is again $R^\delta(\theta)$, given in Section 2.5, albeit computed on the basis of $\{U^\delta_{\tau_j}\}$ and not $\{Y^\delta_{\tau_j}\}$. We can use $\{U^\delta_{\tau_j}\}$ as a discrete-time proposal process within a discrete-time approach. Working as above, the weight assigned to each proposed skeleton is:

$$\frac{1}{p^E_{0,T}(u,v;\theta)}\frac{\prod_{j=0}^{M-1}p^E_{\tau_j,\tau_{j+1}}(U^\delta_{\tau_j},U^\delta_{\tau_{j+1}};\theta)}{\prod_{j=0}^{M-2}q^\delta_{\tau_j,\tau_{j+1}}(U^\delta_{\tau_j},U^\delta_{\tau_{j+1}};\theta_2)}\frac{\prod_{j=0}^{M-2}q^\delta_{\tau_j,\tau_{j+1}}(U^\delta_{\tau_j},U^\delta_{\tau_{j+1}};\theta)}{\prod_{j=0}^{M-2}g^\delta_{\tau_j,\tau_{j+1}}(U^\delta_{\tau_j},U^\delta_{\tau_{j+1}};\theta_2)}\,.$$

A careful calculation shows that the third term equals

$$\sqrt{\frac{T}{\delta}}\exp\left\{-\frac{1}{2}\sum_{j=0}^{M-2}\frac{(\Delta B_{\tau_j})^2}{T-\tau_{j+1}}\right\} \tag{24}$$

where B is the Brownian motion that drives the Euler approximation and hence the same process for all δ subsampled at increasingly high frequency. Note that (24) does not depend on θ, but it might be dependent with the second term in the weight, due to the dependence of both on the same driving Brownian motion. Recalling that $M\delta = T$, $\tau_j = j\delta$, and that ΔB_{τ_j} has the same distribution as δY_j^2, where Y_j for $j = 0,\ldots,M-2$ are iid standard Gaussian, (24) has the same distribution as the following random variable

$$\sqrt{M}\prod_{j=0}^{M-2}\exp\{-\frac{1}{2}\frac{Y_j^2}{M-j-1}\}\,.$$

It is now easy to check that this has mean 1 and variance $\sqrt{2M/(M+1)}-1$, which does not disappear even as $M\to\infty$.

The above demonstration provides a further argument in favor of the so-called modified Brownian bridge as a discrete-time proposal process. Using the Euler discretisation of (17) as a discrete-time proposal process leads to an importance sampling weight which is the product of two terms: the weight associated with the modified Brownian bridge proposal and a random variable independent of the parameters, which marginally has mean 1. If the second term were independent of the first term then the modified Brownian bridge weight would be a *Rao-Blackwellisation* of that based on the Euler discretisation.

3 Markov chain Monte Carlo for discretely observed diffusions

3.1 Data augmentation (DA) framework

Throughout this section we assume a Bayesian framework, although a good deal of the issues we address applies to maximum likelihood inference using variants of the EM algorithm and Monte Carlo maximum likelihood. Recall the notational conventions introduced in Section 2.1. If $\pi(\theta)$ denotes the prior density (with respect to the Lebesgue, say, measure) of the parameters, statistical inference is based on the posterior density

$$\pi(\theta\mid\{V_{t_i}\})\propto\pi(\theta)\prod_{i=0}^{n-1}p_{t_i,t_{i+1}}(V_{t_i},V_{t_{i+1}};\theta)\,.$$

Typically, this function will not be computable due to the unavailability of the transition density terms.

Data augmentation (DA) can be used for Bayesian inference for the parameters of a statistical model when the posterior density is not (easily) computable. This approach has two main components. The first is mathematical and consists of identifying a joint distribution of parameters and auxiliary variables such that it admits the posterior distribution of interest as a marginal. We will be referring to this joint distribution as the auxiliary distribution. The second component of DA is computational and it consists of sampling from the auxiliary distribution using MCMC. Typically, component-wise updating algorithms are used for this purpose, such as the Gibbs sampler and the Metropolis-within-Gibbs, which iteratively sample auxiliary variables and parameters from the conditionals of the auxiliary distribution. The aim is to choose an auxiliary distribution which is possible to sample from. However, this choice is to large extent problem specific.

The first contribution of this section is to show that the transition density identities we obtained in Section 2 provide a general way of constructing auxiliary distributions for diffusions. We first treat reducible diffusions and then move to the general case.

