Efficiency of Delayed-Acceptance Random Walk Metropolis algorithms

Chris Sherlock, Alexandre H. Thiery and Andrew Golightly

July 4, 2019

Abstract

Delayed-acceptance Metropolis-Hastings and delayed-acceptance pseudo-marginal Metropolis-Hastings algorithms can be applied when it is computationally expensive to calculate the true posterior or an unbiased stochastic approximation thereof, but a computationally cheap deterministic approximation is available. An initial accept-reject stage uses the cheap approximation for computing the Metropolis-Hastings ratio; proposals which are accepted at this stage are then subjected to a further accept-reject step which corrects for the error in the approximation. Since the expensive posterior, or the approximation thereof, is only evaluated for proposals which are accepted at the first stage, the cost of the algorithm is reduced and larger scalings may be used.

We focus on the random walk Metropolis (RWM) and consider the delayed-acceptance RWM and the delayed-acceptance pseudo-marginal RWM. We provide a framework for incorporating relatively general deterministic approximations into the theoretical analysis of high-dimensional targets. Justified by diffusion approximation arguments, we derive expressions for the limiting efficiency and acceptance rates in high-dimensional settings. These theoretical insights are finally leveraged to formulate practical guidelines for the efficient tuning of the algorithms. The robustness of these guidelines and predicted properties are verified against simulation studies, all of which are strictly outside of the domain of validity of our limit results.

Keywords: Markov Chain Monte Carlo, delayed acceptance, pseudo-marginal MCMC, particle methods, diffusion limit.

Contents

1 Introduction

The Metropolis-Hastings algorithm is widely used to approximately compute expectations with respect to complicated high-dimensional posterior distributions [GRS96, BGJM11]. The algorithm requires that it be possible to evaluate point-wise the posterior density π up to a fixed but arbitrary constant of proportionality. In many cases each such evaluation can be computationally expensive, prompting the use of a surrogate model to accelerate the computations; for example, a Gaussian process-based approximation is used in [Ras03, FNL11].

The delayed-acceptance Metropolis-Hastings algorithm [CF05, MFS08, HRM+11, CFO11, BGLR15, SGH17, SL17], also called the surrogate transition method [Liu01], the modified Metropolis algorithm [AB01, CB18], preconditioned MCMC [EHL06] and two-stage MCMC [EDGG+05], assumes that the exact posterior π is available up to a constant of integration, but is computationally expensive to evaluate. This framework is particularly relevant to the Bayesian approach to inverse problems [KS06, Stu10] where point estimations of the posterior density typically involve numerically solving sets of partial differential equations. A fast approximation is therefore employed as a first “screening” stage, with proposals which are rejected at the screening stage simply discarded. The correct posterior, π, is only evaluated for proposals which pass the screening stage. A second Metropolis-Hastings accept-reject step, which corrects for the error in the fast approximation, is then calculated so that the desired true posterior is obtained as the limiting distribution of the Markov chain. The delayed-acceptance Metropolis-Hastings algorithm thus provides a principled method to leverage deterministic approximations to the posterior distribution in inverse problem modeling.

The pseudo-marginal Metropolis-Hastings algorithm [Bea03, AR09] allows Bayesian inference to be implemented when only an unbiased stochastic estimate of the target density, possibly up to an unknown normalisation constant, is available. The particle marginal Metropolis-Hastings algorithm [ADH10], a special instance of the pseudo-marginal Metropolis-Hastings algorithm when the unbiased estimate are obtained by using a particle filter, is a popular method for estimating parameters in hidden Markov models [e.g. GW11, KdV12].

The mixing efficiency of a pseudo-marginal algorithm increases with decreasing variability in the stochastic approximation [AR09, AV15]. However, decreasing the variability of the stochastic estimates of the target density typically comes at a computational price. This leads to a trade-off between mixing efficiency and computational expense and suggests that, for a given algorithm and target, there might be an optimal value for the stochasticity of the unbiased estimate, a tuning parameter often easily controlled. When particle filters or importance sampling procedures are used for constructing the unbiased estimate to the target distribution, this trade-off translates into choosing an optimal number of particles. The existing literature on this topic is reviewed in Section 2.2.

The computational expense involved in creating each unbiased stochastic estimate suggests that an initial accept-reject stage using a computationally cheap, deterministic, approximation [vK01, BC14] to the posterior might be beneficial. This motivates the delayed-acceptance pseudo-marginal Metropolis-Hastings algorithm [Smi11, GHS15, SGH17, QTVK18, ER17, VHF16].

Although the theoretical understanding of delayed-acceptance methods is still limited, several results are available. [SL17] compares the ergodicity properties of a delayed-acceptance algorithm with those of the parent MH algorithm, while [FV17] compares the asymptotic variance of the ergodic average from a delayed-acceptance algorithm with the variance of an importance-sampling estimator which takes as its proposal a sample from an MCMC targeting a surrogate. Historically, key insights into the performance and tuning of MCMC algorithms have been obtained by examining the limiting behaviour of a rescaled version of the Markov chain as the dimension of the statespace increases to infinity [RGG97, RR98, RR01, Béd07, BR08, STRR15, ZBK17, DLCMR17, YRR19]. In this article, we focus on random-walk proposals since this class of methods has the advantage of not requiring further information about the target, such as the local gradient or Hessian – if it is not possible to evaluate the posterior density, it is usually also impossible to evaluate these quantities and they are generally more computationally expensive to approximate than the target density itself [PDS11]. We thus concentrate on the delayed-acceptance random walk Metropolis (DARWM) and the delayed-acceptance pseudo-marginal random walk Metropolis (DAPsMRWM) algorithms: we obtain tuning and efficiency insights into these important algorithms through diffusion approximation arguments.

1.1 Contributions

When an accurate approximate posterior distribution is available, the use of well-tuned DARWM and DAPsMRWM algorithms can lead to large computational savings. Unfortunately, the tuning of these methods is delicate: it involves choosing an appropriate scale for the random walk proposals and, for the DAPsMRWM, a computational budget allocated to the creation of unbiased estimates of the posterior distribution. Tuning these parameters by estimating the Effective Sample Size (ESS) it typically impractical since the ESS is notoriously difficult and computationally expensive to estimate. These tuning difficulties have hindered the adoption of these powerful methods.

We examine the efficiency of the DARWM and DAPsMRWM algorithms when employed to explore high-dimensional posterior distributions. We express the efficiency of the methods as a function of the scaling of the random walk proposals and, for the DAPsMRWM, of the quality and computational cost of the unbiased estimates of the posterior distribution. One of our main innovations is to circumvent the difficulty of characterising the infinite variety of problem-specific errors in the cheap approximations to the posterior distribution by assuming that the error is a realisation of a random function – importantly, we empirically demonstrate that, in high-dimensional settings, this framework leads to robust conclusions that can be leveraged to develop efficient tuning guidelines. Under assumptions, we obtain MCMC diffusion limits through homogenization arguments. Despite the flexibility inherent to our specification of the deterministic approximation, the form of the limiting diffusion depends on the random function through just two key scalar properties. These limiting results are used to investigate the overall efficiency of these methods, taking computational time into account. For the DAPsMRWM algorithm, we focus on a specific standard asymptotic regime which occurs for instance when the unbiased stochastic estimates are obtained through a particle filter or when using a product of importance samplers for panel data.

We imagine that a practitioner has tuned a (pseudo-marginal) RWM algorithm, found it too inefficient, and implemented a delayed-acceptance (pseudo-marginal) RWM algorithm. Our analysis shows that the relative efficiency of the optimally tuned delayed-acceptance algorithm when compared to the optimally tuned parent algorithm, as well as the relative changes in the optimal random-walk scaling and computational budget allocated to the creation of unbiased estimates, can be characterised by two parameters **(1)** the relative computational cost of the cheap approximation compared to the cost of the posterior distribution, and **(2)** a measure of the accuracy of the cheap approximation involving the acceptance rate for proposals that have passed the first, screening stage. Crucially, these properties can be estimated easily and robustly, and thus used for tuning the DARWM and DAPsMRWM algorithms.

Three simulation studies verify different aspects of the theory and theoretical predictions. A pivotal result on the relationship between changes in the posterior and changes in the deterministic approximation is verified against a toy Bayesian inverse problem. The scaling and efficiency predictions for the DARWM are verified across a wide range of approximations in tractable settings. Finally, a real statistical example of pseudo-marginal inference on the parameters governing a Markov jump process is shown to fit with the predictions for the DAPsMRWM.

1.2 Organisation

The article proceeds as follows. Section 2 builds up descriptions of the DARWM and DAPsMRWM algorithms through their constituent algorithms, and provides a brief review of the literature on the efficiency of random-walk based algorithms. Section 3 describes the high-dimensional asymptotic regime studied in this article, sets up the models for the two approximations to the posterior and states the assumptions that are made on the posterior itself. In Section 4, we develop an asymptotic analysis of the DAPsMRWM, of which the DARWM is a special case; we formally introduce the expected squared jump distance and obtain asymptotic properties. A diffusion limit that gives theoretical justification for the optimization study presented in the subsequent section is then established. The asymptotic result are leveraged in Section 5 where we discuss the tuning of the DARWM and DAPsMRWM algorithms. The proofs and technical results are gathered in Section 6 and Appendix B. Section 7 provides practical advice and ratifies this against simulation studies. The article concludes with a discussion.

2 Delayed-acceptance Random Walks

Consider a posterior distribution $\pi(\mathrm{d}\mathbf{x})$ on a state-space $\mathcal{X} \subseteq \mathbb{R}^d$. We assume throughout this text that π possesses a density $\pi(\mathbf{x})$ with respect to the Lebesgue measure. The Random-Walk Metropolis-Hastings updating scheme provides a general class of algorithms for obtaining approximate samples from the distribution π by constructing a Markov chain that is reversible with respect to π. Given the current value $\mathbf{x} \in \mathcal{X}$ of the Markov chain, a new value $\mathbf{x}^*$ is proposed from a pre-specified symmetric proposal and accepted with probability $\alpha(\mathbf{x};\mathbf{x}^*) = 1 \wedge [\pi(\mathbf{x}^*)/\pi(\mathbf{x})]$. Upon acceptance, the proposal $\mathbf{x}^* \in \mathcal{X}$ becomes the next current value. Otherwise the current value is left unchanged.

2.1 Delayed-Acceptance strategies

As described in the introduction, there are many situations where π is computationally expensive to calculate while a computationally cheap approximation $\pi_a(\mathbf{x})$ to the density $\pi(\mathbf{x})$ is available and can be leveraged within MCMC schemes using the delayed-acceptance algorithm [Liu01, AB01, CF05, EHL06]. At the k-th iteration and given the current value $\mathbf{x}_k \in \mathcal{X}$ of the parameter, the DARWM generates a proposal $\mathbf{x}^* \in \mathcal{X}$ from a symmetric proposal kernel and proceeds as follows.

1. **Stage One**: compute the approximation $\pi(\mathbf{x}^*)$ and the screening acceptance probability

$$\alpha_1(\mathbf{x}_k;\mathbf{x}^*) \;=\; 1 \wedge \frac{\pi_a(\mathbf{x}^*)}{\pi_a(\mathbf{x}_k)}.$$

With probability $\alpha_1(\mathbf{x}_k; \mathbf{x}^*)$ proceed to Stage Two. Otherwise set $\mathbf{x}_{k+1} = \mathbf{x}_k$ and iterate.

2. **Stage Two**: compute the posterior distribution $\pi(\mathbf{x}^*)$ and the second stage probability

$$\alpha_2(\mathbf{x}_k; \mathbf{x}^*) \;=\; 1 \wedge \frac{\pi(\mathbf{x}^*)\, \pi_a(\mathbf{x}_k)}{\pi(\mathbf{x}_k)\, \pi_a(\mathbf{x}^*)}. \tag{2.1}$$

With probability $\alpha_2(\mathbf{x}_k; \mathbf{x}^*)$, set $\mathbf{x}_{k+1} = \mathbf{x}^*$. Otherwise, set $\mathbf{x}_{k+1} = \mathbf{x}_k$.

This defines a Markov chain that is reversible with respect to the posterior distribution π. Clearly, the more accurate the approximation π_a, the higher the *Stage Two* acceptance probability. The overall acceptance probability is

$$\alpha_{12}(\mathbf{x}_k; \mathbf{x}^*) \;=\; \alpha_1(\mathbf{x}_k; \mathbf{x}^*) \,\times\, \alpha_2(\mathbf{x}_k; \mathbf{x}^*).$$

Pseudo-marginal Metropolis-Hastings algorithms [Bea03, AR09] presume that it is computationally infeasible to evaluate the posterior density $\pi(\mathbf{x})$, even up to a multiplicative constant, but that it is possible to generate a positive and unbiased estimate $\widehat{\pi}(\mathbf{x}; \mathbf{u})$ of it. The quantity $\mathbf{u} \in \mathcal{U}$ represents a sample from a source of randomness necessary to produce the stochastic estimate $\widehat{\pi}(\mathbf{x}; \mathbf{u})$. Without loss of generality, one can assume that the auxiliary variables $\mathbf{u} \in \mathcal{U}$ is sampled from a fixed and known density $\rho(\mathbf{u})$. The pseudo-marginal version of the DARWM can be described as follows. At the k-th iteration, given the current value $\mathbf{x}_k \in \mathcal{X}$ of the parameter and its current stochastic estimate $\widehat{\pi}(\mathbf{x}_k, \mathbf{u}_k)$, the DAPsMRWM [Smi11, GHS15] generates a proposal $\mathbf{x}^* \in \mathcal{X}$ from a symmetric proposal kernel. The *stage one* screening procedure is identical to that of the DARWM. If this screening procedure is successful, a new auxiliary distribution $\mathbf{u}^*$ is generated from $\rho(u)$ and independently from all other sources of randomness to generate a stochastic estimate $\widehat{\pi}(\mathbf{x}^*; \mathbf{u}^*)$ to the (intractable) posterior distribution $\pi(\mathbf{x})$. The modified *stage two* acceptance probability reads

$$\alpha_2(\mathbf{x}_k, \mathbf{u}_k, \mathbf{x}^*, \mathbf{u}^*) = 1 \wedge \frac{\widehat{\pi}(\mathbf{x}^*, \mathbf{u}^*)\, \pi_a(\mathbf{x}_k)}{\widehat{\pi}(\mathbf{x}_k, \mathbf{u}_k)\, \pi_a(\mathbf{x}^*)}.$$

With probability $\alpha_2(\mathbf{x}_k, \mathbf{u}_k, \mathbf{x}^*, \mathbf{u}^*)$ one sets $(\mathbf{x}_{k+1}, \mathbf{u}_{k+1}) = (\mathbf{x}^*, \mathbf{u}^*)$. Otherwise, one sets $(\mathbf{x}_{k+1}, \mathbf{u}_{k+1}) = (\mathbf{x}_k, \mathbf{u}_k)$. Standard arguments show that the DAPsRWM is reversible with respect to the extended density $\pi(\mathbf{x})\, \rho(\mathbf{u})$ on $\mathcal{X} \times \mathcal{U}$. Particle marginal MCMC [ADH10], a special case of pseudo-marginal MCMC where the unbiased estimate of the posterior is obtained using a particle filter, has become the method of choice for Bayesian inference on hidden Markov models [e.g. FS11, GW11, DS19]

2.2 Tuning

The efficiency of a given RWM algorithm varies enormously with the scale of the proposed jumps[RR01, SFR10]. Small proposed jumps lead to high acceptance rates but little movement across the state-space, whereas large proposed jumps lead to low acceptance rates and again to inefficient exploration of the state space. The problem of choosing the optimal scale of the RWM proposal has been tackled for various shapes of target [RGG97, RR01, Béd07, BRS09, SR09, She13] and has led to the following rule of thumb: choose the scale so that the acceptance rate is approximately $\widehat{\alpha}_{rwm} \approx 23\%$. Although nearly all of the theoretical results are based upon limiting arguments in high dimension, the rule of thumb appears to be applicable even in relatively low dimensions [SFR10].

In discussing the literature on optimising pseudo-marginal algorithms it is helpful to define the Standard Asymptotic Regime (SAR), where the noise in the stochastic estimator of the log-posterior is additive, Gaussian with a variance that is independent of x and is inversely proportional to the computation effort required to produce the estimate. The Gaussianity and computational cost can be justified in the case of a particle filter, or a product of importance sampling estimators, by the asymptotic results in [BDMD13, SDDP18].

