Methodology and Computing in Applied Probability
https://doi.org/10.1007/s11009-022-09949-y

The Computational Cost of Blocking for Sampling Discretely Observed Diffusions

Marcin Mider[1] · **Paul A. Jenkins**[2] · **Murray Pollock**[3] · **Gareth O. Roberts**[2]

Received: 22 February 2021 / Revised: 5 November 2021 / Accepted: 10 March 2022

Abstract
Many approaches for conducting Bayesian inference on discretely observed diffusions involve imputing diffusion bridges between observations. This can be computationally challenging in settings in which the temporal horizon between subsequent observations is large, due to the poor scaling of algorithms for simulating bridges as observation distance increases. It is common in practical settings to use a *blocking scheme*, in which the path is split into a (user-specified) number of overlapping segments and a Gibbs sampler is employed to update segments in turn. Substituting the independent simulation of diffusion bridges for one obtained using blocking introduces an inherent trade-off: we are now imputing shorter bridges at the cost of introducing a dependency between subsequent iterations of the bridge sampler. This is further complicated by the fact that there are a number of possible ways to implement the blocking scheme, each of which introduces a different dependency structure between iterations. Although blocking schemes have had considerable *empirical* success in practice, there has been no analysis of this trade-off nor guidance to practitioners on the particular specifications that should be used to obtain a computationally efficient implementation. In this article we conduct this analysis and demonstrate that the expected computational cost of a blocked path-space rejection sampler applied to Brownian bridges scales asymptotically at a cubic rate with respect to the observation distance and that this rate is linear in the case of the Ornstein–Uhlenbeck process. Numerical experiments suggest applicability both of the results of our paper and of the guidance we provide beyond the class of linear diffusions considered.

Keywords Bayesian inference · Blocking · Diffusion · Gaussian process · Markov chain Monte Carlo

Mathematics Subject Classification 60J22 · 65C05

✉ Gareth O. Roberts
gareth.o.roberts@warwick.ac.uk

[1] Max Planck Institute for Mathematics in the Sciences, 04103 Leipzig, Germany

[2] Department of Statistics, University of Warwick, Coventry CV4 7AL, U.K.

[3] School of Mathematics, Statistics and Physics, Newcastle University, Newcastle-upon-Tyne NE1 7RU, U.K.

Published online: 14 July 2022

1 Introduction

Diffusions have been widely applied to model continuous-time phenomena of interest, including molecular dynamics (Boys et al. 2008), neuroscience (Lansky and Ditlevsen 2008), and finance (Karatzas and Shreve 1998). In general, a diffusion on $\mathbb{R}^d$ is a Markov process X defined to be the solution, with law we will denote by $\mathbb{P}$, to a stochastic differential equation of the following form:

$$\mathrm{d}X_t = b(X_t)\mathrm{d}t + \sigma(X_t)\mathrm{d}W_t, \quad X_0 = x_0, \quad t \in [0, T], \tag{1}$$

where $b : \mathbb{R}^d \to \mathbb{R}^d$ and $\sigma : \mathbb{R}^d \to \mathbb{R}^{d\times d'}$ denote the drift and volatility coefficient respectively, and W is a standard d'-dimensional Brownian motion. Throughout we assume standard regularity conditions hold which ensure the existence of a unique, global, weak solution to (1) (see for instance Øksendal 2003).

In practice we will typically only have access to discrete observations of (1), and so for practitioners the statistical problem of interest is to use these observations to draw inference on the parameters of b and σ of (1). A common Bayesian strategy is to augment the parameter space with the space describing the complete underlying diffusion trajectory. A Markov Chain Monte Carlo algorithm can then explore this augmented space by alternating between updates of the parameters and updates of the unobserved sample path connecting observations (sampling of diffusion bridges) (Roberts and Stramer 2001). As a consequence a considerable and methodologically diverse literature has been developed concerned with simulating diffusion bridges (the law of (1) conditioned to terminate at the subsequent observation (for instance, $X_T = x_T$), which we denote by $\mathbb{P}^{(T,x_0,x_T)}$ or generically by $\mathbb{P}^\star$), including Beskos and Roberts (2005); Bladt et al. (2014); Delyon and Hu (2006); Durham and Gallant (2002); Golightly and Wilkinson (2008); Hairer et al. (2011); Roberts and Stramer (2001); Schauer et al. (2017).

One of the common difficulties with Markov Chain Monte Carlo strategies is sampling diffusion bridges between distant observations; the *duration* of the bridge, which we denote by T, is large. This setting naturally arises when the underlying diffusion (1) is *sparsely* observed (or high-dimensional), for instance in shape analysis applications (Arnaudon et al. 2020), or in the case of diffusions on graphs (Freidlin and Wentzell 1993). The problem here is that methodologies for sampling diffusion bridges scale poorly with T, and many of the most widely used approaches have *exponential computational cost* in T. Consequently, addressing the poor scaling in T has drawn considerable interest. One popular approach is the *blocking* scheme introduced by Shephard and Pitt (1997), which has been employed in a number of practical problems with strong empirical evidence of its efficacy (Chib et al. 2004; Golightly and Wilkinson 2008; Kalogeropoulos 2007; Kalogeropoulos et al. 2010; van der Meulen and Schauer 2018; Stramer and Roberts 2007).

Blocking is a conceptually simple idea in which the time domain of the diffusion bridge is overlaid with a set of *temporal anchors* ($0 =: k_0 < k_1 < \cdots < k_m < k_{m+1} := T$), and the values of the bridge are taken for some *initialisation* trajectory at those points (which are known as *knots*, and for which we will denote $X_i := X_{k_i}$ to simplify notation). Simulation from $\mathbb{P}^{(T,x_0,x_T)}$ is then achieved by constructing a Gibbs sampler which alternates between updating knots and updating the segments of the trajectory conditional on the knots, a number of times. For instance, we could begin by simulating from the conditional law $\mathbb{P}^{(k_2-k_0,X_0,X_2)}$ (updating the trajectory between $[t_0, t_2]$ which includes the knot at X_1, conditional on the knots at X_0 and X_2), and then $\mathbb{P}^{(k_3-k_1,X_1,X_3)}$ (updating the trajectory between $[t_1, t_3]$ and containing the knot X_2, conditional on the knots at X_1 and X_3), and so

on *sweeping* across all anchor points. This sweep would then be iterated a number of times to reduce the dependency between the resulting bridge and the initial (or previous) trajectory. In this article we consider the three canonical blocking schemes of Roberts and Sahu (1997) with equidistant anchors: the *checkerboard* scheme, in which the odd and even indexed knots are alternatively updated; the *lexicographic* scheme, in which the knots are updated in temporal order; and the *random* scheme, in which at each step a random knot is updated. We more formally introduce blocking and define these schemes in Sect. 2.

From a computational perspective, blocking substitutes the expensive simulation of a (single independent) draw from $\mathbb{P}^{(T,x_0,x_T)}$, with the cost of simulating repeated sweeps of the $m+1$ shorter (and computationally more efficient) bridges for each segment given by the temporal anchors. Any analysis of this trade-off needs to take into account the serial correlation induced by the blocking strategy.

Despite widespread adoption of blocking in practice to mitigate the computational cost of simulating diffusion bridges (as indicated above), there is little theoretical support for its efficacy. Furthermore, there is little concrete guidance on how to implement, and then appropriately tune (selecting for instance the number and locations of the anchor points), a blocking scheme.

In this article we provide general guidance for implementing blocking schemes by addressing these practical considerations for particular classes of diffusion process. We analyse the computational cost of several rejection sampling algorithms for bridges as a function of block size and bridge duration. In all cases we consider a fixed regular spacing of m anchor points as $m, T \to \infty$, in contrast with the study of the *'in-fill'* asymptotic of Roberts and Stramer (2001) in which T is fixed and $m \to \infty$. We analyse the expected cost of a single iteration of various algorithms, and then to capture the trade-off described above we consider the cost of the algorithm which comprises both the cost of one iteration, and the total number of iterations required to obtain an 'independent' sample. We give a more formal description of what we mean by achieving independence below, in terms of the relaxation time of the underlying Markov chain.

In this article we work under the assumption that the underlying measure is a Gaussian diffusion (i.e. $\mathbb{P}$ is the law of a scaled Brownian motion or the law of the Ornstein–Uhlenbeck process). Under this simplification the Gibbs step for updating the bridge segments can be implemented without error, i.e. without discretising time, for example by means of a rejection sampler directly on the path-space of the diffusion (see Appendix 1 for full details). In this setting we prove that Theorem 1 below holds, as the culmination of the results in Sect. 3. We gather all proofs in the appendices.

Theorem 1 *Suppose $\mathbb{P}^\star$ is the conditional law of a Gaussian diffusion which is sampled by rejection on path-space and using a checkerboard or lexicographic or random blocking scheme. Suppose the m anchors are spaced equidistantly such that $m = c_1 T$ (for some constant $c_1 > 0$). Then the expected computational cost of the blocked rejection sampler, $C_{blocking}(T)$, satisfies*:

$$C_{blocking}(T) = \mathcal{O}(T^3), \quad \text{as} \quad T \to \infty, \tag{2}$$

whenever $\mathbb{P}$ denotes the law of a scaled Brownian motion and

$$C_{blocking}(T) = \mathcal{O}(T), \quad \text{as} \quad T \to \infty, \tag{3}$$

whenever $\mathbb{P}$ denotes the law of the Ornstein–Uhlenbeck process.