3.2 The reducible case

Consider the following measure on parameters and bridges:

$$\prod_{i=1}^{n} R_\eta^{0,i}(\theta, X^{(i)}) \bigotimes_{i=1}^{n} \mathbb{W}(t_{i-1}, t_i, \eta(t_{i-1}, V_{t_{i-1}}; \theta_2), \eta(t_i, V_{t_i}; \theta_2))\, \pi(\mathrm{d}\theta) \tag{25}$$

where

$$R_\eta^{0,i}(\theta, X^{(i)}) = \frac{\mathcal{G}(\eta(t_i, V_{t_i}; \theta_2) - \eta(t_{i-1}, V_{t_{i-1}}; \theta_2); T\, I_d)}{J(t_i, V_{t_i}; \theta_2)} G(t_{i-1}, t_i, X^{(i)}, \alpha, I_d; \theta)$$

is defined as $R_\eta^0(\theta)$ in Section 2.3 but over the time period $[t_{i-1}, t_i]$, $X^{(i)} = (X_t^{(i)}, t_{i-1} \leq t \leq t_i)$ for $i = 1, \ldots, n$ are the auxiliary variables, which are bridges that interpolate the Lamperti-transformed observations, and the dependence of $R_\eta^{0,i}$ on each bridge $X^{(i)}$ is made explicit in the notation. Thus, the measure above is a change of measure from the product measure defined by the composition of the prior measure and independent Brownian bridge measures, with density given by $\prod_{i=1}^n R_\eta^{0,i}(\theta, X^{(i)})$. It is a direct application of (15) that this auxiliary measure admits the posterior measure $\pi(\mathrm{d}\theta \mid \{V_{t_i}\})$ as a marginal. Note that under this measure, the $X^{(i)}$'s conditionally on the parameters are independent and each distributed according to a measure proportional to $R_\eta^{0,i}(\theta, X^{(i)})\, \mathbb{W}(t_{i-1}, t_i, \eta(t_{i-1}, V_{t_{i-1}}; \theta_2), \eta(t_i, V_{t_i}; \theta_2))$.

On the computational side, there are two main problems with this auxiliary distribution. The first is due to the fact that the auxiliary variables are infinite-dimensional. This can be dealt with using approximations, a topic which we addressed in Section 2 and we consider further down. The second, however, is more fundamental and it is due to the fact that for any i, the measures $\mathbb{W}(t_{i-1}, t_i, \eta(t_{i-1}, V_{t_{i-1}}; \theta_2), \eta(t_i, V_{t_i}; \theta_2))$ are singular for different values of θ_2. That is, for each θ_2 the corresponding Brownian bridge measure has non-zero support on sets of paths which have zero probability under the Brownian bridge measures for different values of θ_2. The reason for this phenomenon is that each such measure concentrates all its mass on paths with given endpoints, but these endpoints change with θ_2. This has serious implications to any component-wise updating MCMC algorithm which targets (25). Any such algorithm will typically not be able to change the initial value of θ_2, i.e., it will not be ergodic. If $\theta_2 \to \eta(\cdot, \cdot; \theta_2)$ is many-to-one then the algorithm might be able to make some moves, but they will be very restricted and in any case this situation is uncommon.

The second contribution in this section is the idea that had we been able to obtain a transition density identity as in (15) where the expectation is with respect to a distribution that does not depend on θ, then we would be able to find an auxiliary distribution in which these singularities would not appear. The change of measure in these expectations cannot be achieved using importance sampling techniques. This is again because the measures $\mathbb{W}(t_{i-1}, t_i, \eta(t_{i-1}, V_{t_{i-1}}; \theta_2), \eta(t_i, V_{t_i}; \theta_2))$ are singular for different values of θ_2, hence it is not possible to find a common dominating measure. However, we can use properties of the Brownian bridge to achieve the change of measure. One such useful property is the linear tilting: if $Z^{(i)} \sim \mathbb{W}(t_{i-1}, t_i, 0, 0)$ and

$$X_t^{(i)} = Z_t + \frac{t - t_{i-1}}{t_i - t_{i-1}} \eta(t_i, V_{t_i}; \theta_2) + \frac{t_i - t}{t_i - t_{i-1}} \eta(t_{i-1}, V_{t_{i-1}}; \theta_2)\,, t \in [t_{i-1}, t_i]\,, \tag{26}$$