A relatively tractable lower bound on the efficiency of a pseudo-marginal Metropolis-Hastings algorithm is provided for an unrealistic special case in [PdSSGK12] and then extended considerably in [DPDK15]. Under the SAR, it is shown that the integrated autocorrelation time of the bounding chain is minimised when the variance of the noise in the estimated log-posterior is between 0.92^2 and 1.68^2. [STRR15] examine the behaviour of the pseudo-marginal random walk Metropolis algorithm under various regimes for the noise in the estimate of the posterior. Mixing efficiency is considered in terms of both limiting expected squared jump distance and the speed of a limiting diffusion, and an overall efficiency (ESJD/time) is defined, which takes into account the total computational time. Under the SAR, joint optimisation of this efficiency with respect to the variance of the noise in the log-target and the RWM scale parameter is considered. It is shown that the optimal scaling occurs when the acceptance rate is approximately $\widehat{\alpha}_{pm} \approx 7.0\%$ and the variance of the noise in the estimate of the log-posterior is approximately $\widehat{\sigma}^2_{pm} \approx 1.82^2$. [STRR15] also note that for the two different noise distributions considered in the article, the optimal scaling appears to be insensitive to the noise variance, and even to the distribution. This phenomenon is shown to hold across a large class of noise distributions in [She16].

This article extends [STRR15] to the corresponding delayed-acceptance algorithm, of which the DARWM is a special case. Results on limiting acceptance rates and mixing efficiency are proved, as is a diffusion limit. For the DARWM and for the DAPsMRWM under the Standard Asymptotic Regime (SAR), efficiency is then considered in detail.

3 High dimensional regime

In this section we introduce the high-dimensional asymptotic regime to be analysed in Sections 4, 5 and 6. In Section 3.1, the target distributions are described. In Section 3.2 and 3.3 respectively, we introduce the deterministic and stochastic approximation to the target distribution and the associated notations. We conclude in Section 3.4 with a careful description of the two-stage accept-reject mechanism.

3.1 Product form target distributions

We consider in this article target densities that have a simple product form. A research program along these lines was initiated in the pair of papers [RGG97, RR98]. Although only simple exchangeable product form targets were considered, a range of subsequent theoretical analyses confirmed that the results obtained in these articles also hold for more complex target distributions, such as products of one-dimensional distributions with different variances and elliptically symmetric distributions [RR01, BPS04, SR09, Béd07, SFR10]. Infinite-dimensional extensions were obtained in [MPS12, PST12, PST14]. We consider a target distribution $\pi^{(d)}(d\mathbf{x})$ in $\mathbb{R}^d$ with a density $\pi^{(d)}(\mathbf{x})$ with respect to the Lebesgue measure that can expressed as

$$\pi^{(d)}(\mathbf{x}) = \pi^{(d)}(x_1, \ldots, x_d) = \prod_{i=1}^{d} \pi(x_i) \tag{3.1}$$

for a one-dimensional density $\pi \equiv \pi^{(1)}$ on the real line. Throughout this article we assume that the Markov chain $\{(\mathbf{X}_k, U_k)\}_{k \geq 0}$ is stationary. For any algorithmic index $k \geq 0$, each component of $\mathbf{X}_k$ has distribution π. We consider Gaussian random walk proposals: for a current position $\mathbf{x} \in \mathbb{R}^d$, the proposal $\mathbf{X}^*$ is distributed as

$$\mathbf{X}^* = \mathbf{x} + \lambda^{(d)}\, \mathbf{Z}^{(d)} \qquad \text{with} \qquad \lambda^{(d)} = \left(\frac{\mu}{I}\right) d^{-1/2} \tag{3.2}$$

and a standard centred Gaussian random variable $\mathbf{Z}^{(d)}$ and a tuning parameter $\mu > 0$. The target dependent coefficient $I > 0$ is given by

$$I^2 = \mathbb{E}\left[\partial_x (\log \pi)(X)^2\right] = -\mathbb{E}\left[\partial_{xx} (\log \pi)(X)\right] \tag{3.3}$$

for a scalar random variable $X \overset{\mathcal{D}}{\sim} \pi$. The second equality in Equation (3.3) follows from an integration by parts that is justified, for example, by the regularity Assumption 4 described in Section 4. The constant $I > 0$ is introduced to simplify the statements of the results to follow. The scaling $d^{-1/2}$ ensures that, in the high-dimensional regime $d \to \infty$ the mean acceptance probability of a standard Random Walk Metropolis algorithm with proposals (3.2) and target distribution (3.1) stays bounded away from zero and one. Under mild assumptions, this scaling is optimal [RGG97, Béd07, BRS09, MPS12].

3.2 Deterministic approximation

To circumvent the difficulty of characterising the infinite variety of problem-specific errors in the cheap approximations π_a to the posterior distribution π, we choose model the discrepancy $s(\mathbf{x}) := \log\left(\pi_a(\mathbf{x})/\pi(\mathbf{x})\right)$ as the realisation of a random function. In our setting the target distribution is a d-dimensional product of one-dimensional distributions and we imagine that each of the terms in this product is approximated through an independent realisation of a random function. This means that the deterministic approximation $\pi_a^{(d)}(\mathbf{x}) = \pi^{(d)}(\mathbf{x}) \times \exp\left(s^{(d)}(\mathbf{x})\right)$ to the posterior density $\pi^{(d)}(\mathbf{x})$ possesses a deterministic error, on a logarithmic scale, of the form

$$s^{(d)}(\mathbf{x}) = \sum_{i=1}^{d} \mathcal{S}(x_i, \gamma_i), \tag{3.4}$$

where $\{\gamma_i\}_{i\geq 1}$ is the realisation of an i.i.d sequence of auxiliary random variables $\{\Gamma_i\}_{i\geq 1}$. Without loss of generality, we can assume that these auxiliary random variables are uniformly distributed on the interval $[0,1]$. We assume that the deterministic function $\mathcal{S} : \mathbb{R} \times [0,1] \to \mathbb{R}$ in Equation (3.4) satisfies the regularity Assumption 4 stated below. The following two properties of the function $\mathcal{S}$ directly influence the limiting efficiency of the delayed acceptance algorithm,

$$\beta_1 = \frac{\mathbb{E}\left[\,\partial_{xx}\mathcal{S}(X,\Gamma)\,\right]}{I^2} \in \mathbb{R} \qquad \text{and} \qquad \beta_2 = \left\{ \frac{\mathbb{E}\left[\,\partial_x\mathcal{S}(X,\Gamma)^2\,\right]}{I^2} \right\}^{1/2} \geq 0, \tag{3.5}$$

where expectation is taken over two independent random variables $\Gamma \overset{\mathcal{D}}{\sim} \text{Uniform}([0,1])$ and $X \overset{\mathcal{D}}{\sim} \pi$. An integration by parts and the Cauchy-Schwarz inequality yield that

$$\begin{aligned} I^2\,|\beta_1| = |\mathbb{E}\left[\partial_{xx}\mathcal{S}(X,\Gamma)\right]| \, &= \, |\mathbb{E}\left[\partial_x(\log\pi)(X)\,\partial_x\mathcal{S}(X,\Gamma)\right]| \\ &\leq \mathbb{E}\left[\partial_x(\log\pi)(X)^2\right]^{1/2} \times \mathbb{E}\left[\partial_x\mathcal{S}(X,\Gamma)^2\right]^{1/2} = I^2\beta_2. \end{aligned}$$

The quantity $-\beta_1$ may be interpreted as a measure of the excess curvature in the deterministic approximation, whereas β_2 is a measure of total discrepancy in the gradient. We thus have

$$-\beta_2 \leq \beta_1 \leq \beta_2. \tag{3.6}$$

It seems natural that a good approximation would match the curvature of the target, and indeed a matching of the curvature of an effectively-unimodal target at its mode is the basis of many importance samplers and independence samplers. However in many scenarios, such as the real statistical example considered in Section 7.3, the user has a single approximation and is not at liberty to choose the best from a whole family. Section 7.2 details a short simulation study in $d = 10$ where a Gaussian target is approximated by a logistic density and includes investigations of several different choices for the curvature with the mode fixed at the truth. Both the DAPsMRWM and the DARWM algorithms are considered. The study shows that whilst the best gain in efficiency is obtained when $\beta_1 \approx 0$ (and the mode is in the correct location), a substantial gain in efficiency can still be obtained even when the curvature of the approximation does not match that of the target. In all that follows we therefore consider the general case with $\beta_1 \neq 0$.

We conclude this section by a simple example that provides an intuitive basis for some of our theoretical results in Section 5. Consider a standard centred Gaussian target in $\mathbb{R}^d$ with inverse covariance matrix $\Sigma = \mathbf{Diag}(1/L, \ldots, 1/L)$ and an approximate distribution π_a whose i^{th} coordinate is distributed as $\mathbf{N}\left(a(\gamma_i), b(\gamma_i)^{-1}\right)$ for two arbitrary functions $a : [0,1] \to \mathbb{R}$ and $b : [0,1] \to \mathbb{R}_+$. Algebra shows that

$$\beta_1 = 1 - \mathbb{E}\left[\frac{b(\Gamma)}{L}\right] \qquad \text{and} \qquad \beta_2^2 = \mathbb{E}\left[\left\{1 - \frac{b(\Gamma)}{L}\right\}^2 + \frac{a(\Gamma)^2\, b^2(\Gamma)}{L}\right].$$

This confirms the heuristic that the quantity $-\beta_1$ measures the excess curvature in the deterministic approximation π_a, whereas β_2 is a measure of total discrepancy in the gradient.

3.3 Stochastic approximation

Following [PdSSGK12, STRR15, DPDK15], define $W = W(\mathbf{U};\mathbf{x}) \in \mathbb{R}$ and $W^* = W^*(\mathbf{U}^*;\mathbf{x}^*) \in \mathbb{R}$ implicitly through the equations

$$\widehat{\pi}(\mathbf{x}^*;\mathbf{u}^*) = \pi(\mathbf{x}^*)\, e^{W^*} \qquad \text{and} \qquad \widehat{\pi}(\mathbf{x};\mathbf{u}) = \pi(\mathbf{x})\, e^{W}.$$

The superscript $^{(d)}$ on all variables has been suppressed for simplicity of presentation. We sometimes write $\widehat{\pi}(\mathbf{x}^*;w^*)$ instead of $\pi(\mathbf{x}^*)\, e^{w^*}$ in order to stress the value of w^*. The random variables W^* and W, whose distributions depend on $\mathbf{x}^*$ and $\mathbf{x}$ respectively, are typically intractable and are only introduced to carry out the theoretical analysis of the DAPsMRWM algorithms. Let $\pi_{W^*}(w^* \mid \mathbf{x}^*)$ be the conditional density of $W^* \equiv W^*(\mathbf{U}^*;\mathbf{x}^*)$, deriving from $\rho(\mathbf{u}^*)$. The stationary density of $\widehat{\pi}(\mathbf{x};\mathbf{u})\rho(\mathbf{u})$ from Section 2 translates to a joint target of

$$\pi(\mathbf{x},w) \propto \pi(\mathbf{x})\, \pi_W(w \mid \mathbf{x}) \qquad \text{where} \qquad \pi_W(w \mid \mathbf{x}) = \pi_{W^*}(w \mid \mathbf{x})\, e^{w}. \tag{3.7}$$

The unbiasedness of the estimate $\widehat{\pi}(\mathbf{x}^*)$ yields that $\mathbb{E}\left[\exp\left(W^*\right) \mid \mathbf{x}^*\right] = 1$ for any $\mathbf{x}^* \in \mathcal{X}$. This shows that $\pi_W(w \mid \mathbf{x}) = \exp(w)\pi_{W^*}(w \mid \mathbf{x})$ is a valid conditional density. With this notation, we write the acceptance rate for the PsMMH as $\alpha_1\left(\mathbf{x};w;\mathbf{x}^*;w^*;\widehat{\pi},q\right)$, and the Stage Two acceptance rate for the DAPsMMH as $\alpha_2\left(\mathbf{x};w;\mathbf{x}^*;w^*;\widehat{\pi},\pi_a\right)$. For simplicity, and as in the articles [PdSSGK12, DPDK15, STRR15], we assume the following.

Assumptions 1. The additive noise, W^*, in the estimated log-target at the proposal, $\mathbf{X}^*$, is independent of the proposal value itself. We write π_{W^*} to denote its distribution.

An asymptotic argument justifying this assumption for panel data, where the unbiased estimate is obtained from a product of importance-sampling estimates, and hidden-Markov models, where it is obtained from a particle filter, using the posterior concentration as the number of observations increases to infinity is given in [SDDP18]. Assumption 1 means that, for any value of the proposal $\mathbf{X}^*$, the stochastic estimate of the target $\widehat{\pi}(\mathbf{X}^*)$ can expressed as $\pi(\mathbf{X}^*) \times e^{W^*}$ for a random variable $W^* \overset{\mathcal{D}}{\sim} \pi_{W^*}$ independent of any other source of randomness. From (3.7), in our d-dimensional setting, the process $\{(\mathbf{X},W)\}_{k\geq 0}$ is a Markov chain with invariant distribution $\pi^{(d)} \otimes \pi_W$. The distribution π_W does not vary with the dimension $d \geq 1$), where

$$\frac{d\pi_W}{d\pi_{W^*}}(w) = \exp(w). \tag{3.8}$$

This is Lemma 1 of [PdSSGK12]. In Section 5, we examine the behaviour of the algorithm under the following Gaussian assumption.

Assumptions 2. In addition to being independent of the proposal, $\mathbf{X}^*$, the additive noise in the estimated log-target at the proposal, W^*, is Gaussian:

$$W^* \overset{\mathcal{D}}{\sim} \mathbf{N}\left(-\sigma^2/2, \sigma^2\right). \tag{3.9}$$

In Equation (3.9) the mean is determined by the variance so as to give an unbiased estimate of the posterior, $\mathbb{E}\left[\exp\left(W^*\right)\right] = 1$. It follows from (3.8) that at stationarity, under Assumption 2, we have

$$W \overset{\mathcal{D}}{\sim} \mathbf{N}\left(\sigma^2/2, \sigma^2\right). \tag{3.10}$$

This article focuses on algorithms where the stochastic approximation to the likelihood is computationally expensive. In most scenarios of interest [GW11, KdV12, GHS15, FG14] the stochastic approximation is obtained through Monte-Carlo methods (e.g. importance sampling, particle filter) that converge at the standard $N^{-1/2}$ rate where N designates the number of samples/particles used. For taking into account the computational costs necessary to produce a stochastic estimate of the target-density, we thus assume the following in the rest of this article.

Assumptions 3. When Assumption 2 holds, the computational cost of obtaining an estimate of the log-target density with variance σ^2 is inversely proportional to σ^2.

The article [BDMD13] shows, among other things, that for state-space models (and panel data) the unbiased estimate of the likelihood obtained from standard particle methods [DM04] (or a product of importance sampling estimators) satisfies a log-normal central limit theorem, as the number of observations and particles (or importance samples) goes to infinity, if this number is of the same order as the number of noisy observations. This justifies the Gaussian approximation (3.9) and shows that the log-error is asymptotically inversely proportional to the number of particles used, justifying Assumptions 3. The article [STL17] studies the tuning of pseudo-marginal MCMC methods when the assumption 3 is not appropriate.

3.4 Acceptance probabilities

In this section we give formulae for the different acceptance probabilities when the DAPsMRWM algorithm is used for product-form targets as described in Section 3.1. When the current position of the algorithm is $(\mathbf{x}^{(d)}, w^{(d)}) \in \mathbb{R}^d \times \mathbb{R}$, a proposal $(\mathbf{x}^{(d),*}, w^{(d),*})$ distributed as

$$\mathbf{X}^{(d),*} = \mathbf{x}^{(d)} + \left(\frac{\mu}{I}\right) d^{-1/2}\, \mathbf{Z}^{(d)} \qquad \text{and} \qquad W^{(d),*} \overset{\mathcal{D}}{\sim} \pi_{W^*} \tag{3.11}$$

is generated. In this section and subsequently we will need to refer to three separate quantities and to distinguish for each which parts are fixed and which are random. We therefore define

$$\begin{cases} q_\Delta^{(d)}(\mathbf{x}^{(d)}, \mathbf{x}^{(d),*}) = \log\left[\pi^{(d)}(\mathbf{x}^{(d),*})/\pi^{(d)}(\mathbf{x}^{(d)})\right], \\ s_\Delta^{(d)}(\mathbf{x}^{(d)}, \mathbf{x}^{(d),*}) = s^{(d)}(\mathbf{x}^{(d),*}) - s^{(d)}(\mathbf{x}^{(d)}), \end{cases} \tag{3.12}$$

The deterministic approximation $\pi_a^{(d)}$ to the posterior density is used for a first screening procedure and the stochastic approximation is used for the second part of the accept-reject mechanism. The Stage One and overall acceptance probabilities read

$$\begin{cases} \alpha_1^{(d)}\left(\mathbf{x}^{(d)}, w^{(d)}; \mathbf{x}^{(d),*}, w^{(d),*}\right) & = F\left(q_\Delta^{(d)}(\mathbf{x}^{(d)}, \mathbf{x}^{(d),*}) + s_\Delta^{(d)}(\mathbf{x}^{(d)}, \mathbf{x}^{(d),*})\right) \\ \alpha_{12}^{(d)}\left(\mathbf{x}^{(d)}, w^{(d)}; \mathbf{x}^{(d),*}, w^{(d),*}\right) & = \alpha_1^{(d)}\left(\mathbf{x}^{(d)}, w^{(d)}; \mathbf{x}^{(d),*}, w^{(d),*}\right) \times F\left(w^{(d),*} - w^{(d)} - s_\Delta^{(d)}(\mathbf{x}^{(d)}, \mathbf{x}^{(d),*})\right). \end{cases}$$

where, for the Metropolis-Hastings accept-reject function $F(u) = 1 \wedge \exp(u)$. The proofs readily adapt, under mild regularity assumptions, to the case where $F : \mathbb{R} \to (0, 1]$ is a continuous and increasing function that satisfies the reversibility condition $e^{-u}\, F(u) = F(-u)$ for all $u \in \mathbb{R}$.