Remark 1 Note that in the case of a Brownian bridge there is long range dependency in the path, in the sense that the correlation between X_s and X_t is non-negligible even for $0 \ll s \ll t \ll T$. On the other hand, for the Ornstein–Uhlenbeck process its ergodicity breaks this dependency. For an Ornstein–Uhlenbeck process whose drift is of the form $b(X_t) = -\theta X_t$ and for $T \gg 0$, there is a phase transition in its behaviour as $\theta \to 0$ in that the computational cost of a blocked rejection sampler for Brownian motion is *not* recovered. Recall that in this paper we are working under the assumption that the underlying measure is a Gaussian diffusion, but in most practical settings the *target law* will be more complicated. In such settings it would be typical to use a Gaussian diffusion as a *proposal law* for the non-Gaussian target law. In principle Theorem 1 would suggest that an Ornstein–Uhlenbeck process proposal for a *stationary* target law would be advantageous over a Brownian bridge proposal, although in practice this predicted computational saving would depend on how closely the target process matched the invariant distribution of the Ornstein–Uhlenbeck process.

Theorem 1 contrasts sharply with the case without blocking. We show later in Proposition 1 that, for a d-dimensional Brownian bridge proposal in the absence of blocking, the cost is exponential in T. Although what we prove in Theorem 1 addresses a somewhat idealised setting, the requirement $m = c_1 T$ acts as a concrete guide for choosing the number of blocks. Furthermore, our empirical results in Sect. 4 indicate that the guidance we establish can be more broadly useful beyond the class of linear diffusions. Thus we demonstrate that blocking can lead to significantly improved computational efficiency when conducting inference for discretely observed diffusions.

2 Blocking

In this section we provide a systematic definition of blocking for sampling a diffusion path. Define a set of anchors spread across the time domain: $0 < k_1 < \cdots < k_m < T$ and *knots* as the values of the path taken at the anchors:

$$\mathcal{K}(\omega) := \{X_{k_1}(\omega), \ldots, X_{k_m}(\omega)\}.$$

Each anchor is now assigned to one of $\mathbb{k}$ disjoint subsets $\mathcal{A}_i$, $i = 1, \ldots, \mathbb{k}$, each comprising m_i anchors:

$$\{k_1, \ldots, k_m\} = \dot{\bigcup_{i=1}^{\mathbb{k}}} \{r_{i1}, \ldots, r_{im_i}\} =: \dot{\bigcup_{i=1}^{\mathbb{k}}} \mathcal{A}_i, \quad (\text{with } m_i \in \mathbb{N}_+, i \in \{1, \ldots, \mathbb{k}\}).$$

In particular, $\mathcal{A}_i = \{r_{i1}, \ldots, r_{im_i}\}$ is the set of anchors associated (uniquely) to the ith block. This allows us to group the knots by associating them with the corresponding subsets of anchors:

$$\mathcal{K}_i(\omega) := \{X_r(\omega); r \in \mathcal{A}_i\}, \quad i \in \{1, \ldots, \mathbb{k}\}.$$

For convenience of notation we let $\mathcal{K}_{-i}$ (resp. $\mathcal{A}_{-i}$) denote all knots (resp. anchors) that do not belong to $\mathcal{K}_i$ (resp. $\mathcal{A}_i$):

$$\mathcal{K}_{-i}(\omega) := \bigcup_{j \neq i} \mathcal{K}_j(\omega), \qquad \mathcal{A}_{-i}(\omega) := \dot{\bigcup_{j \neq i}} \mathcal{A}_j(\omega), \qquad i \in \{1, \ldots, \mathbb{k}\},$$

and assign labels to an ordered collection of all anchors in $\mathcal{A}_{-i}$, plus the end-points:

$$\{e_{i0}, \ldots, e_{i(m+1-m_i)}\} = \mathcal{A}_{-i} \cup \{0, T\}, \quad (\text{with } e_{ij} < e_{i(j+1)}), \quad i \in \{1, \ldots, \mathbb{k}\}.$$

Further, define

$$\mathcal{B}_i := \left\{ (e_{ij}, e_{i(j+1)}) \,\middle|\, \exists r \in \mathcal{A}_i \text{ s.t. } r \in [e_{ij}, e_{i(j+1)}] \right\}_{j=0}^{m-m_i},$$

to be only those intervals between the end-points or anchors in $\mathcal{A}_{-i}$, which contain at least one anchor belonging to $\mathcal{A}_i$. The path segments $X|_{\mathcal{B}_i}$, obtained through restricting X to $\mathcal{B}_i$, are termed *blocks*. Finally, in the case $\mathbb{k} = 2$ we say that $\mathcal{A}_1$ and $\mathcal{A}_2$ are *interlaced* if whenever $a, c \in \mathcal{A}_i$, with $a < c$, then there exists $b \in \mathcal{A}_{(i \mod 2)+1}$ s.t. $a < b < c$, $i = 1, 2$.

A sampler for a path equipped with a blocking technique is a Gibbs sampler that updates the full path only one block at a time by drawing from the conditional laws $\mathbb{P}^\star|_{\mathcal{B}_i}(\cdot|\mathcal{K}_{-i})$—i.e. the target laws restricted to blocks $\mathcal{B}_i$ and conditioned on the knots in $\mathcal{K}_{-i}$. For simplicity we refer to this technique as a *blocked sampler* in the remainder of the text, and present general pseudo-code for it in Algorithm 1.

Algorithm 1. Blocked sampler on path-space

Initialise X;
for $n = 1, \ldots, N$ **do**
  for $i = 1, \ldots, \mathbb{k}$ **do**
    Draw $I \sim q(i, \cdot)$ (various choices for q are defined below, in Definitions 1–3);
    Update $X|_{\mathcal{B}_I}$ by sampling $X|_{\mathcal{B}_I} \sim \mathbb{P}^\star|_{\mathcal{B}_I}(\cdot|\mathcal{K}_{-I})$;
return X

There are a number of ways we can update the blocks, and in this article we consider the three canonical blocking schemes of Roberts and Sahu (1997). In particular, we refer to a single, full Gibbs sweep of Algorithm 1 (the inner `for-loop`) as a:

Definition 1 *Checkerboard* blocking update scheme if $\mathbb{k} = 2$, $\mathcal{A}_1$ and $\mathcal{A}_2$ are interlaced, and $q(i, j) := \mathbb{1}_{\{i\}}(j)$.

Definition 2 *Lexicographic* blocking update scheme if $\mathbb{k} = m$, $\mathcal{A}_i := \{k_i\}$, $i \in \{1, \ldots, m\}$, and $q(i, j) := \mathbb{1}_{\{i\}}(j)$.

Definition 3 *Random* blocking update scheme if $\mathbb{k} = m$, $\mathcal{A}_i := \{k_i\}$, $i \in \{1, \ldots, m\}$, and $q(i, j) := \frac{1}{m}\mathbb{1}_{\{1,\ldots,m\}}(j)$.

The above are not exhaustive, but characterise the most widely used, and are tractable enough for analysis. We further simplify various computations by assuming the anchors are *equidistant*, and defer discussion of this assumption and its relaxation to Sect. 5.

Assumption 1 The anchors are placed on an equidistant grid:

$$k_{i+1} - k_i = \frac{T}{m+1} =: \delta_{m,T}, \quad i \in \{1, \ldots, m-1\}.$$

As mentioned in the Introduction, we will study the asymptotic regime in which $\delta_{m,T}$ is fixed as $m, T \to \infty$.

3 Computational Analysis

3.1 Cost of a Single Sweep

We begin by quantifying the computational cost of a rejection sampling algorithm for diffusion bridges in the absence of blocking. The setting here, as given in further detail in Appendix 1, is one in which the target law is absolutely continuous with respect to a d-dimensional Brownian bridge proposal path.

Proposition 1 *Under Assumptions 3–9 enumerated in Appendix 1, the expected computational cost as a function of T of obtaining a single draw with a path-space rejection sampling algorithm, denoted by $C_{\mathtt{rej}}(T)$, is given by*

$$C_{\mathtt{rej}}(T) = f(T)Te^{c_2 T}, \tag{4}$$

where $c_2 > 0$ is some constant independent of T, and the function $f : \mathbb{R}_+ \to \mathbb{R}$ is continuous and such that $f(T) \sim T^{-d/2}$ as $T \to \infty$. In particular, for large enough T there is a constant $c_3 > 0$ such that:

$$C_{\mathtt{rej}}(T) \geq c_3 T^{1-d/2} e^{c_2 T}.$$

Remark 2 Note that Proposition 1 does not stipulate in what way the constant c_3 might vary with dimension. Without further structure it is impossible to characterise this behaviour. However it is highly likely that c_3 will increase at least linearly with dimension (as for instance would be the case for diffusions consisting of d independent components).

Remark 3 If $\mathbb{P}$ is the law of a drifted Brownian motion, then Proposition 1 cannot be applied directly, because Assumption 9 does not hold. However, for this case an easy calculation shows that the acceptance probability of a rejection sampler with Brownian bridge proposals is equal to 1, implying (under Assumption 8) that $C_{\mathtt{rej}}(T)$ is proportional to T.