then $X^{(i)} \sim \mathbb{W}(t_{i-1}, t_i, \eta(t_{i-1}, V_{t_{i-1}}; \theta_2), \eta(t_i, V_{t_i}; \theta_2))$. In this way, $X^{(i)}$ is a function of $Z^{(i)}$, θ_2 and the observations, and we will write $X(Z^{(i)}, \theta_2)$, where the dependence of the transformation on the observations is suppressed for economy. Therefore, the transition density identity (15), applied to a pair of consecutive observations $V_{t_{i-1}}, V_{t_i}$ can we re-written as

$$p_{t_{i-1},t_i}(V_{t_{i-1}}, V_{t_i}; \theta) = \mathbb{E}_{\mathbb{W}(t_{i-1},t_i,0,0)} \left[R_\eta^{0,i}(\theta, X(Z^{(i)}, \theta_2)) \right]\,,$$

which implies that the following auxiliary distribution

$$\prod_{i=1}^{n} R_\eta^{0,i}(\theta, X(Z^{(i)}, \theta_2)) \bigotimes_{i=1}^{n} \mathbb{W}(t_{i-1}, t_i, 0, 0)\, \pi(\mathrm{d}\theta) \tag{27}$$

admits $\pi(\mathrm{d}\theta \mid \{V_{t_i}\})$ as a marginal. In this auxiliary distribution the auxiliary variables are the tilted bridges $Z^{(i)}$, for $i = 1, \ldots, n$. Note that this measure is dominated by the *product measure* given by the prior measure and independent Brownian bridge measures.

This is our proposed auxiliary distribution for reducible diffusions, which is the same as the one obtained in [21], although we have used a different and we believe more transparent and direct argument here. The conditional distribution of θ given the auxiliary variables has density proportional to

$$\pi(\theta) \prod_{i=1}^{n} R_\eta^{0,i}(\theta, X(Z^{(i)}, \theta_2))\,.$$

Conditionally on θ the auxiliary variables are independent and distributed according to a measure proportional to $R_\eta^{0,i}(\theta, X(Z^{(i)}, \theta_2))\mathbb{W}(t_{i-1}, t_i, 0, 0)$.

The conditional independence of the auxiliary variables given θ is the main advantage of a component-wise updating algorithm, which updates parameters and auxiliary variables in separate blocks according to their conditional distributions. Within such an algorithm, the conditional distribution of θ will typically be sampled using a Metropolis-Hastings step. It is a computationally costly step, since for the proposed value of θ, the paths $X^{(i)}$ will have to be reconstructed by tilting the existing paths $Z^{(i)}$. On the other hand, note that only when changing θ_2 the paths $X^{(i)}$ need reconstruction, which suggests that it might be computationally beneficial to update θ_1 and θ_2 in separate steps. For several interesting models, the update of θ_1 conditionally on θ_2 and the auxiliary variables can be done by direct simulation, for example when the drift of (1) is linear in θ_1, see for example [7]. The conditional distribution of the auxiliary variables will also by sampled using a Metropolis-Hastings step. Different proposal distributions can be considered to this effect. One possibility is to propose independent $\widetilde{Z^{(i)}} \sim \mathbb{W}(t_{i-1}, t_i, 0, 0)$ and accept them with probability

$$\min \left\{ R_\eta^{0,i}(\theta, X(\widetilde{Z^{(i)}}, \theta_2)) / R_\eta^{0,i}(\theta, X(Z^{(i)}, \theta_2)),\ 1 \right\}\,.$$

Alternatively, we can use local algorithms to update the auxiliary variables by doing local perturbations to $Z^{(i)}$ which leave the dominating measure $\mathbb{W}(t_{i-1}, t_i, 0, 0)$ invariant, see for example the algorithms reviewed in [5].

Practical implementation of the algorithm requires a finite-dimensional approximation. Using the ideas from Section 2 this amounts to approximating the dominating measure by $\pi(\mathrm{d}\theta) \otimes_{i=1}^n \mathbb{W}^\delta(t_{i-1}, t_i, 0, 0)$, hence the auxiliary variables by the skeletons $\{Z^{(i)}_{\tau_{i,j}}\}$ for $\tau_{i,j}$ equally spaced at distance δ between t_{i-1} and t_i, and each factor $R^{0,i}_\eta$ by $R^{\delta,i}_\eta$, computed as in Section 2.5 with the obvious modifications.