When the current position of the algorithm is $(\mathbf{x}^{(d)}, w^{(d)}) \in \mathcal{X} \times \mathcal{U}$, the first and second stage acceptance rate are defined by

$$\alpha_1^{(d)}\left(\mathbf{x}^{(d)}, w^{(d)}\right) = \mathbb{E}\left[\alpha_1^{(d)}\left(\mathbf{x}^{(d)}, w^{(d)}; \mathbf{X}^{(d),*}, W^{(d),*}\right)\right]$$
$$\alpha_{12}^{(d)}\left(\mathbf{x}^{(d)}, w^{(d)}\right) = \mathbb{E}\left[\alpha_{12}^{(d)}\left(\mathbf{x}^{(d)}, w^{(d)}; \mathbf{X}^{(d),*}, W^{(d),*}\right)\right]$$

for a proposal $(\mathbf{X}^{(d,*)}, W^{(d),*}) \in \mathcal{X} \times \mathcal{U}$ distributed as in (3.11). The conditional second stage acceptance rate is defined through Bayes rule as $\alpha_{2|1}^{(d)}\left(\mathbf{x}^{(d)}, w^{(d)}\right) = \alpha_{12}^{(d)}\left(\mathbf{x}^{(d)}, w^{(d)}\right)/\alpha_1^{(d)}\left(\mathbf{x}^{(d)}, w^{(d)}\right)$. We will eventually be interested in the acceptance rate at Stage One and the overall acceptance rate, which are defined as

$$\alpha_1^{(d)} = \mathbb{E}\left[\alpha_1^{(d)}\left(\mathbf{X}^{(d)}, W^{(d)}\right)\right] \qquad \text{and} \qquad \alpha_{12}^{(d)} = \mathbb{E}\left[\alpha_{12}^{(d)}\left(\mathbf{X}^{(d)}, W^{(d)}\right)\right]$$

with $(\mathbf{X}^{(d)}, W^{(d)}) \overset{\mathcal{D}}{\sim} \pi^{(d)} \otimes \pi_W$, as well as in the conditional Stage Two acceptance rate $\alpha_{2|1}^{(d)} = \alpha_{12}^{(d)}/\alpha_1^{(d)}$.

4 Asymptotic analysis

In this section we investigate the behaviour of the DAPsMRWM, and hence of the DARWM) as a special case, in the high-dimensional regime described in Section 3. We make the following regularity assumptions.

Assumptions 4. The density $\pi : \mathbb{R} \to (0, \infty)$ and the function $\mathcal{S} : \mathbb{R} \times [0, 1] \to \mathbb{R}$ satisfy the following.

1. The function $x \mapsto \log \pi$ is thrice differentiable, with second and third derivative bounded and the quantity $\mathbb{E}\left[(\partial_x \log \pi)^2(X)\right]$ is finite, for $X \overset{\mathcal{D}}{\sim} \pi$.
2. The first three derivatives with respect to the first argument of the function $(x, \gamma) \mapsto \mathcal{S}(x, \gamma)$ exist and are bounded over $(x, \gamma) \in \mathbb{R} \times [0, 1]$.

Assumptions 4 are repeatedly used for controlling the behaviour of second-order Taylor expansions; they could be relaxed in several directions at the costs of increasing technicality in the proofs. The following lemma is pivotal, and is proved in Section 6.1.

Lemma 4.1. *Let the regularity Assumptions 4 hold. Let $\{\gamma_i\}_{i \geq 1}$ be a realisation of the sequence of auxiliary random variable used to described the deterministic approximation (3.4) to the posterior density. Let $\{x_i\}_{i \geq 1}$ be the realisation of an i.i.d sequence marginally distributed as π. For $d \geq 1$, set $\mathbf{x}^{(d)} = (x_1, \ldots, x_d) \in \mathbb{R}^d$ and define the random variable*

$$\mathbf{X}^{(d),*} = \mathbf{x}^{(d)} + \left(\frac{\mu}{I}\right) d^{-1/2}\, \mathbf{Z}^{(d)} \qquad \textit{for} \qquad \mathbf{Z}^{(d)} \overset{\mathcal{D}}{\sim} \boldsymbol{N}(0, I_d)$$

For almost all realisations $\{x_i\}_{i \geq 1}$ and $\{\gamma_i\}_{i \geq 1}$ and $w \in \mathbb{R}$, the following limit

$$\lim_{d \to \infty} \begin{bmatrix} q_\Delta^{(d)}(\mathbf{x}^{(d)}, \mathbf{X}^{(d),*}) \\ s_\Delta^{(d)}(\mathbf{x}^{(d)}, \mathbf{X}^{(d),*}) \end{bmatrix} = \begin{bmatrix} Q_\Delta^\infty \\ S_\Delta^\infty \end{bmatrix} \overset{\mathcal{D}}{\sim} \boldsymbol{N}\left(-\frac{\mu^2}{2} \begin{bmatrix} 1 \\ -\beta_1 \end{bmatrix}, \mu^2 \begin{bmatrix} 1 & -\beta_1 \\ -\beta_1 & \beta_2^2 \end{bmatrix}.\right) \tag{4.1}$$

holds in distributions with parameters β_1 and β_2 defined in (3.5).

That the correlation is $-\beta_1/\beta_2 \in [-1, 1]$ is another manifestation of inequality (3.6). In the Gaussian example described at the end of Section 3.2 with $b(\Gamma) = b > L$ and $a(\Gamma) = 0$, and thus $\beta_1 < 0$, it is readily seen that $q_\Delta^{(d)}(\mathbf{x}^{(d)}, \mathbf{X}^{(d),*})$ and $s_\Delta^{(d)}(\mathbf{x}^{(d)}, \mathbf{X}^{(d),*})$ have the same sign and are positively correlated. In general, Lemma 4.1 shows that if the approximating density has an average excess of (negative) curvature (i.e. $\beta_1 < 0$), the limiting random variables Q_Δ^∞ and S_Δ^∞ are positively correlated.

4.1 Numerical confirmation of Lemma 4.1

Lemma 4.1 is pivotal to the high-dimensional asymptotic analysis to be described in subsequent sections. The product form Assumptions (3.1) and (3.4) from which we derive the bivariate Gaussian distribution in Lemma 4.1 are chosen for convenience. We expect the same conclusions to hold, at least approximately, in much broader settings; for example, we believe that extensions of Lemma 4.1 to non i.i.d target distributions similar to those discussed in [BPS04, Béd07, BR08, SR09, BRS09, PST12] are possible, at the cost of much less transparent proofs. In order to test the (approximate) validity of Lemma 4.1 in more realistic scenarios, and thus test the robustness of the results proved in this article, we consider a toy Bayesian inverse problem [Stu10] where none of the i.i.d assumptions are satisfied. We consider the problem of reconstructing an initial one-dimensional temperature field represented by a continuous function $T(\cdot, t = 0) : [0, 1] \to \mathbb{R}$ from N observations at time $t = \tau$ corrupted by independent Gaussian additive noise with known variance $\sigma^2_{\text{noise}} > 0$. In other words, we collect $\{y_i\}_{i=1}^N$ with $y_i \sim \mathbf{N}\left(T(x_i, t = T), \sigma^2_{\text{noise}}\right)$ for $1 \leq i \leq N$ at some location $x_i \in [0, 1]$. We assume that the evolution of the temperature field is described by the heat equation $\partial_t T = (1/2)\, \partial_{xx} T$ with Dirichlet boundary $T(x = 0, t) = T(x = 1, t) = 0$ for all time $t \in [0, \tau]$. We adopt a Gaussian process prior on the initial and unobserved temperature field and represent this prior as a finite Karhunen-Loève expansion

$$T(x, t = 0) = \sum_{k=1}^{K} \xi_k \, \sin(k \pi x),$$

for independent Gaussian random variables $\xi_k \sim \mathbf{N}(0, \kappa_k)$; the decay of the sequence $\kappa_k > 0$ controls the a-priori smoothness of the initial temperature field. We chose $\kappa_k = 1/k$ and $K = 40$ in our simulations. We have chosen this simple Bayesian inversion problem since a closed form solution for the heat equation is available; this allows a straightforward analysis of the approximation. Our approximate target is obtained through a coarse discretisation of the heat equation on $N_x = 50$ and $N_\tau = 10$ equidistant spatial and temporal points and using a standard fully-implicit finite-difference scheme [e.g. Tho13]. We implemented an exact RWM algorithm in the Fourier domain; i.e. the initial temperature field $T(\cdot, t = 0)$ is represented by its K-dimensional Karhunen-Loève expansion $\underline{c} = (c_1, \ldots, c_K)$. The prior log-density equals, up to an irrelevant additive constant, $-(1/2) \sum_{k=1}^{K} c_k^2/\kappa_k$ and the log-likelihood reads, up to a constant, $-(1/2\sigma_{\text{noise}}^2) \sum_{i=1}^{N} \{y_i - \mathcal{F}(\underline{c})(x_i)\}^2$ where $\mathcal{F}(\underline{c})(x) = \sum_{k=1}^{K} c_k \exp\left(-(k\,\pi)^2\,\tau/2\right) \sin(k\pi x)$. The variance of the RWM proposal was proportional to the prior variance matrix (which is not optimal, but reasonable in our example) with a scaling $\lambda > 0$ chosen so that roughly 25% of the proposals were accepted.

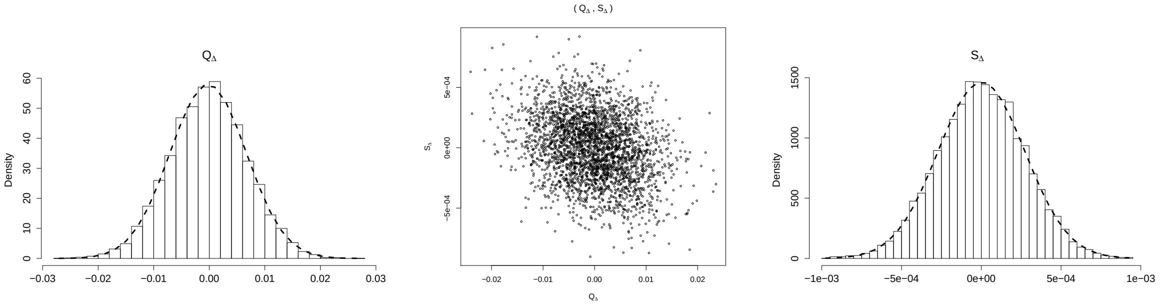

Figure 1: Empirical distribution of $q_\Delta^{(d)}(\mathbf{x}^{(d)}, \mathbf{X}^{(d),*})$ and $s_\Delta^{(d)}(\mathbf{x}^{(d)}, \mathbf{X}^{(d),*})$ evaluated from a current point in the bulk of the target distribution. The dashed lines in the left and right panels show the densities of Gaussian fits to the empirical marginal distributions of $q_\Delta^{(d)}(\mathbf{x}^{(d)}, \mathbf{X}^{(d),*})$ and $s_\Delta^{(d)}(\mathbf{x}^{(d)}, \mathbf{X}^{(d),*})$ respectively.

We focus on the aspects of Lemma 4.1 that are new: the properties of $s_\Delta^{(d)}(\mathbf{X}^{(d)}, \mathbf{X}^{(d),*})$ and its relationship with $q_\Delta^{(d)}(\mathbf{X}^{(d)}, \mathbf{X}^{(d),*})$. The marginal properties of $q_\Delta^{(d)}(\mathbf{X}^{(d)}, \mathbf{X}^{(d),*})$ have been known for some time [RGG97]. We ran 10^5 iterations of the exact RWM Markov chain in the Fourier domain and investigated numerically the distribution of the pair $q_\Delta^{(d)}(\mathbf{X}^{(d)}, \mathbf{x}^{(d),*})$ and $s_\Delta^{(d)}(\mathbf{x}^{(d)}, \mathbf{X}^{(d),*})$ at the final position of the RWM chain (in order to be in the main mass of the target distribution). The Gaussian behaviour of $q_\Delta^{(d)}(\mathbf{x}^{(d)}, \mathbf{X}^{(d),*})$ and $s_\Delta^{(d)}(\mathbf{x}^{(d)}, \mathbf{X}^{(d),*})$ is confirmed, as well as its non-trivial correlation structure (Fig. 1). Furthermore, we repeated the same experiment (results not presented here) at several other locations in the bulk of the target distribution and the distribution of $q_\Delta^{(d)}(\mathbf{x}^{(d)}, \mathbf{X}^{(d),*})$ and $s_\Delta^{(d)}(\mathbf{x}^{(d)}, \mathbf{X}^{(d),*})$ appears approximately independent of the location, as predicted by the theory.

To investigate the validity of Equation (4.1), we computed the quantities $\mathbb{E}[s_\Delta^{(d)}(\mathbf{X}^{(d)}, \mathbf{X}^{(d),*})]/\lambda^2$ and $\text{Var}\left[s_\Delta^{(d)}(\mathbf{X}^{(d)}, \mathbf{X}^{(d),*})\right]/\lambda^2$ and $\text{Corr}\left[q_\Delta^{(d)}(\mathbf{X}^{(d)}, \mathbf{X}^{(d),*}), s_\Delta^{(d)}(\mathbf{X}^{(d)}, \mathbf{X}^{(d),*})\right]$ for several choices of jump scaling $\lambda > 0$; Lemma 4.1 predicts that these quantities are independent of the scaling $\lambda > 0$, as is approximately numerically confirmed in Figure 2.

4.2 Limiting acceptance probability

The following lemma identifies the limiting acceptance rates as the dimension d goes to infinity.

Proposition 4.1. *Let Assumptions 1 and 4 hold. For almost every realisation $\{\gamma_i\}_{i\geq 1}$ of the sequence of auxiliary random variables used to describe the deterministic approximation (3.4) to the posterior density we have*

$$\lim_{d\to\infty} \mathbb{E}\left[\left|\alpha_1^{(d)}\left(\mathbf{X}^{(d)}, W^{(d)}\right) - \alpha_1\right|^2\right] = 0 \qquad \textit{and} \qquad \lim_{d\to\infty} \mathbb{E}\left[\left|\alpha_{12}^{(d)}(\mathbf{X}^{(d)}, W^{(d)}) - \alpha_{12}\right|^2\right] = 0 \tag{4.2}$$

where the limiting acceptance rates are given by

$$\alpha_1 = \mathbb{E}\left[F\left(Q_\Delta^\infty + S_\Delta^\infty\right)\right] \qquad \textit{and} \qquad \alpha_{12} = \mathbb{E}\left[F\left(Q_\Delta^\infty + S_\Delta^\infty\right) \times F\left(W_\Delta - S_\Delta^\infty\right)\right] \tag{4.3}$$

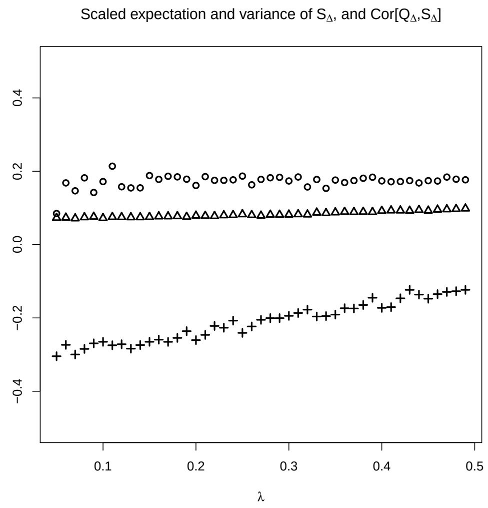

Figure 2: To investigate the validity of Equation (4.1), we computed the quantities $-\mathbb{E}[s_\Delta^{(d)}(\mathbf{X}^{(d)}, \mathbf{X}^{(d),*})_\Delta]/\lambda^2$ ($\circ$) and $\operatorname{Var}\left[s_\Delta^{(d)}(\mathbf{X}^{(d)}, \mathbf{X}^{(d),*})\right]/\lambda^2$ ($\triangle$) and $\operatorname{Corr}\left[q_\Delta^{(d)}(\mathbf{X}^{(d)}, \mathbf{X}^{(d),*}), s_\Delta^{(d)}(\mathbf{X}^{(d)}, \mathbf{X}^{(d),*})\right]$ ($+$) for several choices of jump scaling $\lambda > 0$; Lemma 4.1 predicts that these quantities are independent of the jump size $\lambda > 0$; this is approximately true in this Bayesian inverse problem toy example.

for $(Q_\Delta^\infty, S_\Delta^\infty)$ *as described in* (4.1) *and* $W_\Delta = W^* - W$ *for* $(W^*, W) \overset{\mathcal{D}}{\sim} \pi_{W^*} \otimes \pi_W$. *The dependence of* α_1 *and* α_{12} *upon* $(\mu, \beta_1, \beta_2, \pi_W)$ *is implicit.*

Corollary 4.1. *Under Assumptions 1 and 4 we have* $\lim_{d\to\infty} \alpha_1^{(d)} = \alpha_1$ *and* $\lim_{d\to\infty} \alpha_{12}^{(d)} = \alpha_{12}$.