Now considering a single sweep of the blocking schemes introduced in Sect. 2, note that we have substituted sampling a single diffusion bridge (of length T) with sampling a number of diffusion bridges of shorter time horizon, $2\delta_{m,T}$ (for example, to ensure the point $X_{2\delta_{m,T}}$ is updated one could sample a new bridge of length $2\delta_{m,T}$ connecting $X_{\delta_{m,T}}$ with $X_{3\delta_{m,T}}$). By application of Proposition 1, the expected computational cost of simulating each of these shorter bridges is therefore $C_{\mathtt{rej}}(2\delta_{m,T})$, and hence the expected cost of a single Gibbs sweep is:

$$C_{\mathtt{sweep}}(T, m) := m \cdot C_{\mathtt{rej}}(2\delta_{m,T}) = f(2\delta_{m,T}) \frac{2mT}{m+1} \exp\{2c_2\delta_{m,T}\}. \tag{5}$$

Equation (5) holds for all m and T and follows from (4); however, as the behaviour of $f(t)$ for small t is not immediately transparent, to learn something about $C_{\texttt{sweep}}(T,m)$ when $\delta_{m,T}$ is small, we may use the fact that the acceptance probability of the rejection sampler approaches 1 as the bridge duration decreases to 0. This fact implies that for small enough t, $C_{\texttt{rej}}(t) \sim c_5 t$ and thus if $\delta_{m,T} < c_4$ as $T \to \infty$ for some constant c_4 then

$$C_{\texttt{sweep}}(T,m) \sim c_5 T, \tag{6}$$

for some $c_5 > 0$. For instance, upon setting $m = \lfloor T \rfloor$, the cost in (5) becomes $\mathcal{O}(T)$ as $T \to \infty$. Contrast this with (4) to see that the relative gain in efficiency, $C_{\texttt{rej}}(T)/C_{\texttt{sweep}}(T,m)$ grows exponentially in T and suggests that blocking is to be preferred for large enough T. However, this ignores the costs associated with mixing; we address this in the next subsection.

3.2 Cost of Multiple Sweeps

Direct comparison of the exponential cost $C_{\texttt{rej}}(T)$ of direct rejection sampling (as given by Proposition 1), with the linear cost $C_{\texttt{sweep}}(T,m)$ of a single sweep of a blocking scheme (as given by (6)), does not capture the remnant dependency structure introduced by the blocking scheme. In addition we need to consider the number of sweeps required to render this dependency negligible. In order to do that we first introduce the following notion

Definition 4 (Roberts and Sahu 1997) The [$\mathcal{L}^2$-]*convergence rate* ρ of a Markov chain $\{X^{(n)}; n = 1, \ldots, N\}$ with the transition kernel P and an invariant density π is defined as the minimum number for which for all square π-integrable functions f, and for all $r > \rho$

$$\|P^n f - \pi(f)\|_{\mathcal{L}^2(\pi)} := \int \left[P^n f(X^{(0)}) - \pi(f)\right]^2 \pi(\mathrm{d}X^{(0)}) \leq V_f r^n,$$

where $P^n f(X^{(0)}) := \mathbb{E}_\pi[f(X^{(n)})|X^{(0)}]$, $\pi(f) := \mathbb{E}_\pi[f(X)]$ and V_f is a positive number that depends on f.

We can now capture the cost of reducing the dependency on the past by considering the *relaxation time*, denoted $\mathcal{T} = \mathcal{T}(T,m)$, and defined as:

$$\mathcal{T} = -\frac{1}{\log(\rho)}. \tag{7}$$

It represents the time required by the underlying Markov chain to output a draw from its stationary distribution (Levin and Peres 2017). This makes it possible to compare $C_{\texttt{rej}}(T)$ with the expected computational cost of the blocked rejection sampler as follows:

$$C_{\texttt{blocking}}(T,m) := \mathcal{T}(T,m) \cdot C_{\texttt{sweep}}(T,m). \tag{8}$$

We will later consider the most appropriate choice of blocking scheme, and how to optimise m.

Instead of analysing the chain targeting the law $\mathbb{P}^\star$ it is sufficient to consider a related chain that targets the marginal law of the vector

$$\mathcal{G} := (X_{k_1}, \ldots, X_{k_m})|(X_0, X_T) = \mathcal{K}|(X_0, X_T), \tag{9}$$

which we denote by $\mathbb{G}$. To see this, notice that conditionally on the knots $\mathcal{K}$ being distributed according to $\mathbb{G}$, a path X returned after a single Gibbs sweep of a blocking scheme is distributed exactly according to $\mathbb{P}^\star$. The object of interest becomes a Markov chain with a transition kernel P denoting a single Gibbs sweep, and with stationary distribution $\pi = \mathbb{G}$ (Roberts and Rosenthal 2001).

Throughout, we additionally assume that the following condition holds, which makes the subsequent required calculations tractable.

Assumption 2 The target law $\mathbb{P}^\star$ is such that $\mathcal{G}$ is a Gaussian process.

We discuss this key technical assumption in Sect. 4, where we note that the established results seem to hold empirically more broadly.

Under Assumption 2 and using either the lexicographic or checkerboard updating scheme, a single Gibbs step (i.e. an update $\mathcal{G}|_{\mathcal{B}_l} \sim \mathbb{P}^\star|_{\mathcal{B}_l \cap \{X_{k_1}, \ldots, X_{k_m}\}}(\cdot|\mathcal{K}_{-l})$) has a tractable, Gaussian transition density, and thus so does the entire Gibbs sweep $\mathcal{G}^{(n)} \mapsto \mathcal{G}^{(n+1)}$ with mean and covariance

$$\mu := \mathbb{E}[\mathcal{G}], \qquad \Sigma := \mathbb{C}ov[\mathcal{G}].$$

As a consequence it is possible to explicitly characterise the transition kernel P, as follows.

Lemma 1 *Under the lexicographic and checkerboard updating schemes, the n-step transition kernel P^n of the Markov chain $\{\mathcal{G}^{(l)}; l = 0, \ldots\}$ is Gaussian, with mean and covariance matrix given respectively by*:

$$\mathbb{E}[\mathcal{G}^{(l+n)}|\mathcal{G}^{(l)}] = B^n\mathcal{G}^{(l)} + (I - B)^{-1}(I - B^n)b, \quad \mathbb{C}ov[\mathcal{G}^{(l+n)}|\mathcal{G}^{(l)}] = \Sigma - B^n\Sigma(B^n)^{\mathrm{T}}, \quad (10)$$

with $B \in \mathbb{R}^{m\times m}$ and $b \in \mathbb{R}^m$.

Under the lexicographic or the checkerboard updating schemes $\{\mathcal{G}^{(l)}; l = 0, \ldots\}$ is an AR(1) process, and so the spectral radius $\rho_{\mathtt{spec}}(B)$ of the matrix B must satisfy $\rho_{\mathtt{spec}}(B) < 1$ for the process to converge, and equals the $\mathcal{L}^2$-convergence rate (Amit 1991). This connection extends to the random updating scheme. In the following lemma we derive the spectral radius of each blocking scheme as a function of m and T, which aids in optimising their parameterisation and analysing their scaling. We denote by $\Lambda := \Sigma^{-1}$ the precision matrix of $\mathcal{G}$ and define

$$A := I - \text{ diag } \{\Lambda_{11}^{-1}, \ldots, \Lambda_{mm}^{-1}\}\Lambda.$$

Lemma 2 (Roberts and Sahu 1997) *Under the checkerboard and lexicographic updating schemes, the spectral radius of the matrix B and the $\mathcal{L}^2$-convergence rate of a blocked rejection sampler coincide. More explicitly, under the checkerboard, lexicographic, and random updating schemes respectively the $\mathcal{L}^2$-convergence rates ($\rho_{\mathtt{check}}$, $\rho_{\mathtt{lex}}$, and $\rho_{\mathtt{rand}}$ resp.) are equal to*:

$$\rho_{m,T} := \rho_{\mathtt{check}} = \rho_{\mathtt{lex}} = \rho_{\mathtt{spec}}(B_{\mathtt{lex}}) = \rho_{\mathtt{spec}}(B_{\mathtt{check}})$$
$$= \lambda^2_{\mathtt{max}}(A), \rho_{\mathtt{rand}} = \left[\frac{m - 1 + \lambda_{\mathtt{max}}(A)}{m}\right]^m,$$

where $\lambda_{\texttt{max}}(A)$ *denotes the maximum eigenvalue of the matrix A and where we write* $B_{\texttt{check}}$ (*resp.* $B_{\texttt{lex}}$) *to denote a matrix B corresponding to the checkerboard* (*resp. lexicographic*) *updating scheme.*

$\lambda_{\texttt{max}}(A)$ can be found more explicitly by exploiting the close connection between the precision matrix Λ and the matrix of partial correlations (given precisely in (19), in Appendix 2).