3.3 The irreducible case

The same arguments as for the reducible case suggest the following auxiliary distribution

$$\prod_{i=1}^n R^{0,i}(\theta, V^{(i)}) \bigotimes_{i=1}^n \mathbb{Q}(t_{i-1}, t_i, V_{t_{i-1}}; \theta_2, V_{t_i})\, \pi(\mathrm{d}\theta) \tag{28}$$

where

$$R^{0,i}(\theta, V^{(i)}) = (2\pi(t_i - t_{i-1}))^{-d/2} \left|A(t_i, V_{t_i}; \theta_2)\right|^{1/2} G(t_{i-1}, t_i, V^{(i)}, b, A; \theta) \zeta(t_{i-1}, t_i, V^{(i)}, A; \theta_2)$$
$$\times \exp\left\{-\frac{1}{2t_i - t_{i-1}} (V_{t_i} - V_{t_{i-1}})^T A(t_{i-1}, V_{t_i}; \theta_2)(V_{t_i} - V_{t_{i-1}})\right\}$$

is defined as $R^0(\theta)$ in Section 2.3 but over the time period $[t_{i-1}, t_i]$, $V^{(i)} = (V^{(i)}_t, t_{i-1} \leq t \leq t_i)$ for $i = 1, \ldots, n$ are the auxiliary variables, which are processes that interpolate the observations, and the dependence of $R^{0,i}$ on each process $V^{(i)}$ is made explicit in the notation. It is again immediate consequence of (13) that this auxiliary measure admits $\pi(\mathrm{d}\theta \mid \{V_{t_i}\})$ as a marginal.

Note that under this auxiliary measure and conditionally on θ, the $V^{(i)}$'s are independent and each distributed according to a measure proportional to $R^{0,i}(\theta, V^{(i)})\mathbb{Q}(t_{i-1}, t_i, V_{t_{i-1}}; \theta_2, V_{t_i})$. Recall from Section 2.3 and Equation (12), that this measure is the diffusion bridge measure $\mathbb{P}(t_{i-1}, t_i, V_{t_{i-1}}, V_{t_i}; \theta)$, i.e., the law of a process that solves the SDE (10). Due to the quadratic variation identity (3), these laws for different θ_2's will typically be mutually singular, since different θ_2's will typically imply different quadratic variation processes. Thus, for different θ_2's the conditional auxiliary measures will be mutually singular, hence, as with (25) in the reducible case, component-wise updating algorithms that target the joint auxiliary measure will not be ergodic. This problem was first pointed out by [21].

As with the reducible case, we will overcome this problem by seeking a change of measure in the transition density identity (13). In the reducible case we obtained this change by linearly transforming the Brownian bridge and we then approximated the resulting infinite-dimensional measure for practical purposes. The possibility to identify changes of measure in infinite-dimensional spaces will be harder in more general contexts, for example for irreducible diffusions.

The third main idea we introduce in this section, is that we can exploit the connection between discrete and continuous-time results by reversing the operations of change of measure and approximation. That is, we can first approximate (28), building on the results of Sections 2.4 and 2.5, and then devise a change of measure. For example, we can approximate the dominating measure by $\pi(\mathrm{d}\theta) \otimes_{i=1}^n \mathbb{Q}^\delta(t_{i-1}, t_i, V_{t_{i-1}}; \theta_2, V_{t_i})$, although we could have used an alternative approximation of each $\mathbb{Q}(t_{i-1}, t_i, V_{t_{i-1}}; \theta_2, V_{t_i})$, e.g. by the Euler scheme. Given the advantages of the local linearisation of (17) over alternatives discussed in Sections 2.4 and 2.6, we will concentrate on $\mathbb{Q}^\delta$ in the rest of this section. We also approximate each $R^{0,i}(\theta, V^{(i)})$ by $R^{\delta,i}(\theta, \{Y^{\delta,(i)}_{\tau_{i,j}}\})$ as suggested in Section 2.5.

Component-wise MCMC algorithms on the resulting finite-dimensional auxiliary distribution will be ergodic, but the mixing time will deteriorate as $\delta \to 0$; see Section 2.3 of [21] for a worked example and [24] for a rigorous evaluation of how small δ needs to be for the resulting approximation bias to be insignificant, in models where MCMC can be applied without such approximation error.