Proof of Proposition 4.1 . We prove the first limit in Equation (4.2); the proof of the second limit is analogous. The first limit is equivalent to

$$\lim_{d\to\infty} \mathbb{E}\left[\left\{\mathbb{E}\left[F\left(q_\Delta^{(d)}(\mathbf{X}^{(d)}, \mathbf{X}^{(d),*}) + s_\Delta^{(d)}(\mathbf{X}^{(d)}, \mathbf{X}^{(d),*})\right) \,\middle|\, \mathbf{X}^{(d)}\right] - \mathbb{E}\left[F\left(Q_\Delta^\infty + S_\Delta^\infty\right)\right]\right\}^2\right] = 0.$$

This follows from the dominated convergence theorem, since the function F is bounded and continuous, and from the convergence in distribution proved in Lemma 4.1. □

For the remainder of our discussion of acceptance rates we adopt the Metropolis-Hastings acceptance probability, so $F(u) = 1 \wedge \exp(u)$, and we suppose that Assumption 2 holds: there is additive Gaussian noise in the logarithm of the stochastic approximation. We also make the dependence of the acceptance rate on the approximation parameters, β_1 and β_2, explicit. Standard computations (e.g. Proposition 2.4 of [RGG97]) yield that $\mathbb{E}\left[1 \wedge \exp(\mathbf{N}\left(a, b^2\right))\right] = \Phi(a/b) + \exp\left(a + b^2/2\right)\, \Phi(-b - a/b)$, with $\Phi : \mathbb{R} \to [0,1]$ the standard Gaussian cumulative distribution function. This permits straightforward evaluation of

$$\alpha_1(\mu; \beta_1, \beta_2) = \mathbb{E}\left[1 \wedge \exp(Q_\Delta^\infty + S_\Delta^\infty)\right], \tag{4.4}$$

$$\alpha_{12}(\mu, \sigma^2; \beta_1, \beta_2) = \mathbb{E}\left[\{1 \wedge \exp(Q_\Delta^\infty + S_\Delta^\infty)\}\{1 \wedge \exp(W_\Delta - S_\Delta^\infty)\}\right], \tag{4.5}$$

$$\alpha_{2|1}(\mu, \sigma^2; \beta_1, \beta_2) = \alpha_{12}(\mu, \sigma^2; \beta_1, \beta_2)/\alpha_1(\mu; \beta_1, \beta_2) \tag{4.6}$$

in terms of standard functions and (for α_{12}) a one-dimensional numerical integral, as detailed in Appendix A. The limit as $\beta_1 \to 0$ and $\beta_2 \to 0$ corresponds to the case when there is no deterministic error and leads to the usual [RGG97, MPS12] limiting acceptance rate of $2 \times \Phi(-\mu/2)$. For computing the limiting overall acceptance rate, note that under the Gaussian Assumption 2 we have $W_\Delta \overset{\mathcal{D}}{\sim} \mathbf{N}\left(-\sigma^2, 2\sigma^2\right)$.

The following result, whose proof is deferred to Appendix B.1, shows that it is possible to characterise the (unknown) values of μ and σ^2 in terms of the Stage One and the conditional Stage Two acceptance rates.

Proposition 4.2. *Let Assumptions 1, 2 and 4 hold and let the Metropolis-Hastings accept-reject function be used.*

1. *For any $\beta_2 > 0$ and $\beta_1 < 1$ the Stage One acceptance rate $\alpha_1(\mu; \beta_1, \beta_2)$ is a continuous decreasing bijection in μ from $[0, \infty)$ to $(0, 1]$.*
2. *For any fixed $\mu, \beta_2 > 0$ and β_1, the conditional Stage Two acceptance rate $\alpha_{2|1}(\mu, \sigma; \beta_1, \beta_2)$ is a decreasing bijection in σ from $[0, \infty)$ to $(0, \alpha_{2|1}(\mu, 0; \beta_1, \beta_2)]$.*

For the DARWM algorithm, we have that $W_\Delta \equiv 0$. Equation (4.5) yields

$$\alpha_{12}(\mu, 0; \beta_1, \beta_2) = \mathbb{E}\left[\{1 \wedge \exp(Q^\infty_\Delta + S^\infty_\Delta)\}\{1 \wedge \exp(-S^\infty_\Delta)\}\right],$$

which, as with DAPsMRWM, may be evaluated via a one-dimensional numerical integral. When $\beta_1 = \beta_2^2$, which necessitates $\beta_2 \leq 1$ by (3.6)), we have $\mathrm{Cov}\left[Q^\infty_\Delta + S^\infty_\Delta, -S^\infty_\Delta\right] = 0$ so that the random variables $Q^\infty_\Delta + S^\infty_\Delta$ and $W_\Delta - S^\infty_\Delta$ are independent and

$$\begin{aligned} \alpha_{2|1}(\mu, \sigma^2; \beta_2^2, \beta_2) &= \mathbb{E}\left[1 \wedge \exp\left\{\mathbf{N}\left(-\frac{1}{2}\beta_2^2\mu^2 - \sigma^2, \beta_2^2\mu^2 + 2\sigma^2\right)\right\}\right] \\ &= 2\Phi\left(-\frac{1}{2}\sqrt{\beta_2^2\mu^2 + 2\sigma^2}\right). \end{aligned} \tag{4.7}$$

This is the limiting acceptance probability of a pseudo-marginal RWM algorithm with a scaling of $\beta_2\,\mu$ and a noise variance of σ^2 [STRR15]. Substituting $\sigma^2 = 0$ into (4.7), we find that for the DARWM, $\alpha_{2|1}(\mu, 0; \beta_2^2, \beta_2) = 2\Phi(-\beta_2\mu/2)$, the limiting acceptance probability for a RWM algorithm with a scaling of $\beta_2\,\mu$ [RGG97]. In Section 5 the insights arising from this phenomenon help to motivate our approach to understanding the efficiency and tuning of DARWM and DAPsMRWM algorithms.

4.3 Limiting expected squared jumping distance

A standard measure of efficiency [SR09, BRS09, She13] for local algorithms is the Euclidian Expected Squared Jumping Distance (ESJD); see [RR14b, PG10] for detailed discussions. Theoretical motivations for our use of the ESJD are given by the diffusion approximation proved in Section 4.4. In our d-dimensional setting, it is defined as

$$\mathrm{ESJD}^{(d)} = \mathbb{E}\left[\|\mathbf{X}^{(d)}_{k+1} - \mathbf{X}^{(d)}_k\|^2\right]$$

where the Markov chain $\left\{(\mathbf{X}^{(d)}_k, W^{(d)}_k\right\}_{k\geq 0}$ is assumed to evolve at stationarity and $\|\cdot\|$ is the standard Euclidian norm.

Proposition 4.3. *Let Assumptions 1 and 4 hold. For almost every realisation $\{\gamma_i\}_{i\geq 1}$ of the sequence of auxiliary random variables used to describe the deterministic approximation (3.4) to the posterior density we have*

$$\lim_{d\to\infty} ESJD^{(d)} = \alpha_{12} \times \left(\frac{\mu}{I}\right)^2 \equiv J(\mu) \tag{4.8}$$

where α_{12} is the limiting acceptance rate identified in Proposition 4.1. The dependence of the limiting expected squared jumping distance $J(\mu)$ upon (β, π_W) is implicit.

4.4 Diffusion limit

We are motivated to prove that the DAPsMRWM algorithm in high dimensions can be well-approximated by an appropriate diffusion limit as this provides theoretical underpinning to our use of the ESJD as measure of efficiency [BDM12, RR14b]. The connection between ESJD and diffusions comes from the fact that the asymptotic jumping distance $\lim_{d\to\infty} \mathrm{ESJD}^{(d)} = J(\mu)$ is equal to the square of the limiting process's diffusion coefficient and is proportional to the drift coefficient. By a simple time change argument, the asymptotic variance of *any* Monte Carlo estimate of interest is inversely proportional to $J(\mu)$. Consequently, $J(\mu)$ becomes, at least in the limit, unambiguously the right quantity to optimise.

It is important to stress that the existence of the diffusion limit in this argument cannot be circumvented. MCMC algorithms which have non-diffusion limits can behave in very different ways and ESJD may not be a natural way to compare algorithms. The main result of this section is a diffusion limit for a rescaled version $V^{(d)}$ of the first coordinate process. For time $t \ge 0$ we define the piecewise constant continuous time process

$$V^{(d)}(t) := X^{(d)}_{\lfloor d\times t\rfloor,1} \, .$$

with the notation $\mathbf{X}^{(d)}_k = (X^{(d)}_{k,1}, \ldots, X^{(d)}_{k,d}) \in \mathbb{R}^d$. In general, the process $V^{(d)}$ is not Markovian; the next theorem shows nevertheless that in the limit $d \to \infty$ the process $V^{(d)}$ can be approximated by a Langevin diffusion.

Theorem 4.1. *Let Assumptions 1 and 4 hold. Let $T > 0$ be a finite time horizon and suppose that for all $d \ge 1$ the DAPsMRWM Markov chain starts at stationarity, $(\mathbf{X}^{(d)}_k, W^{(d)}) \overset{\mathcal{D}}{\sim} \pi^{(d)} \otimes \pi_W$. Then, as $d \to \infty$, the sequence of processes $V^{(d)}$ converges weakly to V in the Skorokhod topology on $D([0,T],\mathbb{R})$ where the diffusion process V satisfies the Langevin stochastic differential equation*

$$dV_t = \frac{1}{2}\, J(\mu)\, (\log \pi)'(V_t)\, dt + J^{1/2}(\mu)\, dB_t \tag{4.9}$$

with initial distribution $V_0 \overset{\mathcal{D}}{\sim} \pi$. The process B_t is a standard scalar Brownian motion.

Note that, as with Propositions 4.1 and 4.3, the Gaussian Assumption 2 is not necessary for the conclusion of Theorem 4.1 to hold. The proof can be found in Section 6.3. Theorem 4.1 shows that the rescaled first coordinate process converges to a Langevin diffusion V that is a time-change of the diffusion $d\overline{V}_t = \frac{1}{2}\,(\log \pi)'(\overline{V}_t)\, dt + dB_t$; indeed, $t \mapsto V_t$ has the same law as $t \mapsto \overline{V}_{J(\mu)\, t}$. This reveals that when speed of mixing is measured in terms of the number of iterations of the algorithm, the higher $J(\mu)$, the faster the mixing of the Markov chain. See [RR14a] for a detailed discussion and rigorous results. However any measure of overall efficiency should also take into account the computational time required for each iteration of the algorithm, and this is the subject of the next section.

5 Optimising the efficiency

When examining the efficiency of a standard RWM the computational time is usually either not taken into account or is implicitly supposed to be independent of the choice of tuning parameter(s). In any delayed-acceptance scenario, the computational time depends on the number of acceptances at Stage One; furthermore, in any pseudo-marginal setting the computational time also depends on the variance of the stochastic estimate of $\log \pi$. For this article, we measure the efficiency through a rescaled version of the expected squared jump distance,

$$(\text{Efficiency}) \equiv \frac{(\text{Expected Squared Jump Distance})}{(\text{Averaged one step computing time})}. \tag{5.1}$$

For any increasing function $\mathscr{F}$ the quantity $\mathscr{F}(\mathrm{ESJD})/(\text{Averaged one step computing time})$ is a valid measure of efficiency; the discussion at the start of Section 4.4 reveals nonetheless, because of the diffusion approximation proved in Theorem 4.1, that (5.1) is the essentially unique measure of efficiency valid in the high-dimensional asymptotic regime considered in this article. Proposition 4.3 shows that the limiting ESJD equals $\alpha_{12} \times (\mu/I)^2$ where I, defined in Equation (3.3), is a constant irrelevant for the optimisation of the efficiency discussed in this section; the constant also appears in the same form in the limiting ESJD for the equivalent non-delayed acceptance algorithm, and so it may also safely be ignored when calculating relative efficiencies. We examine the efficiency of the DARWM first, then move on to the DAPsMRWM.

5.1 Delayed-acceptance random walk Metropolis

For the DARWM we define an evaluation of π as taking one unit of time and define η to be the time for an evaluation of π_a: the one-step cost of a DARWM algorithm is $\eta + \alpha_1$. Following Equation (5.1) and eliminating unnecessary constants, the limiting efficiency of the DARWM can be quantified by the following efficiency functional:

$$\mathrm{Eff}_{\mathrm{da}}(\mu) = \frac{\mu^2\,\alpha_{12}(\mu,0)}{\eta + \alpha_1(\mu)}. \tag{5.2}$$

The dependence upon β_1 and β_2 is implicit. Using the same timescale, the efficiency of the RWM is $\mathrm{Eff}_{\mathrm{rwm}}(\mu) := 2\mu^2\Phi(-\mu/2)$ [RGG97], which is optimised at $\mu = \hat{\mu}_{rwm} \approx 2.38$. We may therefore define the relative efficiency of the DARWM algorithm compared with the optimal efficiency of the RWM algorithm

$$\mathrm{Eff}^{\mathrm{rel}}_{\mathrm{da}}(\mu) := \frac{\mathrm{Eff}(\mu)}{\mathrm{Eff}_{rwm}(\hat{\mu}_{rwm})}. \tag{5.3}$$

In the special case of $\beta_1 = \beta_2^2$, and as investigated in and around (4.7), we have that

$$\mathrm{Eff}^{\mathrm{rel}}_{\mathrm{da}}(\mu;\beta_2^2,\beta_2) = \frac{\mu^2}{\hat{\mu}^2_{rwm}}\,\frac{\alpha_1(\mu;\beta_1,\beta_2)\,\Phi(-\mu\beta_2/2)}{(\eta+\alpha_1(\mu;\beta_1,\beta_2))\,\Phi(-\hat{\mu}_{rwm}/2)}.$$

In the limit of an infinitesimal cost to evaluating π_a,*i.e.* $\eta = 0$, the efficiency is maximised at $\hat{\mu}_{da} = \hat{\mu}_{rwm}/\beta_2$, giving an overall relative efficiency of $\mathrm{Eff}^{\mathrm{rel}}_{\mathrm{da}}(\widehat{\mu}_{rwm}) = 1/\beta_2^2$. In reality, $\eta > 0$, and if μ is large enough so that $\alpha_1(\mu) \lesssim \eta$ then $\mu^2\alpha_{12}(\mu;\beta_1,\beta_2)$ will decrease rapidly with μ, as will the efficiency. This suggests that the quantity $\alpha_{2|1}(\widehat{\mu}_{rwm};\beta_1,\beta_2)$ might provide insight into the optimal scaling, $\widehat{\mu}_{da}$, and of the magnitude of $\mathrm{Eff}^{\mathrm{rel}}_{\mathrm{da}}(\widehat{\mu}_{da})$, provided that η is also taken into account. Figure 3 shows $\alpha_{2|1}(\widehat{\mu}_{rwm};\beta_1,\beta_2)$ and $\widehat{\mu}_{da}/\widehat{\mu}_{rwm}$ and $\mathrm{Eff}^{\mathrm{rel}}_{\mathrm{da}}(\widehat{\mu}_{da})$ as functions of β_1 and β_2 when $\eta = 0.01$. The shapes of the contours are almost identical, indicating that *whatever the values of* β_1 *and* β_2, the quantity $\alpha_{2|1}(\widehat{\mu}_{rwm};\beta_1,\beta_2)$ provides information on the optimal increase in scaling (relative to the optimal scaling for the RWM) and the corresponding increase in efficiency. Along the line where $\beta_1 = \beta_2^2$, as predicted, at $\beta_2 = 1$, $\widehat{\mu}_{da} \approx \widehat{\mu}_{rwm}/\beta_2$, but, since $\eta > 0$, as β_2 decreases the optimal scaling does not increase as quickly as this simple formula suggests.

Figure 4 plots $\widehat{\mu}_{da}/\widehat{\mu}_{rwm}$ (left) and $\mathrm{Eff}^{\mathrm{rel}}_{\mathrm{da}}(\widehat{\mu}_{da})$ (right) vs $\alpha_{2|1}(\widehat{\mu}_{rwm})$ over the fine grid of values of (β_1,β_2) used to create Figure 3. It shows that $\alpha_{2|1}(\widehat{\mu}_{rwm})$ combined with η does indeed provide information on the relative increase in scaling needed over $\widehat{\mu}_{rwm}$ and the resulting relative efficiency.