Theorem 2 *We have*

$$\lambda_{\texttt{max}}(A) = 2|c(\delta_{m,T})| \cos\left(\frac{\pi}{m+1}\right),$$

with $c(\delta_{m,T}) := \mathbb{C}orr(X_\delta, X_{2\delta}|X_0, X_{3\delta})$. *In particular*:

$$\rho_{m,T} = 4c^2(\delta_{m,T}) \cos^2\left(\frac{\pi}{m+1}\right), \qquad \rho_{\texttt{rand}} = \left[\frac{m-1+2|c(\delta_{m,T})| \cos\left(\frac{\pi}{m+1}\right)}{m}\right]^m. \quad (11)$$

The form of $c(\delta_{m,T})$ will, in general, depend on the type of a Gaussian process that is being considered. In the following corollaries we present more explicit versions of the statements from Theorem 2 for the two choices of $\mathbb{P}$: scaled Brownian motion σW, with $\sigma > 0$; and, the Ornstein–Uhlenbeck process. Without loss of generality we centre the latter at 0:

$$\mathrm{d}X_t = -\theta X_t \mathrm{d}t + \sigma \mathrm{d}W_t, \quad X_0 = x_0, \quad t \in [0, T]. \quad (12)$$

Corollary 1 *If* $\mathbb{P}$ *is the law of a scaled Brownian motion* $\sigma W, \sigma > 0$, *then*:

$$\rho_{m,T} = \cos^2\left(\frac{\pi}{m+1}\right), \quad \rho_{\texttt{rand}} = \left[\frac{m-1+\cos\left(\frac{\pi}{m+1}\right)}{m}\right]^m.$$

In particular, independently of T, *as* $m \to \infty$

$$\rho_{m,T} = 1 - \left(\frac{\pi}{m+1}\right)^2 + \mathcal{O}(m^{-4}), \qquad \rho_{\texttt{rand}} = 1 - \frac{1}{2}\left(\frac{\pi}{m+1}\right)^2 + \mathcal{O}(m^{-4}).$$

Corollary 2 *If* $\mathbb{P}$ *is the law of the Ornstein–Uhlenbeck process* (12), *then*:

$$\rho_{m,T} = \cos^2\left(\frac{\pi}{m+1}\right) \mathrm{sech}^2(\theta\delta_{m,T}), \quad \rho_{\texttt{rand}} = \left[\frac{m-1+\cos\left(\frac{\pi}{m+1}\right)\mathrm{sech}(\theta\delta_{m,T})}{m}\right]^m.$$

In particular, when $\delta_{m,T} = \delta$ *is set to a constant, as* $m, T \to \infty$:

$$\rho_{m,T} = \operatorname{sech}^2(\theta\delta)\left[1 - \left(\frac{\pi}{m+1}\right)^2 + \mathcal{O}(m^{-4})\right],$$
$$\rho_{\mathtt{rand}} = e^{\operatorname{sech}(\theta\delta)-1}\left[1 - \frac{(1-\operatorname{sech}(\theta\delta))^2}{2(m+1)} + \mathcal{O}(m^{-2})\right].$$

Remark 4 Results obtained by Pitt and Shephard (1999), who studied the discrete-time first-order autoregressive process $\alpha_t = \phi\alpha_{t-1} + \eta_t$, $\eta_t \sim \mathcal{N}(0, \sigma^2)$, observed with Gaussian noise, are closely related to Corollaries 1 and 2. In the context of their model, where $c(\delta_{m,T}) = \phi/(1+\phi^2)$, they derive the expression (11) for $\rho_{m,T}$ as well as bounds on $\rho_{m,T}$ which exhibit the same asymptotic behaviour as in Corollary 2.

We can now combine the above results with (7) to find the relaxation time:

Theorem 3 *Suppose we use one of the checkerboard, lexicographic, and random updating schemes. If $\mathbb{P}$ is the law of a scaled Brownian motion σW, then we have*:

$$\mathcal{T}(m) = \mathcal{O}(m^2), \qquad m \to \infty.$$

If $\mathbb{P}$ is the law of the Ornstein–Uhlenbeck process in (12), and additionally the sequence $T(m)$ is chosen so that $m = c_1 T$ for some constant $c_1 > 0$, then we have:

$$\mathcal{T}(m) = \mathcal{O}(1), \qquad m \to \infty.$$

Remark 5 Note that if $\mathbb{P}$ is the law of the Ornstein–Uhlenbeck process in (12) then Theorem 3 holds for only the sequence $T(m)$ where $m = c_1 T$, but in the case of scaled Brownian motion there is no such constraint; see Remark 1.

Remark 6 From the proof of Theorem 3, one can show that for the Ornstein–Uhlenbeck process (12):

$$\mathcal{T}(m) = -\frac{1}{\log(\rho_{m,T})} \to -\frac{1}{2\log(\operatorname{sech}(\theta\delta))}, \qquad m \to \infty.$$

This provides insight into the influence of θ and δ on mixing.

We can minimize the cost of blocking $C_{\mathtt{blocking}}(T, m)$ over the remaining parameter, m, using Theorem 3 and (8). This leads to Theorem 1, which is the main result of this paper (as presented in Sect. 1, with accompanying proof in Appendix 2).

4 Numerical Experiments

Consider a target process defined to be the solution of the following stochastic differential equation (with law $\mathbb{P}$):

$$\mathrm{d}X_t = (2 - 2\sin(8X_t))\mathrm{d}t + \frac{1}{2}\mathrm{d}W_t, \quad X_0 = 0, \quad t \in [0, T]. \tag{13}$$

This diffusion exhibits highly multimodal behaviour, and so in practice it is challenging to simulate trajectories of $\mathbb{P}$ (and in particular the conditioned bridge law $\mathbb{P}^\star$ over large time horizons). It is possible to simulate trajectories exactly by means of path-space rejection

sampling (as detailed in Appendix 1). However, *X is not* a Gaussian process (it violates Assumption 2), and so Theorem 1 does not hold in a rigorous sense. As such (13) makes an interesting case to investigate the practical limitations of Theorem 1. Because it is not an ergodic diffusion, out of the two theoretical results from Theorem 1 the ones for the Brownian motion are expected to be more relevant. As we show below, the empirical results would suggest the theory holds more broadly.

We consider six problems (increasing in difficulty) of simulating paths according to the laws $\mathbb{P}^{(T,x_0,x_T)}$, with parameters $\mathbb{P}^{(0.2,0,0.1)}$, $\mathbb{P}^{(0.4,0,0.85)}$, $\mathbb{P}^{(0.5,0,0.85)}$, $\mathbb{P}^{(1,0,0.95)}$, $\mathbb{P}^{(2,0,2.5)}$, $\mathbb{P}^{(4,0,4.85)}$. The values of the end-points were chosen by fixing T, simulating multiple paths according to (13) and picking x_T to be some point in the vicinity of the (largest) mode as these are the bridges we will most commonly be interested in. For $T = 0.2$, the plotted paths resemble Brownian bridges, but as T increases the non-linear dynamics become pronounced: the diffusion is effectively attracted to a ladder of values and it is repelled at the intermediate points, leading to multimodal behaviour of the trajectories. Drawing paths from the last three laws using path-space rejection sampling but without blocking (an *unmodified rejection sampler*) is computationally infeasible.

For each of the six examples we ran a blocked rejection sampler with checkerboard updating scheme for 10^5 iterations and with various numbers of knots. For the first three problems we also employed an unmodified rejection sampler. We recorded the time required to sample a single path (which for a blocked rejection sampler is counted as one execution of the inner `for-loop` of Algorithm 1) and plotted it in Fig. 1 against the number of used knots. Code sufficient for reproducing these results can be found at https://github.com/mmider/blocking.

For $T = 0.2$ the unmodified rejection sampler clearly outperforms any blocking scheme. This is unsurprising as paths under $\mathbb{P}^{(0.2,0,0.1)}$ closely resemble Brownian bridges (and indeed every diffusion behaves as a drifted Brownian motion on a small-enough time-scale). However, as T increases, this pattern changes and blocking reduces the cost of obtaining any single sample path. In particular, notice a steep, exponential reduction in cost that is especially pronounced for $(T, x_T) = (1, 0.95)$ (this would be illustrated even more emphatically by $(T, x_T) = (2, 2.5)$ and $(T, x_T) = (4, 4.85)$ had the corresponding

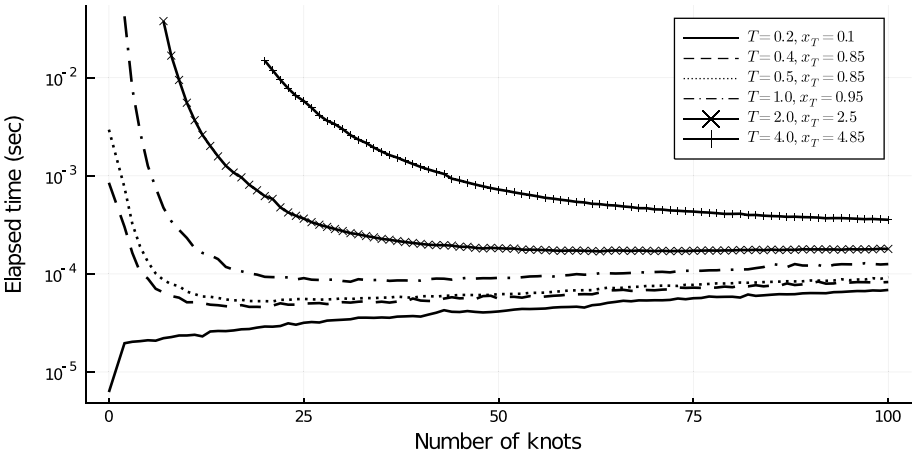

Fig. 1 Time (in seconds; log-transformed) required to sample a single path of the sine diffusion (13) as a function of the number of used knots

experiments with a lower number of knots been run; however, their costs are prohibitively high and had to be omitted).