To combat the algorithmic deterioration with δ, we can perform a change of measure in the finite-dimensional auxiliary distribution, which is rather straightforward. One possibility is to work with $\pi(\mathrm{d}\theta) \otimes_{i=1}^{n} \mathbb{W}^{\delta}(t_{i-1}, t_i, 0)$. Let $\{Z^{(i)}_{\tau_{i,j}}\}$ be the new auxiliary variables, which under each dominating measure $\mathbb{W}^{\delta}(t_{i-1}, t_i, 0)$ are Brownian skeletons. Let $Y^{\delta}(\{Z^{(i)}_{\tau_{i,j}}\}, \theta_2)$ denote the discrete-time process produced by (18) where $Z^{(i)}_{\tau_{i,j}} - Z^{(i)}_{\tau_{i,j-1}}$ are used as the noise increments. We can then define the new auxiliary distribution

$$\prod_{i=1}^{n} R^{\delta,i}(\theta, Y^{\delta}(\{Z^{(i)}_{\tau_{i,j}}\}, \theta_2)) \bigotimes_{i=1}^{n} \mathbb{W}^{\delta}(t_{i-1}, t_i, 0)\, \pi(\mathrm{d}\theta)\,. \tag{29}$$

The conditional density of θ given the auxiliary variables is proportional to

$$\pi(\theta) \prod_{i=1}^{n} R^{\delta,i}(\theta, Y^{\delta}(\{Z^{(i)}_{\tau_{i,j}}\}, \theta_2))$$

and the auxiliary variables conditionally on θ are independent, each distributed according to a measure proportional to $R^{\delta,i}(\theta, Y^{\delta}(\{Z^{(i)}_{\tau_{i,j}}\}, \theta_2))\mathbb{W}^{\delta}(t_{i-1}, t_i, 0)$. The simulation from these conditionals can be done along the suggestions made for the reducible case. In particular, an independence sampler is one option for sampling the auxiliary variables, according to which proposals are generated according to $\{\widetilde{Z^{(i)}_{\tau_{i,j}}}\} \sim \mathbb{W}^{\delta}(t_{i-1}, t_i, 0)$ and accepted with probability

$$\min\left\{ R^{\delta,i}(\theta, Y^{\delta}(\widetilde{\{Z^{(i)}_{\tau_{i,j}}\}}), \theta_2))/R^{\delta,i}(\theta, Y^{\delta}(\{Z^{(i)}_{\tau_{i,j}}\}), \theta_2)), 1 \right\}.$$

We can now link this presentation to previous results in the literature on MCMC methods for partially observed irreducible diffusions. The specific approximation of $\mathbb{Q}$ by $\mathbb{Q}^{\delta}$, together with the independence sampler for the auxiliary variables is closely related with the algorithm of [12]. [12] start from a discrete-time missing data perspective, and inspired by the modified Brownian bridge proposal they (effectively) consider the auxiliary distribution

$$\prod_{i=1}^{n} R^{\delta,i}_{E}(\theta, \{Y^{\delta,(i)}_{\tau_{i,j}}\}) \bigotimes_{i=1}^{n} \mathbb{Q}^{\delta}(t_{i-1}, t_i, V_{t_{i-1}}; \theta_2, V_{t_i})\, \pi(\mathrm{d}\theta)\,.$$

where $R^{\delta,i}_{E}(\theta, \{Y^{\delta,(i)}_{\tau_{i,j}}\})$ is defined analogously to R^{δ}_{E} in Section 2.6 but over $[t_{i-1}, t_i]$. To combat the anticipated mixing problems they perform the transformation discussed above and work instead with the auxiliary distribution (29) but with $R^{\delta,i}(\theta, Y^{\delta}(\{Z^{(i)}_{\tau_{i,j}}\}, \theta_2))$ replaced by $R^{\delta,i}_{E}(\theta, Y^{\delta}(\{Z^{(i)}_{\tau_{i,j}}\}, \theta_2))$. Proposition 3 shows that the two auxiliary distributions, that of [12] and the one we discussed above, are slightly different. However, it is useful to realise that this is only one of the possible algorithms for posterior inference, other approximation or sampling schemes can be considered.