5.2 Delayed acceptance pseudo-marginal RWM

For the DAPsMRWM we define an evaluation of $\widehat{\pi}$ with $\sigma^2 = 1$ as taking one unit of time, and $\eta > 0$ is defined to be the time for an evaluation of π_a on this scale. Under Assumption 3, the average time needed to compute the stochastic approximation is inversely proportional to the variance, σ^2, of the estimate of the log-target, which leads to an average computational time for a single iteration of the algorithm of

$$\text{(Averaged one-step computing time)} \;=\; \eta + \alpha_1/\sigma^2.$$

As discussed in Section 3.3, Assumption 3 is reasonable when using particle MCMC to perform inference on the parameters of a hidden-Markov model, or when analysing panel data using a product of importance sampling estimators. We therefore simplify notation and refer the resulting efficiency as that of a Delayed-Acceptance Particle Marginal method. Our efficiency functional is, therefore,

$$\mathrm{Eff}_{\mathrm{dapm}}(\mu,\sigma^2) = \frac{\mu^2\,\sigma^2\,\alpha_{12}(\mu,\sigma)}{\eta\,\sigma^2 + \alpha_1(\mu)}. \tag{5.4}$$

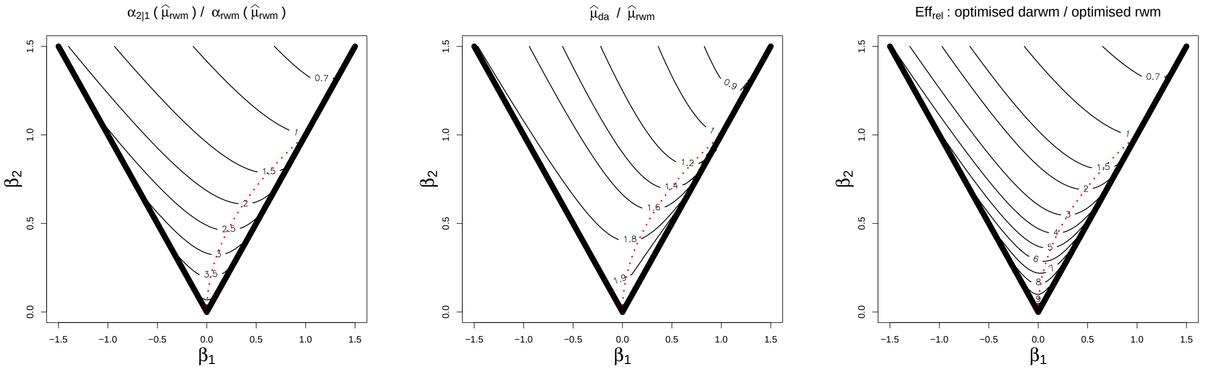

Figure 3: Contour plots of $\alpha_{2|1}(\widehat{\mu}_{rwm}, 0; \beta_1, \beta_2)$ (left), $\widehat{\mu}_{da}/\widehat{\mu}_{rwm}$ (centre) and $\text{Eff}^{\text{rel}}_{\text{da}}(\widehat{\mu}_{da})$ (right), as a function of β_1 and β_2 for $\eta = 0.01$. The red, dotted line satisfies $\beta_1 = \beta_2^2$.

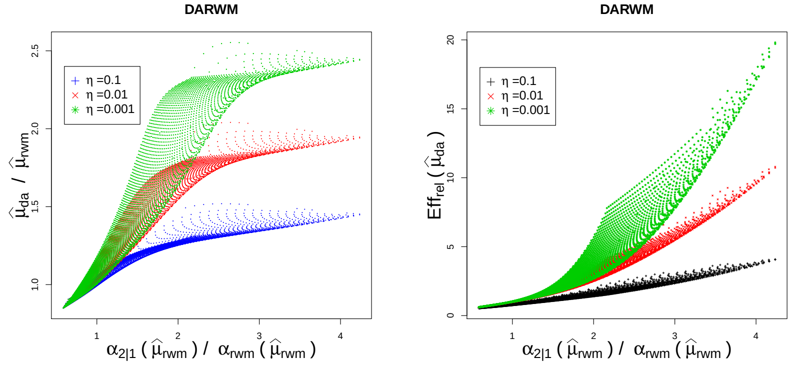

Figure 4: Scatter plots of $\widehat{\mu}_{da}/\widehat{\mu}_{rwm}$ (left) and $\text{Eff}^{\text{rel}}_{\text{da}}(\widehat{\mu}_{da})$ (right) vs $\alpha_{2|1}(\widehat{\mu}_{rwm})$, partitioned by η.

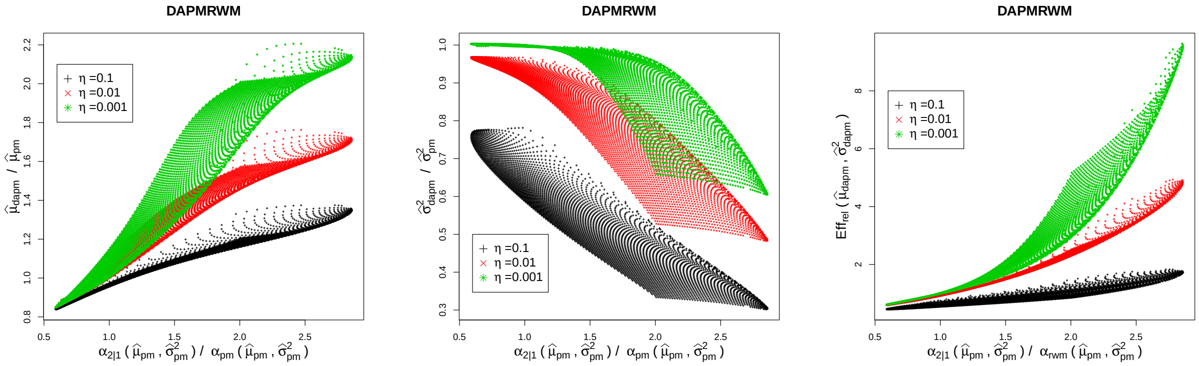

Figure 5: Scatter plots of $\widehat{\mu}_{dapm}/\widehat{\mu}_{pm}$ (left), $\widehat{\sigma}^2_{dapm}/\widehat{\sigma}^2_{pm}$ (centre) and $\mathrm{Eff}^{\mathrm{rel}}_{\mathrm{dapm}}(\widehat{\mu}_{dapm}, \widehat{\sigma}^2_{dapm})$, vs $\alpha_{2|1}(\widehat{\mu}_{pm}, \widehat{\sigma}^2_{pm})$ (right), partitioned by η.

Theorem 5.1, which is proved in Appendix B.2, shows that the efficiency functional $\mathrm{Eff}(\mu, \sigma^2)$ possesses intuitive limiting properties: too large or too small a jump size and/or stochastic variability in the estimation of the target is sub-optimal.

Theorem 5.1. *Let the regularity Assumption 4, the cost Assumption 3 and the Gaussian Assumption 2 hold. Suppose further that the Metropolis-Hastings accept-reject function has been used.*

1. *For a fixed variance $\sigma^2 > 0$ we have $\mathrm{Eff}(\mu, \sigma^2) \to 0$ as $\mu \to 0$ or $\mu \to \infty$.*
2. *For a fixed jump size $\mu > 0$ we have $\mathrm{Eff}(\mu, \sigma^2) \to 0$ as $\sigma^2 \to 0$ or $\sigma^2 \to \infty$.*

Using the same time scale as in (5.4), the equivalent efficiency function for the Particle-Marginal RWM is $\mathrm{Eff}_{pm}(\mu, \sigma^2) := 2\mu^2\sigma^2\Phi\left(-\frac{1}{2}\sqrt{\mu^2 + 2\sigma^2}\right)$, and this is maximised at $\hat{\mu}_{pm} \approx 2.562$ and $\hat{\sigma}^2_{pm} \approx 3.283$ [STRR15]. We may therefore define the relative efficiency of the DAPsMRWM algorithm compared with the maximum achievable efficiency of the Particle-Marginal RWM as follows,

$$\mathrm{Eff}^{\mathrm{rel}}_{\mathrm{dapm}}(\mu, \sigma^2) := \frac{\mathrm{Eff}_{\mathrm{dapm}}(\mu, \sigma^2)}{\mathrm{Eff}_{\mathrm{pm}}(\hat{\mu}_{pm}, \hat{\sigma}^2_{pm})}. \tag{5.5}$$

An argument analogous to the one used for analyzing the DARWM suggests that $\alpha_{2|1}(\widehat{\mu}_{pm}\widehat{\sigma}^2_{pm})$ and $\eta > 0$ together should be informative on $\widehat{\mu}_{dapm}$ and $\widehat{\sigma}^2_{dapm}$ and $\mathrm{Eff}^{\mathrm{rel}}_{\mathrm{dapm}}(\widehat{\mu}_{dapm}, \widehat{\sigma}^2_{dapm})$. Analogous contour plots to those in Figure 3, provided in Appendix A.1, show the same key property. Figure 5 shows scatter plots of $\widehat{\mu}_{dapm}$, $\widehat{\sigma}^2_{dapm}$ and $\mathrm{Eff}^{\mathrm{rel}}_{\mathrm{dapm}}$ against $\alpha_{2|1}(\widehat{\mu}_{pm}, \widehat{\sigma^2}_{pm})$ segregated by η. Again the combination of known quantities provides insight on the optimal relative tunings of the DA parameters compared with their non-DA optimal values. The efficiency plot also makes clear that it is not worth implementing a DAPsMRWM algorithm if π_a is only ten times faster to evaluate than $\widehat{\pi}$ is with $\sigma^2 = 1$.

As discussed in Section 2.2, an alternative tuning methodology relies on the property of the Particle-Marginal MRWM algorithm that the optimal μ for a given σ^2, $\widehat{\mu}(\sigma^2)$, is almost independent of σ^2 [STRR15, She16]. This effectively reduces a two-dimensional optimisation problem to two one-dimensional problems. Figure 6, which is typical of many other such figures that we produced, shows contour plots of $\mathrm{Eff}^{\mathrm{rel}}_{\mathrm{dapm}}$ as a function of μ and σ^2 for specific combinations of $\beta_2 \geq 0$, $|\beta_1| < \beta_2$ and $\eta > 0$. Each plot shows a single mode and also shows that for a particular variance, the optimal scaling $\hat{\mu}(\sigma)$ is insensitive to the value of σ, except when, approximately, $\sigma < 1$, at which point the optimal scaling increases. Provided the noise variance is not made too small, therefore, μ and σ may also be tuned independently for the DAPsMRWM.

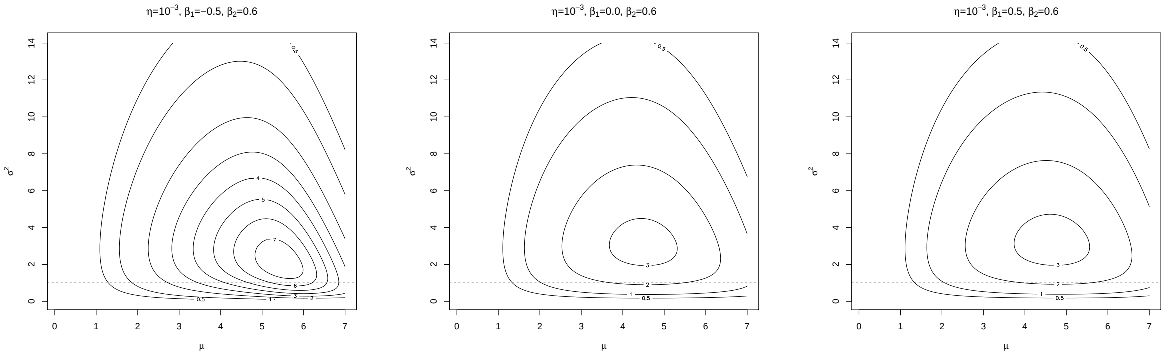

Figure 6: Contour plots of the asymptotic efficiency relative to the optimal efficiency of the equivalent pseudo-marginal RWM algorithm, $\mathrm{Eff}^{\mathrm{rel}}_{\mathrm{dapm}}$, as a function of the scaling, μ, and the variance of the noise in the log-target, σ^2, for different choices of $\beta_1,\ \beta_2$ at $\eta = 10^{-3}$. For comparability, all contours are at $0.5, 1, 2, 3, 4, 5, 6, 7, 8, 10, 12, 15$. The horizontal dashed line denotes $\sigma^2 = 1$.

6 Proofs

It will be helpful to introduce i.i.d sequences $\{X_i\}_{i\geq 1}$ and $\{\Gamma_i\}_{i\geq 1}$ respectively marginally distributed as π and π_Γ, and corresponding realisations of them, $\{x_i\}_{i\geq 1}$ and $\{\gamma_i\}_{i\geq 1}$. Similarly, we consider an i.i.d sequence $\{Z_{i,k}\}_{i,k\geq 0}$ of standard Gaussian $\mathbf{N}(0,1)$ random variables, $\{U_k\}_{k\geq 0}$ an i.i.d sequence of random variables uniformly distributed on $[0,1]$, W a random variable distributed as π_W and $\{W_k^*\}_{k\geq 0}$ an i.i.d sequence distributed as π_{W^*}. For any dimension $d \geq 1$ we set $\mathbf{X}_0^{(d)} = (X_1, \ldots, X_d) \in \mathbb{R}^d$ and $W_0^{(d)} = W$ and $X_{k,j}^{(d),*} = X_{k,j}^{(d)} + (\mu/I)\, d^{-1/2}\, Z_{k,j}$; we recursively define

$$(\mathbf{X}_{k+1}^{(d)}, W_{k+1}^{(d)}) = \begin{cases} (\mathbf{X}_k^{(d),*}, W_k^*) & \text{if} \quad U_k < \alpha^{(d)}\left(\mathbf{X}_k^{(d)}, W_k^{(d)}; \mathbf{X}_k^{(d),*}, W_k^*\right) \\ (\mathbf{X}_k^{(d)}, W_k^{(d)}) & \text{otherwise,} \end{cases}$$

for a proposal $\mathbf{X}_d^{(d),*} = (X_{k,1}^{(d),*}, \ldots, X_{k,d}^{(d),*})$. Indeed, the process $(\mathbf{X}_k^{(d)}, W_k^{(d)})$ is a DAPsMRWM Markov chain started at stationarity and targeting $\pi^{(d)} \otimes \pi_W$. We denote by $\mathcal{F}_k$ the σ-algebra generated by the family of random variables $\left\{\mathbf{X}_t^{(d)}, W_t^{(d)} \mid t \leq k\right\}$ and use the notation $\mathbb{E}_k[\,\cdot\,]$ for designating the conditional expectation $\mathbb{E}[\,\cdot \mid \mathcal{F}_k]$. Similarly, we use the notation $\mathbb{E}_{\mathbf{x},w}[\,\cdot\,]$ instead of $\mathbb{E}[\,\cdot \mid (\mathbf{X}_0^{(d)}, W_0^{(d)}) = (\mathbf{x}, w)]$. Finally, we set

$$\begin{array}{lll} q_\Delta^{(d)} = q_\Delta^{(d)}(\mathbf{x}^{(d)}, \mathbf{x}^{(d),*}), & s_\Delta^{(d)} = s_\Delta^{(d)}(\mathbf{x}^{(d)}, \mathbf{x}^{(d),*}), & w_\Delta^{(d)} = w^{(d),*} - w^{(d)}, \\ \mathsf{Q}_\Delta^{(d)} = q_\Delta^{(d)}(\mathbf{x}^{(d)}, \mathbf{X}^{(d),*}), & \mathsf{S}_\Delta^{(d)} = s(d)_\Delta(\mathbf{x}^{(d)}, \mathbf{X}^{(d),*}), & \mathsf{W}_\Delta^{(d)} = W^{(d),*} - w^{(d)}, \\ Q_\Delta^{(d)} = q(d)_\Delta(\mathbf{X}^{(d)}, \mathbf{X}^{(d),*}), & S_\Delta^{(d)} = s(d)_\Delta(\mathbf{X}^{(d)}, \mathbf{X}^{(d),*}), & W_\Delta^{(d)} = W^{(d),*} - W^{(d)}. \end{array} \tag{6.1}$$

and use the shorthand notation $\ell(x) \equiv \log \pi(x)$.