Figure 1, though helpful in confirming Proposition 1, does not take into account the cost due to decreased speed of mixing—the main motivation for the developments presented in Sect. 3. To incorporate also this cost we plot in Fig. 2 the time-adjusted effective sample size (taESS), with

$$\text{taESS} := [\text{effective sample size}]/[\text{elapsed time in seconds to sample an entire chain}]$$

and ESS was computed according to Gelman et al. (2013, Section 11.5) against the (half-) length of blocks (i.e. $\delta_{m,T}$). As defined in Fig. 2, taESS is approximately equal to a number of independent samples that can be drawn in one second. Clearly, the larger taESS is the more efficient the algorithm is.

First, for any experiment we expect there to be a point for which increasing the number of knots any further will only lead to a decrease in taESS—this corresponds to all costs being dominated by the cost due to a slowdown in mixing and it is clearly illustrated by sharp dips of curves on the left side of Fig. 2. Second, for examples for which the target law is sufficiently different from the law of Brownian bridges we expect that some level of blocking will improve the overall computational cost. This is also confirmed by the declines of taESS curves toward the right side of Fig. 2. We note that under the most difficult sampling regimes it was impractical to run the algorithm with even fewer blocks due to excessive execution times—had the examples been run and the curves continued, the decline in performance would have been even starker. Additionally, Fig. 2 is suggestive of there being an optimal value of $\delta_{m,T}$ (somewhere around $\delta_{m,T} \approx 0.1$), that is almost independent of T and m and that yields the highest taESS in each experiment. This is consistent with the results of Sect. 3, where an optimal number of knots was found to be $m = c_1 T$ for some $c_1 > 0$, which implies the claim about the dependence of the optimal $\delta_{m,T}$ on T and m.

Finally, we verify the bound from (2) empirically. To this end, notice that taESS^{-1} is approximately equal to the amount of time needed to obtain a single independent sample. This is consistent with the characterisation of the computational cost of a blocked rejection sampler as given in (8). Theorem 1 asserts that this cost scales at a cubic rate in the

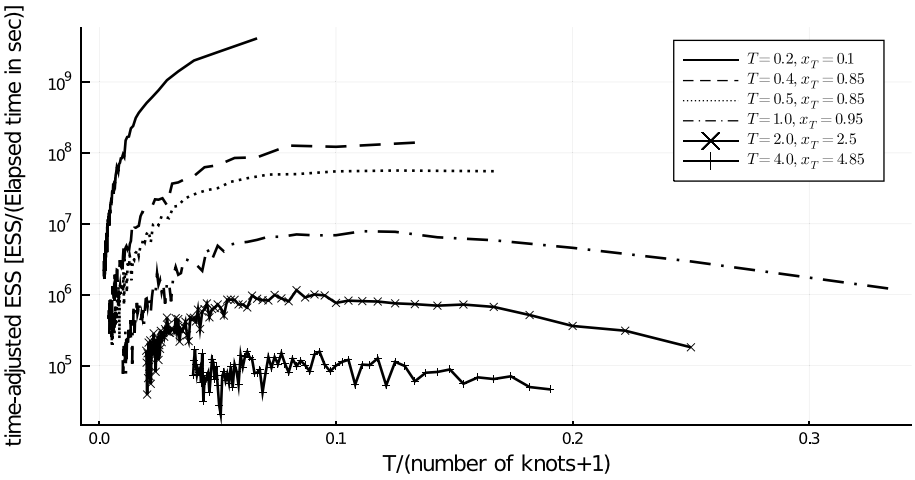

Fig. 2 Time-adjusted effective sample size vs half-length of blocks (i.e. $\delta_{m,T}$)

duration of the bridge, so long as $\delta_{m,T}$ is set to a constant when $T \to \infty$. Consequently, taESS(T) should be at most a cubic function of T and if plotted on a log-log scale, this would be equivalent to taESS(T) tracking some line with slope 3. Figure 3 gives this precise plot, showing that the prediction (2) is indeed satisfied.

5 Discussion

In this article we have analysed and provided practical guidance for using blocking schemes when conducting Bayesian inference for discretely observed diffusions. We achieved this by studying the computational cost of diffusion bridge sampling algorithms. We have shown rigorously that the computational cost of rejection sampling on path-space (modified with blocking) targeting the law of scaled Brownian motion scales as $\mathcal{O}(T^3)$ as $T \to \infty$, and as $\mathcal{O}(T)$ in the case of the Ornstein–Uhlenbeck process, so long as the number of equidistant anchors is $m = c_1 T$ (for some $c_1 > 0$). In Remark 1 we discussed the practicality of exploiting the computational saving achievable in the case of the Ornstein–Uhlenbeck process. Furthermore, using the example of a non-linear sine diffusion we provide empirical evidence which would suggest that the conclusions about Brownian motion hold also for non-ergodic diffusions outside of a restrictive class of Gaussian processes.

Our theory indicates that choosing too few knots results in the computational cost being dominated by the exponential cost for imputing diffusion bridges between successive knots (see Proposition 1). As such our guideline of choosing $m = c_1 T$, (for some $c_1 > 0$) is useful for ensuring the robustness of blocking schemes and a reasonable heuristic for practitioners. Note that although choosing too many knots is likely to be penalized less than choosing too few, choosing an excessive number of knots can negatively impact the mixing of the underlying chain.

Naturally, for more general target laws $\mathbb{P}$ it might be useful to consider using irregularly spaced anchors (and so relaxing Assumption 1). Heuristically, we may wish to place more knots in areas in which the proposal law does not approximate the target law well. Developing more general theory to support the use of an irregular spacing

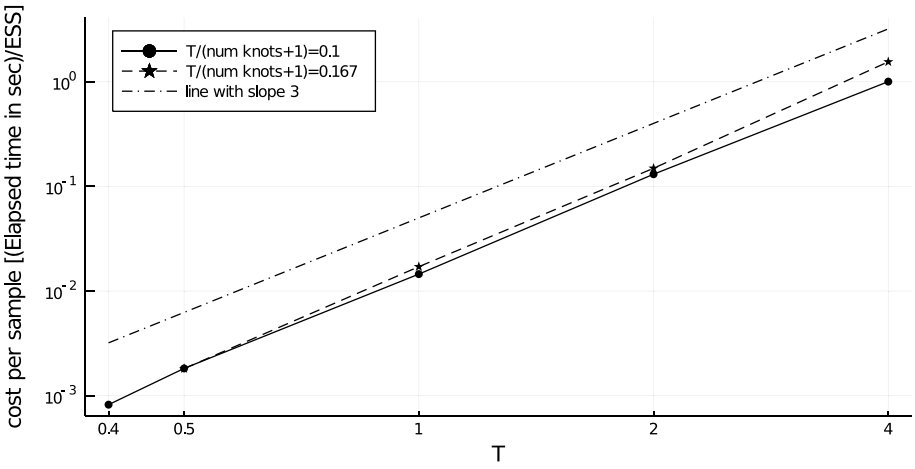

Fig. 3 Computational cost as a function of time for the sine example

of anchors is likely to require more knowledge of the specific diffusion under study. Of course, from a methodological perspective this motivates future research looking at how to place knots by assessing proposal-target discrepancy, or developing adaptive schemes.

Finally, it is worth recalling that within the context of Bayesian inference for discretely observed diffusion processes, the full chain in this setting is a Gibbs sampler that alternates between updating the unknown parameters and imputing the unobserved path. Since the mixing time of the unobserved path influences the mixing time of the parameter chain, then in light of the work in this paper it may as a future extension also be possible to study the mixing behaviour of the parameter chain.

Appendix 1

Rejection Sampling on Path-space

In this article we have restricted our attention to the class of diffusion bridges which can be sampled by means of *path-space rejection sampling*. In particular, to sample from $\mathbb{P}^\star$ we sample trajectories from an accessible and absolutely continuous *proposal law* (denoted $\mathbb{Q}^*$), and accept with probability proportional to the Radon-Nikodým derivative of $\mathbb{P}^\star$ to $\mathbb{Q}^*$ (Beskos and Roberts 2005; Beskos et al. 2006; Beskos et al. 2008; Pollock et al. 2016). To find an appropriate $\mathbb{Q}^*$ we impose the following common assumption (Kloeden and Platen 2013, Section 4.4)

Assumption 3 There exists $\eta : \mathbb{R}^d \to \mathbb{R}^d$ such that $\nabla\eta = \sigma^{-1}$.

Under Assumption 3 the process $Y := \{\eta(X_t), t \in [0,T]\}$ satisfies the following stochastic differential equation,

$$\mathrm{d}Y_t = \alpha(Y_t)\mathrm{d}t + \mathrm{d}W_t, \quad Y_0 = y_0 := \eta(x_0), \quad t \in [0,T],$$

for a known, closed-form drift α. With unit volatility, the law of Y is now absolutely continuous with respect to Brownian motion, and so Brownian motion is a viable proposal law for sampling from $\mathbb{P}$ (and the law induced by the Brownian bridge is a viable proposal for $\mathbb{P}^\star$).