3.4 Robustness of posterior distributions and MCMC algorithms to approximations

In this paper we have managed to construct on the same probability space stochastic processes, likelihood estimates and MCMC algorithms that can be obtained either for the

continuous-time model and then be approximated or for a discrete-time approximation of the model. This joint construction provides an ideal framework for studying the properties of these processes and inference algorithms as $\delta \to 0$, i.e., as the approximation to the continuous-time limit improves. We have obtained such concrete results for the stochastic processes involved in this framework, see Theorem 1. What is beyond the scope of this paper, and is missing from the literature, is a similar study of the properties of the algorithms. We are particularly interested in the stability properties of the finite-dimensional MCMC algorithms discussed in this section. In the reducible case there exists a limiting algorithm, which we approximated, but there exist no general concrete results on the convergence properties of the Markov chain that operates on the infinite-dimensional space, neither for that of the approximating chains. In the irreducible case, we do not even have a formal result establishing the limit of the finite-dimensional algorithms to an infinite-dimensional one.

Consider first the case of reducible diffusions, for which an infinite-dimensional sampler is well-defined. It is known that approximating a Markov chain kernel does not guarantee good approximation of the corresponding limiting distribution. Moreover, even when the limiting distribution of the discrete chains do converge to the desired continuous time limit, the MCMC algorithm on the discretisation space may not share similar convergence characteristics (for example geometric ergodicity) as the idealised chain operating in the continuous time limit. See for example [22, 4] for a discussion of these issues, counter-examples where the expected convergences fail to hold, and positive results showing robustness of both the target distribution and the convergence properties of the MCMC chain. For irreducible diffusions we face the additional challenge to establish that a limit is well defined.

Unfortunately, results to guarantee robustness are not readily available from existing theory in the literature, particularly as the state spaces of the approximating chains are invariably changing as discretisations become increasingly finer. However it seems clear that methods based on total variation bounds and associated coupling arguments (as used for example in the papers mentioned above) ought to provide some reassurance, perhaps in the presence of weak additional regularity conditions. The embedding of the discrete-time approaches within a continuous-time framework and the results on convergence of the stochastic processes involved are important steps in this direction. We also have some preliminary results on the convergence of the weight R^δ to its limit R^0. Giving precise statements of results in this direction is beyond the scope of this paper, but it will be the focus of ongoing work.

4 Discussion

In this paper we introduce a generic framework for data augmentation for diffusions. There are some key ideas that underpin this framework. One is the embedding of different stochastic processes that have been used for simulation of conditioned diffusions within a common framework, that of different discretisations of (11). We demonstrated the strong convergence of appropriate discretisations of these processes to their desired limits.

We also emphasise important identities for the diffusion transition density and the respective discretisations. These are developed by simple albeit not completely trivial stochastic analysis tools, in particular the decomposition of measure (8) and the Cameron-Martin-Girsanov theorem. The embedding of discrete and continuous-time approaches in a common framework allowed us to study the connection between different estimators of the likelihood for diffusions. Proposition 3 shows that apparently different schemes are closely related and the discussion that follows it makes a case for the optimality of some of those.

A third key idea is that these identities can be used to find auxiliary distributions within a data augmentation framework for Bayesian inference for partially observed diffusions. This allows us to obtain expressions for the joint distribution of parameters and auxiliary variables in a rather simple and automatic way bypassing complications that can be found in previous works on this topic. The resulting auxiliary distributions do not lend themselves to Gibbs sampling, due to the mutual singularity of the conditional distributions of the auxiliary variables. Nevertheless, another key idea in this paper is a change of measure which results in auxiliary distributions that can be sampled by the Gibbs sampler or other component-wise updating MCMC algorithms.

Our methods and theoretical results rely on certain assumptions on the coefficients of (1). Those in Appendix B for the proof of Theorem 1 are merely technical. The assumptions that σ and A are bounded, which can be used to establish that the importance sampling approximation (12) is valid, are probably unnecessarily strong. The assumption of ellipticity of (1) is of course more restrictive and not only of technical nature. In general, it is hopeless to study importance sampling for hypoelliptic diffusion bridges without assumptions on the structure of σ. However, for specific (and at the same time practically relevant) type of hypoelliptic diffusions, e.g. integrated diffusions, importance sampling representations such as those obtained in Section 2.3 should be possible, as well as the transition density identities that then form the backbone of the MCMC methodology. It is also interesting and challenging to study such representations and identities for other Markov processes, e.g. jump diffusions. These questions form part of our current research agenda.

The derivation of such representations and identities is to some extent model specific and it might require careful probabilistic analysis. However, and we think this is the appeal of this work, once these results are obtained they can be quite easily utilised within MCMC algorithms for parameter estimation when the underlying processes are partially observed.