6.1 Proof of Lemma 4.1

The Law of Large Numbers and the separability of $L^1(\pi \otimes \pi_\Gamma)$ readily yield that for almost every realisations $\{x_i\}_{i\geq 1}$ and $\{\gamma_i\}_{i\geq 1}$, the following holds,

$$\lim_{n\to\infty} n^{-1} \sum_{i=1}^{n} \varphi(x_i, \gamma_i) = \int \varphi(x, \gamma)\ (\pi \otimes \pi_\Gamma)\,(dx, d\gamma) \qquad \text{for all} \quad \varphi \in L^1(\pi \otimes \pi_\Gamma). \tag{6.2}$$

We can thus safely assume in the remainder of this section that Equation (6.2) holds for the realisation $\{\gamma_i\}_{i\geq 1}$ of the auxiliary random variables used to describe the deterministic approximation (3.4) . By the

Cramer-Wold device, for proving Lemma 4.1 it suffices to establish that for any coefficient $c_Q, c_S \in \mathbb{R}$ the sequence $c_Q\, Q_\Delta^{(d)}(\mathbf{x}^{(d)}) + c_S\, S_\Delta^{(d)}(\mathbf{x}^{(d)})$ converges in law towards $c_Q\, Q_\Delta^\infty + c_S\, S_\Delta^\infty$; the boundedness assumption on the derivatives of the functions $x \mapsto \ell(x)$ and $x \mapsto \mathcal{S}(x,u)$ and a second order Taylor expansion show that this is equivalent to proving that the sum

$$\frac{\mu}{\sqrt{I^2\, d}} \sum_{i=1}^d \Big\{ c_Q\, \ell'(x_i) + c_S\, \partial_x \mathcal{S}(x_i, \gamma_i) \Big\}\, Z_i + \frac{1}{2}\, \frac{\mu^2}{I^2\, d} \sum_{i=1}^d \Big\{ c_Q\, \ell''(x_i) + c_S\, \partial_{xx}\mathcal{S}(x_i, \gamma_i) \Big\}$$

converges in law towards $c_Q\, Q_\Delta^\infty + c_S\, S_\Delta^\infty$. Definition (3.5) of the coefficient β_1 and β_2 yields that for almost every realisation $\{x_i\}_{i \geq 1}$ and $\{\gamma_i\}_{i \geq 1}$ we have

$$\frac{1}{I^2\, d} \sum_{i=1}^d \Big(\ell'(x_i)^2,\, \ell''(x_i),\, \partial_x \mathcal{S}(x_i,\gamma_i)^2,\, \partial_{xx}\mathcal{S}(x_i,\gamma_i),\, \ell'(x_i)\, \partial_x \mathcal{S}(x_i,\gamma_i) \Big) \;\to\; \Big(1,\, -1,\, \beta_2^2,\, \beta_1,\, -\beta_1\Big), \tag{6.3}$$

from which the conclusion directly follows since $c_Q\, Q_\Delta^\infty + c_S\, S_\Delta^\infty$ has a Gaussian distribution with mean $\mu^2\,(c_S \beta_1 - c_Q)/2$ and variance $\mu^2\,(c_Q^2 + c_S^2\, \beta_2^2 - 2\, c_Q\, c_S\, \beta_1)$.

6.2 Proof of Proposition 4.3

The quantity $\text{ESJD}^{(d)}$ can also be expressed as

$$\begin{aligned} \text{ESJD}^{(d)} &= \left(\frac{\mu^2}{I^2\, d}\right) \sum_{j=1}^d \mathbb{E}\left[\left(Z_j^{(d)}\right)^2 \times F\left(Q_\Delta^{(d)} + S_\Delta^{(d)}\right) \times F\left(W_\Delta^{(d)} - S_\Delta^{(d)}\right) \right] \\ &= \frac{\mu^2}{I^2}\, \mathbb{E}\left[\left(Z_1^{(d)}\right)^2 \times F\left(Q_\Delta^{(d)} + S_\Delta^{(d)}\right) \times F\left(W_\Delta^{(d)} - S_\Delta^{(d)}\right) \right] \end{aligned}$$

for $Q_\Delta^{(d)}$, $S_\Delta^{(d)}$, $W_\Delta^{(d)}$ defined in (6.1); the second equality follows from the exchangeability, at stationarity, of the d coordinates of the Markov chain. One can decompose $Q_\Delta^{(d)}$ and $S_\Delta^{(d)}$ as a sum of a term that is independent of $Z_1^{(d)}$ and a negligible term; we have $Q_\Delta^{(d)} = Q_\Delta^{(d),\perp} + \log\left[\pi(\mathbf{X}_1^{(d),*})/\pi(\mathbf{X}_1^{(d)})\right]$ and $S_\Delta^{(d)} = S_\Delta^{(d),\perp} + \mathcal{S}(\mathbf{X}_1^{(d),*}, \gamma_1) - \mathcal{S}(\mathbf{X}_1^{(d)}, \gamma_1)$ with

$$Q_\Delta^{(d),\perp} = \sum_{j=2}^d \log\left[\pi(\mathbf{X}_j^{(d),*})/\pi(\mathbf{X}_j^{(d)})\right] \qquad \text{and} \qquad S_{\Delta,\perp}^{(d)} = \sum_{j=2}^d \mathcal{S}(\mathbf{X}_j^{(d),*}, \gamma_j) - \mathcal{S}(\mathbf{X}_j^{(d)}, \gamma_j).$$

Note that $Q_\Delta^{(d),\perp}$ and $S_\Delta^{(d),\perp}$ are independent of $Z_1^{(d)}$. Under Assumption 4, the moments of order two of the differences $Q_\Delta^{(d)} - Q_\Delta^{(d),\perp}$ and $S_\Delta^{(d)} - S_\Delta^{(d),\perp}$ are finite and converges to zero as $d \to \infty$. The Cauchy-Schwarz inequality and the fact that F is bounded and Lipschitz yield that $\text{ESJD}^{(d)}/(\mu/I)^2$ can also be expressed as

$$\begin{aligned} &\mathbb{E}\Big[\left(Z_1^{(d)}\right)^2 \times F\left(Q_\Delta^{(d),\perp} + S_\Delta^{(d),\perp}\right) \times F\left(W_\Delta^{(d)} - S_\Delta^{(d),\perp}\right) \Big] \\ &\qquad + \mathbb{E}\left[\left(Z_1^{(d)}\right)^2 \times F\left(Q_\Delta^{(d),\perp} + S_\Delta^{(d),\perp}\right) \times \left\{ F\left(W_\Delta^{(d)} - S_\Delta^{(d)}\right) - F\left(W_\Delta^{(d)} - S_\Delta^{(d),\perp}\right) \right\} \right] \\ &\qquad + \mathbb{E}\left[\left(Z_1^{(d)}\right)^2 \times \left\{ F\left(Q_\Delta^{(d)} + S_\Delta^{(d)}\right) - F\left(Q_\Delta^{(d),\perp} + S_\Delta^{(d),\perp}\right) \right\} \times F\left(W_\Delta^{(d)} - S_\Delta^{(d)}\right) \right] \\ &= \mathbb{E}\left[F\left(Q_\Delta^{(d),\perp} + S_\Delta^{(d),\perp}\right) \times F\left(W_\Delta^{(d)} - S_\Delta^{(d),\perp}\right) \right] + o(1) \\ &= \mathbb{E}\left[F\left(Q_\Delta^\infty + S_\Delta^\infty\right) \times F\left(W_\Delta - S_\Delta^\infty\right) \right] + o(1) = \alpha_{12} + o(1), \end{aligned}$$

as required. We have used the fact that for almost every realisation of the auxiliary random variable $\{\Gamma_j\}_{j \geq 1}$ the sequence $\left(Q_\Delta^{(d),\perp}, S_\Delta^{(d),\perp}\right)$ converges in distribution to $(Q_\Delta^\infty, S_\Delta^\infty)$, which readily follows from Lemma 4.1.

6.3 Proof of Theorem 4.1

The proof is a generalisation of the generator approach of [RGG97, Béd07] coupled with an homogenization argument. We introduce the subsampled processes $\widetilde{\mathbf{X}}^{(d)}$ and $\widetilde{W}^{(d)}$ defined by

$$\widetilde{\mathbf{X}}^{(d)}_k = \mathbf{X}^{(d)}_{k \times T^{(d)}} \qquad \text{and} \qquad \widetilde{W}^{(d)}_k = W^{(d)}_{k \times T^{(d)}}$$

for an intermediary time scale defined as $T^{(d)} = \lfloor d^\gamma \rfloor$ where γ is an arbitrary exponent such that $\gamma \in (0, 1/4)$. One step of the process $\widetilde{\mathbf{X}}^{(d)}$ (resp. $\widetilde{W}^{(d)}$) corresponds to $T^{(d)}$ steps of the process $\mathbf{X}^{(d)}$ (resp. $W^{(d)}$). We then define an accelerated version $\widetilde{V}^{(d)}$ of the subsampled process $\widetilde{X}^{(d)}$. In order to prove a diffusion limit for the process $X^{(d)}$, one needs to accelerate time by a factor of d; consequently, in order to prove a diffusion limit for the process $\widetilde{X}^{(d)}$, one needs to accelerate time by a factor $d/T^{(d)}$ and thus define $\widetilde{V}^{(d)}$ by

$$\widetilde{V}^{(d)}(t) := \widetilde{X}^{(d)}_{\lfloor td/T^{(d)} \rfloor, 1}.$$

The proof then consists of showing that the sequence $\widetilde{V}^{(d)}$ converges weakly in the Skorohod topology towards the limiting diffusion (4.9) and verifying that $\|\widetilde{V}^{(d)} - V^{(d)}\|_{\infty, [0,T]}$ converges to zero in probability; this is enough to prove that the sequence $V^{(d)}$ converges weakly in the Skorohod topology towards the limiting diffusion (4.9). We denote by $\mathscr{L}$ the generator of the limiting diffusion (4.9). Similarly, we define $\mathscr{L}^{(d)}$ and $\widetilde{\mathscr{L}}^{(d)}$ the approximate generators of the first coordinate processes $X^{(d)}_1$ and $\widetilde{X}^{(d)}_1$; for any smooth and compactly supported test function $\varphi : \mathbb{R} \to \mathbb{R}$, vector $\mathbf{x} = (x_1, \ldots, x_d) \in \mathbb{R}^d$ and scalar $x, w \in \mathbb{R}$ we have

$$\begin{cases} \mathscr{L}^{(d)}\varphi(\mathbf{x}, w) & = \mathbb{E}_{\mathbf{x},w}[\varphi(X^{(d)}_{1,1}) - \varphi(X^{(d)}_{0,1})]/\delta \\ \widetilde{\mathscr{L}}^{(d)}\varphi(\mathbf{x}, w) & = \mathbb{E}_{\mathbf{x},w}[\varphi(\widetilde{\mathcal{X}}^{(d)}_{1,1}) - \varphi(\widetilde{\mathcal{X}}^{(d)}_{0,1})]/(T^{(d)} \times \delta) \qquad \text{for} \qquad \delta \equiv 1/d \\ \mathscr{L}\varphi(x) & = \frac{1}{2}\, J(\mu) \times (\ell'(x)\varphi'(x) + \varphi''(x))\,. \end{cases}$$

Note that although φ is a scalar function, the functions $\mathscr{L}^{(d)}\varphi$ and $\widetilde{\mathscr{L}}^{(d)}\varphi$ are defined on $\mathbb{R}^d \times \mathbb{R}$. The law of iterated conditional expectation yields the important identity between the generators $\mathscr{L}^{(d)}$ and $\widetilde{\mathscr{L}}^{(d)}$,

$$\widetilde{\mathscr{L}}^{(d)}\varphi(\mathbf{x}, w) = \frac{1}{T^{(d)}}\, \mathbb{E}_{\mathbf{x},w}\left[\sum_{k=0}^{T^{(d)}-1} \mathscr{L}\varphi\left(\mathbf{X}^{(d)}_k, W^{(d)}_k\right)\right]. \tag{6.4}$$

For clarity, the proof of Theorem 4.1 is divided into several steps.

6.3.1 The finite dimensional marginals of $\widetilde{V}^d$ converge to those of the diffusion (4.9)

Since the limiting process is a scalar diffusion, the set of smooth and compactly supported functions is a core for the generator of the limiting diffusion ([EK86],Theorem 2.1, Chapter 8); in the sequel, one can thus work with test functions belonging to this core only. Because the processes are started at stationarity, it suffices to show ([EK86],Chapter 4, Theorem 8.2, Corollary 8.4) that for any smooth and compactly supported function $\varphi : \mathbb{R} \to \mathbb{R}$ the following limit holds,

$$\lim_{d \to \infty} \mathbb{E}\left[\left|\widetilde{\mathscr{L}}^{(d)}\varphi(X_1, \ldots, X_d, W) - \mathscr{L}\varphi(X_1)\right|^2\right] = 0. \tag{6.5}$$

The proof of Equation (6.5) spans the remaining of this section and is based on an asymptotic expansion that we now describe. For every $x, w \in \mathbb{R}$ we define the approximated generator $\mathcal{A}\varphi : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ by

$$\mathcal{A}\varphi(x, w) = \left(\frac{\mu}{I}\right)^2 \left\{A(w)\,\ell'(x) + \left(\frac{1}{2}\alpha_{12} + [A(w) - B(w)]\,\partial_x \mathcal{S}(x, \gamma_1)\right)\, \varphi^{''}(x)\right\} \tag{6.6}$$

where $A, B : \mathbb{R} \to (0; \infty)$ are two bounded and continuous functions defined by

$$\begin{cases} A(w) & = \mathbb{E}\big[F'(Q^\infty_\Delta + S^\infty_\Delta) \times F(W^* - w - S^\infty_\Delta)\big] \\ B(w) & = \mathbb{E}\big[F(Q^\infty_\Delta + S^\infty_\Delta) \times F'(W^* - w - S^\infty_\Delta)\big] \end{cases} \tag{6.7}$$

for $W^* \overset{\mathcal{D}}{\sim} \pi_{W^*}$ and $F'(u) = e^u \, \mathbb{I}_{u<0}$ and $(Q_\Delta^\infty, S_\Delta^\infty)$ as defined in (4.1). The functions $A, B : \mathbb{R} \to \mathbb{R}_+$ are such that

$$\mathbb{E}\left[A(W)\right] = \mathbb{E}\left[B(W)\right] = \frac{1}{2}\,\alpha_{12}. \tag{6.8}$$

The proof of (6.8) can be found in Appendix B.3. It follows from (6.8) that for any fixed $x \in \mathbb{R}$ we have

$$\mathbb{E}\big[\mathcal{A}\varphi(x, W)\big] = \mathscr{L}\varphi(x) \tag{6.9}$$

for a random variable $W \overset{\mathcal{D}}{\sim} \pi_W$.

Lemma 6.1. *Let Assumptions 4 hold. We have*

$$\lim_{d\to\infty} \ \mathbb{E}\left[\left|\mathscr{L}^{(d)}\varphi(X_1, \ldots, X_d, W) - \mathcal{A}\varphi(X_1, W)\right|^2\right] \ = \ 0. \tag{6.10}$$

The proof of Lemma (6.1) consists in second order Taylor expansion and an averaging argument; details are in Section B.4. For proving Equation (6.5), note that identity (6.4) and Jensen's inequality yield the quantity inside the limit described in Equation (6.5) is less than two times the expectation of

$$\left\{\frac{\sum_{k=0}^{T^{(d)}} \mathscr{L}^{(d)}\varphi\left(\mathbf{X}_k^{(d)}, W_k^{(d)}\right) - \mathcal{A}\varphi\left(\mathbf{X}_{k,1}^{(d)}, W_k^{(d)}\right)}{T^{(d)}}\right\}^2 + \left\{\frac{\sum_{k=0}^{T^{(d)}} \mathcal{A}\varphi\left(\mathbf{X}_{k,1}^{(d)}, W_k^{(d)}\right) - \mathscr{L}\varphi\left(\mathbf{X}_{0,1}^{(d)}\right)}{T^{(d)}}\right\}^2.$$

The expectation of the first term is less than $\mathbb{E}\left[\left|\mathscr{L}^{(d)}\varphi(X_1, \ldots, X_d, W) - \mathcal{A}\varphi(X_1, W)\right|^2\right]$ and Lemma (6.1) shows that this quantity goes to zero as $d \to \infty$. To finish the proof it thus remains to verify that the expectation of the second term also converges to zero; to prove so, note that the second term is less than two times

$$\frac{\sum_{k=0}^{T^{(d)}} \left|\mathcal{A}\varphi(X_{k,1}^{(d)}, W_k^{(d)}) - \mathcal{A}\varphi(X_{0,1}^{(d)}, W_k^{(d)})\right|^2}{T^{(d)}} + \left\{\sum_{k=0}^{T^{(d)}} \mathcal{A}\varphi\left(X_{0,1}^{(d)}, W_k^{(d)}\right) - \mathscr{L}\varphi(\left(X_{0,1}^{(d)}\right) T^{(d)}\right\}^2. \tag{6.11}$$

Under the assumptions of Theorem 4.1, it is straightforward to verify that the function $\mathcal{A}\varphi$ is globally Lipschitz in the sense that there exists a constant $\|\mathcal{A}\varphi\|_{\text{Lip}}$ such that for every $x_1, x_2, w \in \mathbb{R}$ we have $|\mathcal{A}\varphi(x_1, w) - \mathcal{A}\varphi(x_2, w)| \leq \|\mathcal{A}\varphi\|_{\text{Lip}} \times |x_1 - x_2|$; it follows that the expectation of the first term in (6.11) converges to zero. For proving that the second term also converges to zero, we make use of the following ergodic averaging Lemma whose proof can be found in Section B.5.