In order to avoid unnecessary inflation of notation we assume throughout the article that $\sigma \equiv 1$ in (1), so that $\alpha \equiv b$, η becomes an identity map, and $X \equiv Y$. The general case of σ (that satisfies Assumption 3) follows without additional effort.

Assumption 4 α is at least once continuously differentiable.

Assumption 5 There exists a potential function $A : \mathbb{R}^d \to \mathbb{R}$ such that $\nabla A = \alpha$.

Assumption 6 The function $\phi(y) := \frac{1}{2}\left(\|\alpha(y)\|^2 + \Delta A(y)\right)$ is bounded from below by some $\Phi := \inf\{\phi(y) : y \in \mathbb{R}^d\} \in \mathbb{R}$.

Under Assumptions 4–6 we have (Beskos and Roberts 2005, Section 3):

$$\frac{\mathrm{d}\mathbb{P}^\star}{\mathrm{d}\mathbb{Q}^*}\left(\{\eta^{-1}(Y_t), t \in [0,T]\}\right) \propto \exp\left\{-\int_0^T (\phi(Y_t) - \Phi)\mathrm{d}t\right\} =: \mathrm{p}(Y) \leq 1. \tag{14}$$

It follows that sampling from $\mathbb{P}^\star$ can be accomplished using Algorithm 2. Note that computing the integral in (14) required for Algorithm 2 can be achieved either (i) approximately, by simulating a candidate Y° over a fine mesh and computing the integral in (14) numerically; or (ii) exactly, via an additional randomisation step that utilises a Poisson point process (Beskos and Roberts 2005; Beskos et al. 2006; Beskos et al. 2008). We refer to these two methods as, respectively, *approximate* and *exact* path-space rejection samplers. If we contrast $U \sim \texttt{Bern}(\mathrm{p}(Y))$ with that of $\widetilde{U} \sim \texttt{Bern}(\tilde{\mathrm{p}}(Y))$ where $\tilde{\mathrm{p}}$ is the (unbiased) estimator resulting from the additional randomisation step, then since $\mathbb{E}[\tilde{\mathrm{p}}] = \mathrm{p}$ we have $U \overset{d}{=} \widetilde{U}$ and so $\mathbb{V}(U) = \mathbb{V}(\widetilde{U})$, and there is no additional introduction of variance from the use $\tilde{\mathrm{p}}$ in place of p (Łatuszyński et al. 2011).

Algorithm 2. Rejection sampling on path-space

while `True` **do**
  Draw $Y^\circ \sim \mathbb{W}^{(T,\eta(x_0),\eta(x_T))}$, i.e. a d-dimensional Brownian bridge joining $\eta(x_0)$ and $\eta(x_T)$ on $[0,T]$;
  Draw $U \sim$ `Unif([0,1])`;
  if $U \leq p(Y^\circ)$ **then**
    Set $X \leftarrow \{\eta^{-1}(Y_t^\circ), t \in [0,T]\}$;
    return X

To prove Proposition 1, which quantifies the computational cost of Algorithm 2, we impose the following natural assumptions on the cost of simulating a proposal trajectory.

Assumption 7 The cost of generating any proposal sample X is independent of the value of $\mathrm{p}(X)$ as defined in (14).

Assumption 8 The cost $\mathbb{c}(X)$ of simulating X has expectation growing linearly in T: $\mathbb{E}[\mathbb{c}(X)] = c_5 T$, $c_5 \in \mathbb{R}_+$.

Assumptions 7 and 8 are always satisfied if rejection sampling is performed with the approximate method described above, so long as the mesh width is kept constant as $T \to \infty$. For the exact method, Assumption 7 will in general be violated (for instance, if the number of simulated Poisson points is 0, then conditional on this information $\mathrm{p}(X) = 1$ a.s.) but for $T \to \infty$ it is a reasonable approximation. Assumption 8 is satisfied if ϕ is bounded.

We can now derive the cost of a single draw using a path-space rejection sampler as follows:

Lemma 3 *Under Assumptions 3–8:*

$$C_{\texttt{rej}}(T) = c_6 \frac{q_T(x_0, x_T)}{p_T(x_0, x_T)} T e^{-\Phi T},$$

where $c_6 > 0$ is some constant independent of T, $p_T(x_0, x_T)$ is the transition density under $\mathbb{P}$ for going from x_0 to x_T over the interval $[0, T]$ and $q_T(x_0, x_T)$ is the same transition density, but under the proposal law $\mathbb{Q}$ instead.

Proof Denote by $X^{(i)}$, $i \in \{1, 2, \dots\}$, independent samples from $\mathbb{Q}^*$ and by $\mathbb{c}(X^{(i)})$ the cost of sampling path $X^{(i)}$, $i \in \{1, 2, \dots\}$. Rejection sampling requires a geometrically distributed number of simulations (with a randomly distributed parameter at each trial), so its expected cost is

$$\begin{aligned}\mathcal{C}_{\texttt{rej}}(T) &:= \mathbb{E}\left[\sum_{i=1}^{\infty}\left(\sum_{j=1}^{i}\mathbb{c}(X^{(j)})\right)\cdot \mathrm{p}(X^{(i)})\prod_{j=1}^{i-1}\left(1-\mathrm{p}(X^{(j)})\right)\right]\\ &= \mathbb{E}\left[\mathbb{c}(X^{(1)})\right]\sum_{i=1}^{\infty} i\mathbb{E}\left[\mathrm{p}(X^{(i)})\right]\prod_{j=1}^{i-1}\left(1-\mathbb{E}\left[\mathrm{p}(X^{(j)})\right]\right)\\ &= \frac{\mathbb{E}[\mathbb{c}(X)]}{\mathbb{E}\left[\mathrm{p}(X)\right]}, \end{aligned} \tag{15}$$

where the measures with respect to which the expectations above are taken should be clear from the context (and include Brownian bridge measures, products of Brownian bridge measures, and any additional randomness needed to simulate events of probability $\mathrm{p}(X^{(i)})$). We now have:

$$\mathbb{E}_{\mathbb{Q}^*}\left[\mathrm{p}(X)\right] = \mathbb{E}_{\mathbb{Q}^*}\left[\exp\left\{-\int_0^T(\phi(X_t)-\Phi)\mathrm{d}t\right\}\right] \tag{16}$$

$$\begin{aligned}&= \mathbb{E}_{\mathbb{Q}^*}\left[\exp\left\{\left[A(X_T)-A(X_0)\right]-\left[A(X_T)-A(X_0)\right]-\int_0^T(\phi(X_t)-\Phi)\mathrm{d}t\right\}\right]\\ &= \exp\left\{-A(X_T)+A(X_0)+\Phi T\right\}\frac{p_T(x_0,x_T)}{q_T(x_0,x_T)}\mathbb{E}_{\mathbb{Q}^*}\left[\frac{\mathrm{d}\mathbb{P}^\star}{\mathrm{d}\mathbb{Q}^*}(X)\right]\\ &= c_7\frac{p_T(x_0,x_T)}{q_T(x_0,x_T)}e^{\Phi T}, \end{aligned} \tag{17}$$

where $c_7 := \exp\left\{-A(x_T)+A(x_0)\right\}$ and where the third equality followed from Dacunha-Castelle and Florens-Zmirou (1986, Eq (3.1)). The result now follows by substituting (17) into (15) and noting that, by Assumption 8, we have $\mathbb{E}[\mathbb{c}(X)] = c_5 T$. □

To better understand the scaling with T of the ratio of transition densities under the laws $\mathbb{P}$ and $\mathbb{Q}$ in Lemma 3 we impose the following final assumption, which allows us to establish Lemmata 4 and 5 required for proving Proposition 1.

Assumption 9 The target diffusion is ergodic and defined on $\mathbb{R}^d$.

Lemma 4 *Under Assumption 9, for*

$$f : T \to \frac{q_T(x_0,v)}{p_T(x_0,v)}, \quad v \in \mathbb{R}^d,$$

we have that $f(T) \sim T^{-d/2}$ as $T \to \infty$ and d denotes the dimension of the process.

Proof q and p are well-behaved densities which are bounded and bounded away from zero (see for example Rogers (1985)). This implies that f is continuous. As the target diffusion is

ergodic, $p_T(x_0, v) \to \hat{p}(v)$ as $T \to \infty$, where $\hat{p}$ is the stationary density of the diffusion law. On the other hand $q_T(x_0, v)$ is just a Gaussian density with variance $T^d I$, which for $T \to \infty$ behaves as $\sim T^{-d/2}$. □

Lemma 5 *Assumption 9 implies that* $\Phi < 0$.

Proof From Lemma 4 we have that the RHS of (17) is $\sim T^{d/2} e^{\Phi T}$ for $T \to \infty$. As the LHS of (17) represents an expected probability, we must have $\Phi < 0$ for this expression to take values in [0, 1]. □

We are now in a position to prove Proposition 1.