More generally, a lot of the ideas in this paper can be considered in broader terms, without explicit reference to diffusions per se, but of relevance to other high or infinite-dimensional Bayesian inference contexts. In particular, let $\pi_0(\mathrm{d}\theta)$ be the prior distribution of the parameters of a statistical model. Then, Bayesian inference is based upon the posterior distribution, say $\pi_1(\mathrm{d}\theta) \propto L(\theta)\pi_0(\mathrm{d}\theta)$, where $L(\theta)$ is the likelihood. Implicit in this construction is the statistical model from which the likelihood is derived from, as well as the dependence of $L(\theta)$ on observed data, which has been suppressed for notational economy. Data augmentation methods come into play when $L(\theta)$ is intractable or expensive to compute, but auxiliary variables V that take values on a space $\mathcal{V}$, and an expanded distribution $\pi(\mathrm{d}\theta, \mathrm{d}V)$ can be identified, such that $\pi(\mathrm{d}\theta, \mathrm{d}V)$ admits $\pi_1(\mathrm{d}\theta)$ as a marginal. The aim is then to sample from this auxiliary distribution using component-wise updating MCMC algorithms. The first challenge with DA is to identify such auxiliary expansions. The approach we developed in Section 2 in terms of likelihood identities provides one generic way to achieve this. In particular, suppose we have that $L(\theta) = \mathbb{E}_{\mathbb{Q}_\theta}[G(V;\theta)]$, where the equality holds $\pi_0 - a.s$, G is positive, and the expectation is taken with respect to a distribution $\mathbb{Q}_\theta$ on $\mathcal{V}$, which as the notation suggests might depend on θ. Then, we can work with the auxiliary expansion $\pi(\mathrm{d}\theta, \mathrm{d}V) \propto G(V;\theta)\mathbb{Q}_\theta(\mathrm{d}V)\pi_0(\mathrm{d}\theta)$, which by construction admits π_1 as a marginal. The second challenge with DA arises when the measures $\mathbb{Q}_\theta$ for different θ's are mutually singular, i.e., for $\theta_1 \neq \theta_2$, $\mathbb{Q}_{\theta_i}(\mathrm{d}V) > 0 \implies \mathbb{Q}_{\theta_j}(\mathrm{d}V) = 0$ for $i \neq j$. In that case, any component-wise updating MCMC algorithm is not ergodic. The change of measure approach introduced in Section 3 can provide the solution in such frameworks. Finally, when $\mathcal{V}$ is an infinite-dimensional space an approximation will be needed for practical implementation. Again, our results can be helpful in analyzing methods that either discretise a limiting algorithm or they start by a finite-dimensional

approximation of the statistical model.

Supplementary Materials

Appendix: Appendix A gives details for the derivation of the likelihood (12). Appendix B gives the proof of Theorem 1 and Proposition 1. Appendix C gives the proof of Proposition 3. (data_aug_diffusions_appendix.pdf).

Acknowledgements

Papaspiliopoulos would like to acknowledge financial support by the Spanish government through the research grant MTM2009-09063. Roberts would like to thank EPSRC for support through grants EP/20620/01 and EP/S61577/01. The authors thank Anders C. Jensen and an anonymous referee for helpful suggestions.

References

[1] E. Allen. *Modeling with Itô stochastic differential equations*, volume 22. Springer Verlag, 2007.

[2] A. Beskos, O. Papaspiliopoulos, and G.O. Roberts. Monte Carlo maximum likelihood estimation for discretely observed diffusion processes. *Annals of Statistics*, 37:223–245, 2009.

[3] A. Beskos, O. Papaspiliopoulos, G.O. Roberts, and P. Fearnhead. Exact and computationally efficient likelihood-based estimation for discretely observed diffusion processes. *JR Statist. Soc. B*, 68(part 3):333–382, 2006.

[4] Laird Breyer, Gareth O. Roberts, and Jeffrey S. Rosenthal. A note on geometric ergodicity and floating-point roundoff error. *Statist. Probab. Lett.*, 53(2):123–127, 2001.

[5] S. L. Cotter, G. O. Roberts, A. M. Stuart, and D. White. MCMC methods for functions: modifying old algorithms to make them faster. Submitted, 2012.

[6] B. Delyon and Y. Hu. Simulation of conditioned diffusion and application to parameter estimation. *Stochastic Process. Appl.*, 116(11):1660–1675, 2006.