Lemma 6.2. *Let $h : \mathbb{R} \to \mathbb{R}$ be a bounded and measurable test function. We have*

$$\lim_{d\to\infty} \ \mathbb{E}\left[\left|\frac{\sum_{k=0}^{T^{(d)}-1} h(W_k^{(d)})}{T^{(d)}} - \mathbb{E}\left[h(W)\right]\right|^2\right] = 0,$$

for a random variable $W \overset{\mathcal{D}}{\sim} \pi_W$ independent from any other sources of randomness.

Identity (6.9), a standard conditioning argument and Lemma 6.2 yield that the expectation of the second term in Equation (6.11) also converges to zero; this finishes the proof of the convergence of the finite dimensional marginals of $\widetilde{V}^d$ to those of the limiting diffusion (4.9).

6.3.2 The sequence $\widetilde{V}^d$ converges weakly towards the diffusion (4.9)

The finite dimensional marginals of the sequence process $\widetilde{V}^d$ converges to those of the diffusion (4.9). To prove that the sequence $\widetilde{V}^d$ actually converges to the diffusion (4.9), it thus suffices to verify that the sequence $\widetilde{V}^d$ is relatively weak compact in the Skorohod topology: since the process $\widetilde{V}^{(d)}$ is started at stationarity and the space of smooth functions with compact support is an algebra that strongly separates points, ([EK86], Chapter 4, Corollary 8.6) states that it suffices to show that for any smooth and compactly supported test function φ the sequence $d \mapsto \mathbb{E}\big|\widetilde{\mathscr{L}}^{(d)}\varphi(X_1, \ldots, X_d, W)\big|^2$ is bounded. Equation (6.5) shows that it suffices to verify that $\mathbb{E}\big|\mathscr{L}\varphi(X)\big|^2 < \infty$ for $X \overset{\mathcal{D}}{\sim} \pi$, which is obvious since φ is assumed to be smooth with compact support.

6.3.3 The sequence V^d converges weakly towards the diffusion (4.9)

Because the sequence $\widetilde{V}^d$ converges weakly to the diffusion (4.9), it suffices to prove that the difference $\|V^d - \widetilde{V}^d\|_{\infty,[0,T]}$ goes to zero in probability. To this end, it suffices to prove that the supremum

$$\sup\left\{ \left| X^{(d)}_{kT^{(d)}+i,1} - X^{(d)}_{kT^{(d)},1} \right| \;:\; k \times T^{(d)} \leq d \times T,\; i \leq T^{(d)} \right\}$$

converges to zero in probability. Since $\left| X^{(d)}_{kT^{(d)}+i,1} - X^{(d)}_{kT^{(d)},1} \right|$ is less than a constant times

$$\frac{1}{d^{1/2}}\left\{ \left|Z_{kT^{(d)},1}\right| + \ldots + \left|Z_{(k+1)T^{(d)}-1,1}\right| \right\},$$

standard Gaussian concentration gives the conclusion. This ends the proof of Theorem 4.1.

7 Practical advice and simulation studies

Theorem 5.1 suggests that our goal of finding the optimal scaling $\widehat{\mu}_{da} > 0$ or, for the DAPsMRWM, $\widehat{\mu}_{dapm} > 0$ and $\widehat{\sigma}^2_{dapm} > 0$), is sensible. We leverage our theory in section 7.1 to describe practical advice to this end. We conclude this section by empirically verifying these guidelines.

7.1 Practical advice

The values β_1, β_2 and I arise from an idealisation of the form of the target distribution, and the dependence of quantities of interest on these parameters arises from a limiting argument as $d \to \infty$. In reality, the quantities β_1 and β_2 and I might not exist. Even if they did exist, their values would not be known. We therefore base our practical advice on features that appear to be approximately independent of the specific values of β_1 and β_2, and for which I is irrelevant. Specifically, we focus on the quantites described in Figures 4, 5 and 6. Importantly, these quantities can straightforwardly and robustly be estimated from short MCMC trajectories.

DARWM: as described in Section 2.2, scaling analyses of the RWM led to the commonly used practical advice of tuning the scaling so that the acceptance rate is approximately $\widehat{\alpha}_{rwm}$. We assume that the practitioner has already found a scaling, $\widehat{\lambda}_{rwm}$ that is approximately optimal for the basic RWM, and noted $\alpha_{rwm}(\widehat{\lambda})$. Discovering that the efficiency is still too low, they have implemented a delayed-acceptance version of the algorithm which they now wish to tune.

Standard diagnostics give the relative computational cost, η, for the evaluations of π_a and π. The user should then run the DARWM algorithm with a scaling of $\widehat{\lambda}_{rwm}$, noting $\alpha_{2|1}(\widehat{\lambda}_{rwm})$. Figure 4 then gives the ratio $\widehat{\lambda}_{da}/\widehat{\lambda}_{rwm} = (\widehat{\mu}_{da}/I)/(\widehat{\mu}_{rwm}/I)$ that will be approximately optimal, as well as the estimated gain in efficiency.

DAPsMRWM: as described in Section 2.2 there are two possible tuning strategies based on the fact that the effect of altering the number of particles is approximately orthogonal to the effect of altering the scaling: *either* conditional on a given scaling, tune the number of particles to optimise efficiency, then with this number of particles, tune the scaling to optimise efficiency *or* change the number of particles to achieve the approximately optimal variance value, $\widehat{\sigma}^2_{pm}$, and tune the scaling to achieve an approximately optimal acceptance rate, $\widehat{\alpha}_{pm}$. For the first stage in the second option, the variance should be evaluated by running the PsMRWM algorithm with $\lambda = 0$ at some representative value $\mathbf{x}^{(0)}$, such as an approximate posterior mean, median or mode, as well as several other values, $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(k)}$, from the approximate posterior.

Figure 6 suggests that the first strategy should still apply directly to the DAPsMRWM, provided the variance is kept at $\sigma^2 \geq 1$. Alternatively, if the user already has an approximately optimal scaling and variance for the pseudo-marginal RWM, then, running the DAPsMRWM with these parameter values provides $\alpha_{2|1}(\widehat{\lambda}_{pm}, \widehat{\sigma}^2_{pm})$, and also η (if the relative CPU time is η^*, then $\eta \approx \eta^*/\widehat{\sigma}^2_{pm}$). Figure 5 may then be used to adjust λ and σ^2 for the DAPsMRWM algorithm.

To improve the efficiency of a (pseudo-marginal) RWM algorithms it is usual to make the jump proposal matrix reflect the overall shape of the posterior [RR01]. One frequently used strategy [SFR10] is to set the proposal covariance matrix to be proportional to an estimate of the target covariance matrix, $\widehat{V} := \widehat{\mathrm{Var}}(\mathbf{X})$, obtained from a preliminary run, for example whilst finding $\widehat{\lambda}_{rwm}$, or $\widehat{\lambda}_{pm}$ and $\widehat{\sigma}^2_{pm}$.

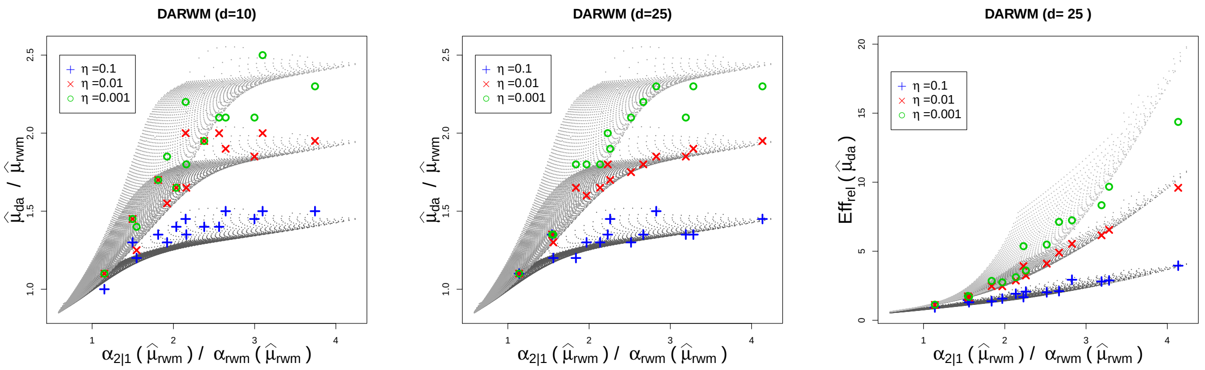

Figure 7: Scatter plots of $\widehat{\mu}_{da}/\widehat{\mu}_{rwm}$ ($d = 10$, left and $d = 25$, centre) and $\mathrm{Eff}^{\mathrm{rel}}_{\mathrm{da}}(\widehat{\mu}_{da})$ ($d = 25$, right), vs $\alpha_{2|1}(\widehat{\mu}_{rwm})$, partitioned by η.

7.2 Simulation study for the DARWM

We consider a scenario where the true target is a product of standard Gaussians and the deterministic approximation is a product of logistic densities with a mode at φ_1 and inverse-scale parameter φ_2,

$$\pi(x) \propto \exp\left\{-\frac{1}{2}\sum_{i=1}^{d} x_i^2\right\} \qquad \text{and} \qquad \hat{\pi}_a(x) \propto \prod_{i=1}^{d} \frac{e^{\varphi_2(x_i-\varphi_1)}}{\left(1+e^{\varphi_2(x_i-\varphi_1)}\right)^2}. \tag{7.1}$$

We consider fourteen scenarios: ten different combinations of values for (φ_1, φ_2), three approximations where the values of φ_1 or φ_2 vary from component to component, and the 'perfect approximation', $\pi_a = \pi$; see Appendix C for further details.

Empirical effective sample sizes (ESSs) for each of the d components are calculated using the `coda` package in `R` [PBCV06]; the overall ESS is taken to be the average of the ESSs over the d individual components. All algorithms were run for 10^6 iterations.

We first obtained the optimal scaling, $\widehat{\lambda}_{rwm}$, for a RWM targeting π by optimising the empirical ESS, and evaluated $\alpha_{rwm}(\widehat{\lambda}_{rwm})$ as well as the empirical ESS at this tuning. Then we ran the DA algorithm with this scaling to find $\alpha_{2|1}(\widehat{\lambda}_{rwm})$. Next, we artifically induced three different values of η: 0.1, 0.01, 0.001 and evaluated the efficiency, (empirical ESS-100) / CPU time) over a grid of possible scalings, λ, to find the optimal scaling. The regularisation penalty is needed because for very poorly mixing chains the empirical ESS tends to overestimate the true efficiency.

Figure 7 reproduces Figure 4 but in three shades of grey, then plots $\widehat{\lambda}_{da}/\widehat{\lambda}_{rwm}$ ($d = 10$ and $d = 25$) and the relative efficiency ($d = 25$) against $\alpha_{2|1}(\widehat{\lambda}_{da})/\alpha_{rwm}(\widehat{\lambda}_{rwm})$. At $d = 10$ the theory sometimes slightly overestimates the increase in scaling that is required, although (not shown) the predicted range of gains in efficiency is accurate except when η is small and $\alpha_{2|1}(\widehat{\mu}_{rwm})/\alpha_{rwm}(\widehat{\mu}_{rwm})$ is large, but by $d = 25$ the theoretical prediction of the ratio is quite accurate, as is the predicted efficiency gain. Essentially, with a larger scaling and a smaller dimension the diffusion approximation is less accurate.

7.3 Simulation study for the DAPsMRWM

To illustrate the advice for the DAPsMRWM, and provide a check on its validity, we consider a Lotka-Volterra predator-prey model [BWK08]. The model describes the continuous time evolution of $\mathbf{U}_t = (U_{1,t}, U_{2,t})$ where $U_{1,t}$ (prey) and $U_{2,t}$ (predator) are non-negative integer-values processes. Starting from an initial value, which is assumed known for simplicity, $\mathbf{U}_t$ evolves according to a Markov jump process (MJP) parameterised by rate constants $\mathbf{c} = (c_1, c_2, c_3)$ and characterised by transitions over $(t, t + dt]$ of the form

$$\begin{aligned}
\mathbb{P}\left(U_{1,t+dt} = u_{1,t} + 1, U_{2,t+dt} = u_{2,t} | u_{1,t}, u_{2,t}\right) &= c_1 u_{1,t} dt + o(dt), \\
\mathbb{P}\left(U_{1,t+dt} = u_{1,t} - 1, U_{2,t+dt} = u_{2,t} + 1 | u_{1,t}, u_{2,t}\right) &= c_2 u_{1,t} u_{2,t} dt + o(dt), \\
\mathbb{P}\left(U_{1,t+dt} = u_{1,t}, U_{2,t+dt} = u_{2,t} - 1 | u_{1,t}, u_{2,t}\right) &= c_3 u_{2,t} dt + o(dt).
\end{aligned}$$

The process is easily simulated via the Gillespie algorithm [Gil77] and the pseudo-marginal RWM scheme is straightforward to apply [GW11]. We assume that the MJP is observed with Gaussian error every time unit for n time units, $t = 1, \dots, n$:

$$\mathbf{Y}_t \overset{\mathcal{D}}{\sim} \mathbf{N}\left(\begin{bmatrix} u_{1,t} \\ u_{2,t} \end{bmatrix}, \begin{bmatrix} s_1^2 & 0 \\ 0 & s_2^2 \end{bmatrix} \right).$$

As all of the parameters of interest must be strictly positive, we consider inference for

$$\mathbf{x} = (\log(c_1), \log(c_2), \log(c_3), \log(s_1), \log(s_2)).$$

The DAPsMRWM scheme requires that a computationally cheap approximation of the MJP is available. We follow [GHS15] by constructing a linear noise approximation (LNA) (see e.g. [vK01]). Under the LNA

$$\mathbf{U}_t \overset{\mathcal{D}}{\sim} \mathbf{N}\left(\mathbf{z}_t + \mathbf{m}_t\, ,\, \mathbf{V}_t\right)$$

where $\mathbf{z}_t$, $\mathbf{m}_t$ and $\mathbf{V}_t$ satisfy a coupled ODE system

$$\left\{ \begin{array}{lcl} \dot{\mathbf{z}}_t &=& \mathbf{S}\,\mathbf{h}(\mathbf{z}_t, \mathbf{c}) \\ \dot{\mathbf{m}}_t &=& \mathbf{F}_t \mathbf{m}_t \\ \dot{\mathbf{V}}_t &=& \mathbf{V}_t \mathbf{F}_t^T + \mathbf{S}\mathrm{diag}\left\{\mathbf{h}(\mathbf{z}_t, \mathbf{c})\right\} \mathbf{S}^T + \mathbf{F}_t \mathbf{V}_t \end{array} \right. \tag{7.2}$$

For the Lotka-Volterra model, the rate vector $\mathbf{h}(\mathbf{z}_t, \mathbf{c})$, stoichiometry matrix $\mathbf{S}$ and Jacobian matrix $\mathbf{F}_t$ are given by

$$\mathbf{h}(\mathbf{z}_t, \mathbf{c}) = (c_1 z_{1,t}, c_2 z_{1,t} z_{2,t}, c_3 z_{2,t}),$$

$$\mathbf{S} = \begin{pmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \end{pmatrix}, \qquad \mathbf{F}_t = \begin{pmatrix} c_1 - c_2 z_{2,t} & -c_2 z_{1,t} \\ c_2 z_{2,t} & c_2 z_{1,t} - c_3 \end{pmatrix}.$$

Appendix D describes an algorithm for evaluating the posterior (up to proportionality) under the LNA. For further details regarding the LNA and its use as an approximation to a MJP, we refer the reader to [FGS14] and [GHS15]. Data were simulated using an initial value $\mathbf{u}_0 = (71, 79)$ for $n = 50$ time units with $\mathbf{c} = (1.0, 0.005, 0.6)$ and $s_1 = s_2 = 8$. These parameters were assumed to be independent *a priori* with independent proper Uniform densities on the interval $[-8, 8]$ ascribed to X_i, $(i = 1, \dots, 5)$. For a pseudo-marginal RWM scheme [STRR15] suggests that for a Gaussian target (where, for each principal component, I is known) the scaling should be $\mathbf{V}_{\text{Gauss}} = \left(2.56^2/d\right) \times \mathrm{Var}(\mathbf{X})$ to optimise efficiency. We refer to the scaling relative to this proposal as γ; i.e. we propose Gaussian jumps with a variance of $\mathbf{V}_{\text{prop}} = \gamma^2 \widehat{\mathbf{V}}_{\text{Gauss}}$, where $\mathrm{Var}(\mathbf{X})$, has been replaced with an approximation, $\widehat{\mathrm{Var}}(\mathbf{X})$, created from an initial run. In this example we found that the pseudo-marginal RWM was optimised at $\gamma \approx 1.2$. [STRR15] suggests that the optimal number of particles should lead to a variance in $\log \widehat{\pi}$ of approximately 3.3. We found that the optimal number of particles was $m = 180$, which occurred when the $\mathrm{Var}[\log \widehat{\pi}(x_*)]$ (with x_* an initial estimate of the componentwise posterior median) was approximately 2.9. The mean acceptance probability at this optimal tuning was $\alpha_{pm} \approx 8.0\%$ and the empirical efficiency, measured in terms of minimum (over each parameter component) effective sample size per second, was 0.067.