Proof (Proposition 1) This follows directly by combining Lemmata 3, 4 and 5. □

Appendix 2

Proofs

Proof (Lemma 1) The chain $\{\mathcal{G}^{(l)}; l = 0, \dots\}$ coincides with the chains considered in Roberts and Sahu (1997). In particular, the 1-step transition kernel under lexicographic and checkerboard updating schemes is stated explicitly as Roberts and Sahu (1997, Lemma 1). We provide a proof for completeness.

$\{\mathcal{G}^{(l)}; l = 0, \dots\}$ behaves like an AR(1) process, therefore

$$\mathcal{G}^{(l+1)} = B\mathcal{G}^{(l)} + \epsilon, \qquad \epsilon \sim \mathcal{N}(b, V),$$

for some B, b, and V and

$$\mathcal{G}^{(l+n)} = B^n \mathcal{G}^{(l)} + \epsilon^{(n)}, \qquad \epsilon^{(n)} \sim \mathcal{N}(b^{(n)}, V^{(n)}), \tag{18}$$

with $b^{(n)} := (I + B + \cdots + B^{n-1})b = (I - B)^{-1}(I - B^n)b$ and some $V^{(n)}$ that we are about to derive. Under either scheme b and B can be found in closed form (which we omit for brevity). If the chain has reached stationarity, i.e. if $\mathcal{G}^{(l)} \sim \mathcal{N}(\mu, \Sigma)$, then also $\mathcal{G}^{(l+n)} \sim \mathcal{N}(\mu, \Sigma)$. On the other hand, if $\mathcal{G}^{(l)} \sim \mathcal{N}(\mu, \Sigma)$, then by (18)

$$\mathcal{G}^{(l+n)} | \mathcal{G}^{(l)} \sim \mathcal{N}\big(B^n \mathcal{G}^{(l)} + (I - B)^{-1}(I - B^n)b, B\Sigma B^{\mathrm{T}} + V^{(n)}\big).$$

Consequently:

$$\Sigma = B^n \Sigma (B^n)^{\mathrm{T}} + V^{(n)},$$

and this yields (10). □

Proof (Lemma 2) Since the chain $\{\mathcal{G}^{(l)}; l = 0, \dots\}$ coincides with the chains considered in Roberts and Sahu (1997), the statement of Roberts and Sahu (1997, Theorem 1) applies under checkerboard and lexicographic updating schemes: i.e. the $\mathcal{L}^2$ convergence rates under the two regimes are given by $\rho_{\texttt{check}} = \rho_{\texttt{spec}}(B_{\texttt{lex}})$ and $\rho_{\texttt{lex}} = \rho_{\texttt{spec}}(B_{\texttt{lex}})$ respectively. Due to tridiagonal structure of the precision matrix Λ (which follows from the Markov property of the process $\mathcal{G}$; see also a short explanation in the proof of Theorem 2

that leads up to (20)), Roberts and Sahu (1997, Corollary 3) implies that the two spectral radii coincide, i.e. $\rho_{\mathtt{spec}}(B_{\mathtt{lex}}) = \rho_{\mathtt{spec}}(B_{\mathtt{check}})$. By the same token, Roberts and Sahu (1997, Theorem 5) applies as well, yielding $\rho_{\mathtt{spec}}(B_{\mathtt{check}}) = \lambda^2_{\mathtt{max}}(A)$. Finally, the $\mathcal{L}^2$ convergence rate of the random updating scheme follows from Roberts and Sahu (1997, Theorem 2). □

Proof (Theorem 2) The precision matrix Λ of any random vector $\mathcal{G}$ with non-degenerate covariance matrix can be related to a matrix of partial correlations via (Lauritzen 1996, p. 130):

$$\mathbb{C}orr(\mathcal{G}^{[i]}, \mathcal{G}^{[j]} | \mathcal{G} \backslash \{\mathcal{G}^{[i]}, \mathcal{G}^{[j]}\}) = -\frac{\Lambda^{[i,j]}}{\sqrt{\Lambda^{[i,i]}\Lambda^{[j,j]}}}. \tag{19}$$

By the definition of $\mathcal{G}$ in (9), it is easy to see that $\mathbb{C}orr(\mathcal{G}^{[i]}, \mathcal{G}^{[j]} | \mathcal{G} \backslash \{\mathcal{G}^{[i]}, \mathcal{G}^{[j]}\}) = 0$ whenever $|i-j| > 1$; that by symmetry $\Lambda^{[i,i+1]} = \Lambda^{[i+1,i]}$, $(i = 1, \ldots, m)$; and that $\Lambda^{[i,i]} = \Lambda^{[j,j]}$, $(i, j = 1, \ldots, m)$, because $\mathbb{V}ar(\mathcal{G}^{[i]} | \mathcal{G} \backslash \mathcal{G}^{[i]}) = (\Lambda^{[i,i]})^{-1}$, $(i = 1, \ldots, m)$ (Roberts and Sahu 1997, p. 296). In addition, under Assumption 2, the covariance matrix depends only on time and not on the state variable, thus combining this with Assumption 1: $\mathbb{C}orr(\mathcal{G}^{[i]}, \mathcal{G}^{[i+1]} | \mathcal{G} \backslash \{\mathcal{G}^{[i]}, \mathcal{G}^{[i+1]}\}) =: c(\delta_{m,T})$, $(i = 1, \ldots, m-1)$. Consequently, Λ is a Toeplitz matrix whose non-zero entries are related via $\Lambda^{[i,i+1]} = \Lambda^{[i+1,i]} = -\Lambda^{[i,i]} c(\delta_{m,T})$, $(i = 1, \ldots, m)$. The form of matrix A now follows:

$$A = \begin{pmatrix} 0 & c(\delta_{m,T}) & 0 & \ldots & 0 \\ c(\delta_{m,T}) & 0 & c(\delta_{m,T}) & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & c(\delta_{m,T}) & 0 & c(\delta_{m,T}) \\ 0 & \ldots & 0 & c(\delta_{m,T}) & 0 \end{pmatrix}. \tag{20}$$

The eigenvalues of Toeplitz matrices may be found in closed form (Smith 1985; Kulkarni et al. 1999) and in particular, those of matrix A are given by

$$-2c(\delta_{m,T}) \cos\left(\frac{\pi l}{m+1}\right), \qquad l = 1, \ldots, m.$$

Depending on the sign of $c(\delta_{m,T})$ the maximal eigenvalue of A is therefore given by:

$$\lambda_{\mathtt{max}}(A) = \begin{cases} -2c(\delta_{m,T}) \cos\left(\frac{\pi m}{m+1}\right) = 2c(\delta_{m,T}) \cos\left(\frac{\pi}{m+1}\right), & \text{if } c(\delta_{m,T}) > 0, \\ -2c(\delta_{m,T}) \cos\left(\frac{\pi}{m+1}\right) & \text{if } c(\delta_{m,T}) < 0, \end{cases}$$

and the result concerning $\lambda_{\mathtt{max}}(A)$ follows. The remaining statements follow as well by substituting the expression for $\lambda_{\mathtt{max}}(A)$ into Lemma 2. □

Proof (Corollary 1) By Theorem 2, only $c(\delta_{m,T})$ needs to be computed. This follows from standard properties of Brownian motion and bridges:

$$c(\delta_{m,T}) = \frac{\mathbb{C}ov\left(X_\delta, X_{2\delta} \middle| X_0, X_{3\delta}\right)}{\sqrt{\mathbb{V}ar(X_\delta | X_0, X_{3\delta}) \mathbb{V}ar(X_{2\delta} | X_0, X_{3\delta})}} = \frac{\frac{1}{3}\delta\sigma^2}{\sqrt{\left(\frac{2}{3}\delta\sigma^2\right)^2}} = \frac{1}{2}. \tag{21}$$

The asymptotic behaviour of $\rho_{m,T}$ follows immediately from Taylor expansion of $\cos^2(x)$ around 0. For the asymptotic behaviour of $\rho_{\texttt{rand}}$, notice that by Taylor expansions of $\cos(x)$ around 0, $\log(1-x)$ around 0, and $\exp(x)$ around 0 respectively:

$$\begin{aligned}
\rho_{\texttt{rand}} &= \exp\left\{ m \log\left[m^{-1}\left\{ m - 1 + \cos\left(\frac{\pi}{m+1} \right) \right\} \right] \right\} \\
&= \exp\left\{ m \log\left[1 - \frac{1}{2m}\left(\frac{\pi}{m+1} \right)^2 + \mathcal{O}(m^{-5}) \right] \right\} \\
&= \exp\left\{ -\frac{1}{2}\left(\frac{\pi}{m+1} \right)^2 + \mathcal{O}(m^{-4}) \right\} \\
&= 1 - \frac{1}{2}\left(\frac{\pi}{m+1} \right)^2 + \mathcal{O}(m^{-4}).
\end{aligned}$$