[7] S. Ditlevsen, A.C. Jensen, M. Kessler, and O. Papaspiliopoulos. A Markov Chain Monte Carlo approach to parameter estimation in the FitzHugh-Nagumo model. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, 86(4):art. no. 041114, 2012.

[8] G. B. Durham and A. R. Gallant. Numerical techniques for maximum likelihood estimation of continuous-time diffusion processes. *J. Bus. Econom. Statist.*, 20(3):297–338, 2002. With comments and a reply by the authors.

[9] K.D. Elworthy. *Stochastic differential equations on manifolds*. Cambridge University Press, 1982.

[10] D. Florens-Zmirou. Approximate discrete-time schemes for statistics of diffusion processes. *Statistics*, 20(4):547–557, 1989.

[11] E. Gobet and C. Labart. Sharp estimates for the convergence of the density of the euler scheme in small time. *Elect. Comm. in Probab.*, 13:352–363, 2008.

[12] A. Golightly and D. J. Wilkinson. Bayesian inference for nonlinear multivariate diffusion models observed with error. *Computational Statistics and Data Analysis*, 52(3):1674–1693, 2008.

[13] J Jacod. Statistics and high frequency data. In *Statistical Methods for Stochastic Differential Equations*, pages 191–309. Monographs on Statistics and Applied Probability, Chapman and Hall, 2012.

[14] P.E. Kloeden and E. Platen. *Numerical Solution of Stochastic Differential Equations*. Springer-Verlag, 1995.

[15] G. N. Milstein. *Numerical integration of stochastic differential equations*, volume 313 of *Mathematics and its Applications*. Kluwer Academic Publishers Group, Dordrecht, 1995. Translated and revised from the 1988 Russian original.

[16] O. Papaspiliopoulos and G. O. Roberts. Importance sampling techniques for estimation of diffusion models. In *Statistical Methods for Stochastic Differential Equations*, pages 311–337. Monographs on Statistics and Applied Probability, Chapman and Hall, 2012.

[17] Omiros Papaspiliopoulos, Gareth O. Roberts, and Martin Sköld. Non-centered parameterizations for hierarchical models and data augmentation. In *Bayesian statistics, 7 (Tenerife, 2002)*, pages 307–326. Oxford Univ. Press, New York, 2003. With a discussion by Alan E. Gelfand, Ole F. Christensen and Darren J. Wilkinson, and a reply by the authors.

[18] Omiros Papaspiliopoulos, Gareth O. Roberts, and Martin Sköld. A general framework for the parametrization of hierarchical models. *Statist. Sci.*, 22(1):59–73, 2007.

[19] Asger Roer Pedersen. Consistency and asymptotic normality of an approximate maximum likelihood estimator for discretely observed diffusion processes. *Bernoulli*, 1(3):257–279, 1995.

[20] Asger Roer Pedersen. A new approach to maximum likelihood estimation for stochastic differential equations based on discrete observations. *Scand. J. Statist.*, 22(1):55–71, 1995.

[21] G. O. Roberts and O. Stramer. On inference for partially observed nonlinear diffusion models using the Metropolis-Hastings algorithm. *Biometrika*, 88(3):603–621, 2001.

[22] Gareth O. Roberts, Jeffrey S. Rosenthal, and Peter O. Schwartz. Convergence properties of perturbed Markov chains. *J. Appl. Probab.*, 35(1):1–11, 1998.

[23] L. C. G. Rogers and David Williams. *Diffusions, Markov processes, and martingales. Vol. 1.* Cambridge Mathematical Library. Cambridge University Press, Cambridge, 2000. Foundations, Reprint of the second (1994) edition.

[24] G. Sermaidis, O. Papaspiliopoulos, G.O. Roberts, A. Beskos, and P. Fearnhead. Markov chain Monte Carlo for exact inference for diffusions. *Scandinavian Journal of Statistics*, to appear, 2012.

[25] Isao Shoji. A note on convergence rate of a linearization method for the discretization of stochastic differential equations. *Commun. Nonlinear. Sci. Numer. SI.*, 16(7):2667–2671, 2011.

[26] Isao Shoji and Tohru Ozaki. A statistical method of estimation and simulation for systems of stochastic differential equations. *Biometrika*, 85:240–243, 1998.

[27] S.E. Shreve. *Stochastic Calculus for Finance II: Continuous-Time models*. Springer Finance, 2008.