□

Proof (Corollary 2) For the Ornstein–Uhlenbeck process we have:

$$\begin{aligned}
\mathbb{C}ov\left[\begin{pmatrix} Y_s \\ Y_t \end{pmatrix} \middle| Y_0, Y_T \right] &= \frac{\sigma^2}{\theta} \begin{pmatrix} e^{-\theta s}\sinh(\theta s) & e^{-\theta t}\sinh(\theta s) \\ e^{-\theta t}\sinh(\theta s) & e^{-\theta t}\sinh(\theta t) \end{pmatrix} \\
&\quad - \frac{\sigma^2}{\theta} \begin{pmatrix} e^{-\theta T}\frac{\sinh^2(\theta s)}{\sinh(\theta T)} & e^{-\theta T}\frac{\sinh(\theta s)\sinh(\theta t)}{\sinh(\theta T)} \\ e^{-\theta T}\frac{\sinh(\theta s)\sinh(\theta t)}{\sinh(\theta T)} & e^{-\theta T}\frac{\sinh^2(\theta t)}{\sinh(\theta T)} \end{pmatrix}, \quad 0 < s < t < T,
\end{aligned}$$

which for the particular choice of $(s,t,T) = (\delta, 2\delta, 3\delta)$ simplifies to:

$$\mathbb{C}ov\left[\begin{pmatrix} Y_\delta \\ Y_{2\delta} \end{pmatrix} \middle| Y_0, Y_{3\delta} \right] = \frac{\sigma^2}{\theta}\frac{\sinh^2(\theta\delta)}{\sinh(3\theta\delta)} \begin{pmatrix} 2\cosh(\theta\delta) & 1 \\ 1 & 2\cosh(\theta\delta) \end{pmatrix}. \tag{22}$$

It now follows from direct substitution of the relevant terms of (22) into the definition of the partial correlation that:

$$c(\delta_{m,T}) = \frac{\mathbb{C}ov\left(X_\delta, X_{2\delta} | X_0, X_{3\delta}\right)}{\sqrt{\mathbb{V}ar(X_\delta | X_0, X_{3\delta})\mathbb{V}ar(X_{2\delta} | X_0, X_{3\delta})}} = \frac{1}{2\cosh(\theta\delta)}. \tag{23}$$

The remaining steps are analogous to the proof of Corollary 1. □

Proof (Theorem 3) The result follows easily from (7) and Corollaries 1 and 2. For example if $\mathbb{P}$ denotes the law of a scaled Brownian motion then

$$\mathcal{T}(m) = -\frac{1}{\log \rho_{m,T}} = -\left[\log\left(1 - \frac{\pi^2}{(m+1)^2} + \mathcal{O}(m^{-4}) \right) \right]^{-1} = -\left[\frac{\pi^2}{(m+1)^2} + \mathcal{O}(m^{-4}) \right]^{-1} = \mathcal{O}(m^2)$$

as $m \to \infty$, as required. Related expressions for the relaxation time of the Ornstein–Uhlenbeck process, and for the relaxation time relating to $\rho_{\texttt{rand}}$, follow similarly. □

Proof (Theorem 1) This follows from Theorem 3 and (8). We minimize the cost of blocking $C_{\texttt{blocking}}(T, m)$ over the remaining hyperparameter m and derive its final form as a function of T.

1. If $\mathbb{P}$ is the law of the Ornstein–Uhlenbeck process, then the restriction $T = c_1 m$ from Theorem 3 already constraints the choice of m. Additionally, under the choice of small enough c_1: $\delta_{m,T} < c_4$ for all m and T, and thus, $C_{\texttt{blocking}}(T, m) \sim \mathcal{T}(m)T = \mathcal{O}(T)$.
2. On the other hand, if $\mathbb{P}$ is the law of the scaled Brownian motion, then the fastest growing contribution is that from the exponential term in (8) and in order to annul it, we should take $m = c_8 T$.

□

Acknowledgements We are grateful to two anonymous referees whose constructive comments have led to substantial improvements over a previous version of this paper.

Funding MM was supported as a doctoral student at the Department of Statistics, University of Warwick under Engineering and Physical Sciences Research Council (EPSRC) grant EP/L016710/1, and his work was concluded while being supported by the Max Planck Institute for Mathematics in the Sciences, Leipzig. PJ, MP, and GOR were supported by The Alan Turing Institute under the EPSRC grant EP/N510129/1. GOR was additionally supported under the EPSRC grants EP/K034154/1, EP/K014463/1, and EP/R018561/1.

Availability of Data and Material Code sufficient for reproducing the results in this article are available in the Github repository https://github.com/mmider/blocking.

Declarations

Conflicts of Interest The authors declare that they have no conflict of interest.

References

Amit Y (1991) On rates of convergence of stochastic relaxation for Gaussian and non-Gaussian distributions. J Multivar Anal 38(1):82–99

Arnaudon A, van der Meulen F, Schauer M, Sommer S (2020) Diffusion bridges for stochastic Hamiltonian systems with applications to shape analysis. arXiv preprint arXiv:200200885

Beskos A, Roberts GO (2005) Exact simulation of diffusions. Ann Appl Probab 15(4):2422–2444

Beskos A, Papaspiliopoulos O, Roberts GO (2006) Retrospective exact simulation of diffusion sample paths with applications. Bernoulli 12(6):1077–1098

Beskos A, Papaspiliopoulos O, Roberts GO (2008) A factorisation of diffusion measure and finite sample path constructions. Methodol Comput Appl Probab 10(1):85–104

Bladt M, Sørensen M et al (2014) Simple simulation of diffusion bridges with application to likelihood inference for diffusions. Bernoulli 20(2):645–675

Boys RJ, Wilkinson DJ, Kirkwood TBL (2008) Bayesian inference for a discretely observed stochastic kinetic model. Stat Comput 18(2):125–135

Chib S, Pitt MK, Shephard N (2004) Likelihood based inference for diffusion driven models
Dacunha-Castelle D, Florens-Zmirou D (1986) Estimation of the coefficients of a diffusion from discrete observations. Stochastics: an International Journal of Probability and Stochastic Processes 19(4):263–284
Delyon B, Hu Y (2006) Simulation of conditioned diffusion and application to parameter estimation. Stochastic Processes and their Applications 116(11):1660–1675
Durham GB, Gallant AR (2002) Numerical techniques for maximum likelihood estimation of continuous-time diffusion processes. J Bus Econ Stat 20(3):297–338
Freidlin MI, Wentzell AD (1993) Diffusion processes on graphs and the averaging principle. Ann Probab 2215–2245
Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, Rubin DB (2013) Bayesian data analysis. CRC Press
Golightly A, Wilkinson DJ (2008) Bayesian inference for nonlinear multivariate diffusion models observed with error. Comput Stat Data Anal 52(3):1674–1693
Hairer M, Stuart AM, Voss J et al (2011) Sampling conditioned hypoelliptic diffusions. Ann Appl Probab 21(2):669–698
Kalogeropoulos K (2007) Likelihood-based inference for a class of multivariate diffusions with unobserved paths. Journal of Statistical Planning and Inference 137(10):3092–3102
Kalogeropoulos K, Roberts GO, Dellaportas P (2010) Inference for stochastic volatility models using time change transformations. Ann Stat 38(2):784–807
Karatzas I, Shreve SE (1998) Methods of mathematical finance, vol 39. Springer
Kloeden PE, Platen E (2013) Numerical solution of stochastic differential equations, vol 23. Springer Science & Business Media
Kulkarni D, Schmidt D, Tsui SK (1999) Eigenvalues of tridiagonal pseudo-Toeplitz matrices. Linear Algebra Appl 297(1–3):63–80
Lansky P, Ditlevsen S (2008) A review of the methods for signal estimation in stochastic diffusion leaky integrate-and-fire neuronal models. Biol Cybern 99(4–5):253
Łatuszyński K, Kosmidis I, Papaspiliopoulos O, Roberts GO (2011) Simulating events of unknown probabilities via reverse time martingales. Random Struct Algoritm 38(4):441–452
Lauritzen SL (1996) Graphical models, vol 17. Clarendon Press
Levin DA, Peres Y (2017) Markov chains and mixing times, vol 107. American Mathematical Soc
Øksendal B (2003) Stochastic differential equations. Springer
Pitt MK, Shephard N (1999) Analytic convergence rates and parameterization issues for the Gibbs sampler applied to state space models. J Time Ser Anal 20(1):63–85
Pollock M, Johansen A, Roberts G (2016) On the exact and ϵ-strong simulation of (jump) diffusions. Bernoulli 22(2):794–856
Roberts GO, Rosenthal JS (2001) Markov chains and de-initializing processes. Scand J Stat 28(3):489–504
Roberts GO, Sahu SK (1997) Updating schemes, correlation structure, blocking and parameterization for the Gibbs sampler. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 59(2):291–317
Roberts GO, Stramer O (2001) On inference for partially observed nonlinear diffusion models using the metropolis-hastings algorithm. Biometrika 88(3):603–621
Rogers L (1985) Smooth transition densities for one-dimensional diffusions. Bull Lond Math Soc 17(2):157–161
Schauer M, Van Der Meulen F, Van Zanten H et al (2017) Guided proposals for simulating multi-dimensional diffusion bridges. Bernoulli 23(4A):2917–2950
Shephard N, Pitt MK (1997) Likelihood analysis of non-Gaussian measurement time series. Biometrika 84(3):653–667
Smith GD (1985) Numerical solution of partial differential equations: finite difference methods. Oxford University Press
Stramer O, Roberts GO (2007) On Bayesian analysis of nonlinear continuous-time autoregression models. J Time Ser Anal 28(5):744–762
van der Meulen F, Schauer M (2018) Bayesian estimation of incompletely observed diffusions. Stochastics 90(5):641–662