The DAPsMRWM with $\gamma = 1.2$ and $m = 180$ gave $\alpha_{2|1} \approx 20.7\%$, so that $\alpha_{2|1}/\alpha_{pm} \approx 2.6$; timing diagnostics gave $\eta = 0.0014$. For this combination, Figure 5 suggests increasing the scaling by a factor of around 2.1, decreasing the variance by a factor of between 0.7 and 0.8, and that this should lead to an increase in efficiency of a factor of between 6 and 7. The tuning suggestions translate to $\gamma \approx 2.5$ and $m \approx 225 - 255$. Alternatively, Figure 6 suggests that provided $\sigma^2 > 1$, m and γ may be tuned independently.

To confirm that the practical advice is reasonable and to test some of the other predictions of our theory, the number of particles m was varied between 80 and 2000 and, for each m, the scaling γ was varied

	m	80	100	150	200	250	300	500	800	2000
γ	σ^2	8.30	5.86	3.53	2.52	1.83	1.50	0.89	0.52	0.20
1	mESS/s	0.0750	0.0808	0.0810	0.108	0.118	0.119	0.119	0.113	0.0661
	$\hat{\alpha}_1$	0.256	0.255	0.257	0.255	0.257	0.254	0.254	0.255	0.258
	$\hat{\alpha}_{2\|1}$	0.0651	0.0883	0.170	0.237	0.289	0.341	0.447	0.547	0.692
2	mESS/s	0.140	0.218	0.296	0.289	0.319	0.278	0.262	0.181	0.127
	$\hat{\alpha}_1$	0.0556	0.0514	0.0489	0.0503	0.0517	0.0520	0.0513	0.0517	0.0505
	$\hat{\alpha}_{2\|1}$	0.0619	0.0895	0.163	0.213	0.286	0.313	0.438	0.522	0.674
2.5	mESS/s	0.142	0.226	0.338	0.381	0.325	0.318	0.330	0.282	0.142
	$\hat{\alpha}_1$	0.0244	0.0237	0.0234	0.0259	0.0264	0.0234	0.0241	0.0230	0.0250
	$\hat{\alpha}_{2\|1}$	0.0600	0.0815	0.159	0.218	0.252	0.312	0.434	0.523	0.675
3	mESS/s	0.160	0.294	0.364	0.441	0.401	0.419	0.364	0.277	0.156
	$\hat{\alpha}_1$	0.0143	0.0123	0.0119	0.0114	0.0131	0.0120	0.0114	0.0124	0.0121
	$\hat{\alpha}_{2\|1}$	0.0416	0.101	0.152	0.233	0.274	0.320	0.426	0.516	0.673
3.5	mESS/s	0.107	0.225	0.331	0.402	0.374	0.390	0.348	0.307	0.162
	$\hat{\alpha}_1$	0.00629	0.00789	0.00763	0.00684	0.00669	0.00663	0.00725	0.00634	0.00694
	$\hat{\alpha}_{2\|1}$	0.0550	0.0869	0.170	0.237	0.273	0.312	0.424	0.534	0.673
4	mESS/s	0.107	0.174	0.176	0.291	0.308	0.319	0.351	0.292	0.162
	$\hat{\alpha}_1$	0.00343	0.00318	0.00401	0.00388	0.00372	0.00357	0.00377	0.00402	0.00418
	$\hat{\alpha}_{2\|1}$	0.0680	0.105	0.151	0.215	0.287	0.310	0.407	0.500	0.681
4.5	mESS/s	0.0728	0.159	0.150	0.267	0.310	0.300	0.300	0.258	0.153
	$\hat{\alpha}_1$	0.00220	0.00183	0.00207	0.00247	0.00230	0.00256	0.00224	0.00249	0.00226
	$\hat{\alpha}_{2\|1}$	0.0527	0.111	0.143	0.213	0.265	0.280	0.424	0.491	0.658

Table 1: Minimum effective sample size (mESS) per second, stage 1 acceptance probability $\hat{\alpha}_1$ and stage 2 acceptance probability $\hat{\alpha}_{2|1}$ as functions of the number of particles m and scaling γ. The variance (σ^2) of the estimated log-posterior at the median is also shown for each choice of m.

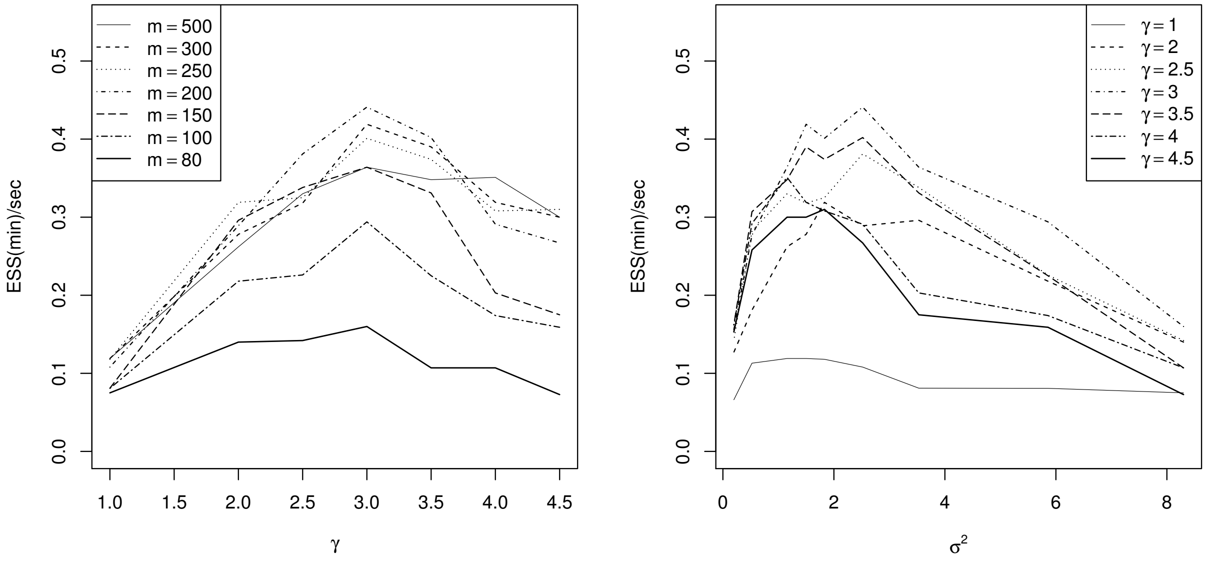

Figure 8: Empirical efficiency measured as the effective sample size per CPU second. The left-hand panel gives the efficiency plotted against γ for various numbers of particles. The right-hand panel gives the efficiency plotted against σ (estimated at the posterior median), for various scalings.

between 1 and 4.5. For each (m, γ) pair, a long MCMC run (of at least 4×10^5 iterations) was performed. Figure 8 shows empirical efficiency as a function of the scaling γ (with a varying number of particles m) and as a function of the number of particles (for various scalings γ) and provides empirical evidence of the insensitivity of the optimal choice of scaling, γ, to the value of σ^2, for values of $\sigma^2 >= 0.89$; furthermore, for variances below 0.89 the optimal scaling increases, as predicted by our theory. Table 1 shows empirical efficiency, as well as Stage 1 and conditional Stage 2 acceptance rates; it shows that $\gamma = 2.5$ gives close to the optimal efficiency, with $\widehat{\gamma} \approx 3.1$, and $\widehat{m} \approx 220 - 250$ as predicted. The empirical efficiency gain from using the DAPsMRWM algorithm compared to the pseudo-marginal RWM algorithm was $0.441/0.067 \approx 6.6$, which is in the centre of the range predicted by the theory. Finally, Proposition 4.2 proves that, subject to assumptions, the Stage 2 acceptance probability decreases as the variance in the log-posterior (σ^2) increases and the Stage 1 acceptance probability decreases as the scaling increases; these patterns are observed in our experiments (see Table 1).

8 Discussion

We have provided a theoretical analysis of the delayed-acceptance pseudo-marginal random walk Metropolis algorithm (DAPsMRWM) in the limit as the dimension, d, of the parameter space tends to infinity. Our analysis also applies to the delayed-acceptance random walk Metropolis (DARWM).

As with many other analyses [RGG97, RR98] we assume that the target has an iid product form. We then follow [STRR15] and [DPDK15] in assuming that the noise in the unbiased estimate of the posterior is additive on the logarithmic scale, with a distribution which is independent of the current position. We also assume that a cheap deterministic approximation is available for each component of the product, and that the error in each such approximation is a realisation of a random function. Individual realisations of the error are subject to only minor regularity conditions. As such, the error model is reasonably general and should capture the main characteristics of many real, deterministic approximations. This is verified for a toy Bayesian inverse problem. We examine the above model as dimension $d \to \infty$ and we obtain limiting forms for the Stage One and the conditional Stage Two acceptance rates and the expected squared jump distance. We also obtain a diffusion approximation for the first component of the target, which justifies the use of expected squared jump distance as a measure of efficiency.

For the DARWM, and for the DAPsMRWM subject to the assumption of the Standard Asymptotic Regime, introduced in Section 2.2, we obtain simplified forms for the acceptance rates and for the efficiency in terms of both the mixing of the Markov chain and of the computational time. The Standard Asymptotic Regime applies, for example in cases where likelihood estimates are obtained using a particle filter, or a product of importance sampling estimates. We show that when compared to the optimally tuned non-DA algorithm, the relative changes in the efficiency, optimal scaling and optimal variance can be characterised by the relative cost of the cheap approximation to the full evaluation and by its accuracy. The accuracy can be expressed in terms of the conditional stage two acceptance rate of the DA algorithm at the parameter value that was optimal for the non-DA algorithm. For the DAPsMRWM, the theory also shows that, except for small values of σ^2, the optimal scaling $\mu > 0$ is almost independent of the variance σ^2 and, hence, of the number of particles used. Consequently, as an alterantive tuning route, the two-dimensional optimisation over the jump scaling and the number of particles can be reduced to two one-dimensional optimisations – this results greatly simplifies the practical tuning of the DAPsMRWM. The theoretical work also suggests that even for a very accurate approximation the DAPsMRWM is only worth implementing if the cheap approximation is at least ten times quicker to compute than the target itself when $\sigma^2 = 1$.

The theoretical work supports the intuition that, provided the cheap deterministic approximation is fast and reasonably accurate, the DAPsMRWM and DARWM algorithms should be optimally efficient when μ is much larger than (and the overall acceptance rate is much lower than) that of the equivalent (pseudo-marginal) RWM algorithm.

A Explicit expressions for the acceptance probabilities

Define $G(a, b) := \mathbb{E}\left[1 \wedge \exp(\mathbf{N}\left(a, b^2\right))\right] = \Phi(a/b) + \exp\left(a + b^2/2\right) \Phi(-b - a/b)$ with $\Phi : \mathbb{R} \to [0, 1]$ the standard Gaussian cumulative distribution function. Then

$$\alpha_1(\mu, \sigma; \beta_1, \beta_2) = G\Big(-\frac{\mu^2}{2}(1 - \beta_1),\, \mu^2 \left(1 + \beta_2^2 - 2\beta_1\right)\Big). \tag{A.1}$$

Further, we may rewrite

$$Q_\Delta^\infty = -\frac{1}{2}\mu^2 + \mu\frac{\beta_1}{\beta_2}\xi + \mathbf{N}\left(0, \mu^2 - \mu^2\frac{\beta_1^2}{\beta_2^2}\right)$$
$$S_\Delta^\infty = \frac{\beta_1}{2}\mu^2 - \mu\beta_2\xi,$$

where $\xi \sim \mathbf{N}(0, 1)$ is independent of any other source of variability. Thus

$$\alpha_{12}(\mu, \sigma^2; \beta_1, \beta_2) = \mathbb{E}\left[G\left(-\frac{\mu^2}{2}(1 - \beta_1) + \mu\left(\frac{\beta_1}{\beta_2} - \beta_2\right)\xi, \mu^2 - \mu^2\frac{\beta_1^2}{\beta_2^2}\right) G\left(-\frac{\beta_1}{2}\mu^2 - \sigma^2 + \mu\beta_2\xi, 2\sigma^2\right)\right]. \tag{A.2}$$

A.1 Contour plots against β_1 and β_2 for the DAPsMRWM

Figure 9 shows contour plots of $\alpha_{2|1}(\widehat{\mu}_{pm}, \widehat{\sigma}^2_{pm})$, $\widehat{\mu}_{dapm}/\widehat{\mu}_{pm}$, $\widehat{\sigma}^2_{dapm}/\widehat{\sigma}^2_{pm}$ and $\mathrm{Eff}_{\mathrm{rel}}(\widehat{\mu}_{dapm}, \widehat{\sigma}^2_{dapm})$, as a function of β_1 and β_2 for $\eta = 0.01$.

B Proof of technical results

In this section we denote by $\Phi(x) = \int_{-\infty}^{x} \varphi(u)\, du$ the cumulative Gaussian function with $\varphi(u) = e^{-u^2/2}/\sqrt{2\pi}$. The bound $1 - \Phi(x) < \varphi(x)/x$ for $x > 0$ is used in several places.

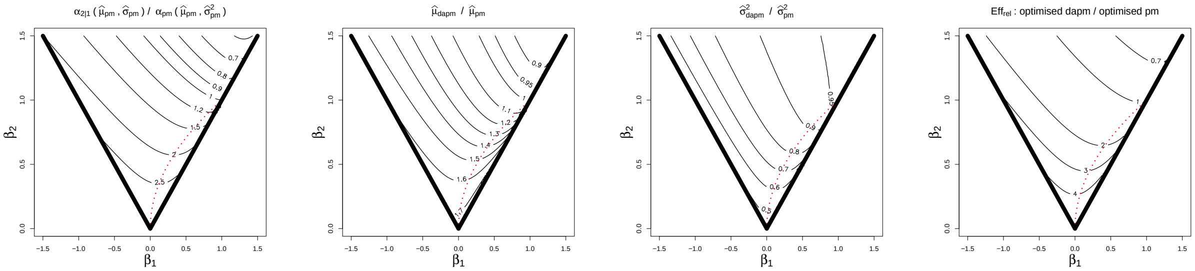

Figure 9: Contour plots of $\alpha_{2|1}(\widehat{\mu}_{pm}, \widehat{\sigma}^2_{pm}; \beta_1, \beta_2)$ (left), $\widehat{\mu}_{dapm}/\widehat{\mu}_{pm}$, $\widehat{\sigma}^2_{dapm}/\widehat{\sigma}^2_{pm}$ and $\mathrm{Eff}_{\mathrm{rel}}(\widehat{\mu}_{dapm}, \widehat{\sigma}^2_{dapm})$ (right), as a function of β_1 and β_2 for $\eta = 0.01$. The red, dotted line satisfies $\beta_1 = \beta_2^2$.

B.1 Proof of Proposition 4.2

The only not entirely trivial parts of this proposition involve establishing that α_1 and $\alpha_{2|1}$ are decreasing in μ and σ respectively. For proving that $\alpha_1 = G\Big(-\frac{\mu^2}{2}(1-\beta_1),\, \mu^2\left(1+\beta_2^2-2\beta_1\right)\Big)$ is decreasing as a function of μ when $\beta_1 < 1$, note that since $|\beta_1| < \beta_2$, $1+\beta_2^2-2\beta_1 \geq (1-\beta_1)^2$; hence it suffices to show that for any positive constant $c > 0$ the function $h : \mu \mapsto G(-\mu^2, c^2\mu^2)$ is decreasing. Since $h(\mu) = \mathbb{E}\left[F(-\mu^2 + c\,\mu\,\xi)\right]$ for a random variable $\xi \overset{\mathcal{D}}{\sim} \mathbf{N}(0,1)$ and $F'(x) = e^x\,\mathbb{I}(x<0)$ it follows that

$$h'(\mu) = \int_{z\in\mathbb{R}} F'(-\mu^2 + c\,\mu\,z)\,(-2\,\mu + c\,z)\,\varphi(z)\,dz = \int_{z<\mu/c} F'(-\mu^2 + c\,\mu\,z)\,(-2\,\mu + c\,z)\,\varphi(z)\,dz.$$

This quantity is negative since $-2\,\mu + c\,z < 0$ on the event $\{z : z < \mu/c\}$. Proving that $\alpha_{2|1}$ is decreasing as a function of σ readily follows from the fact that for any fixed $a \in \mathbb{R}$ the derivative of the function $\sigma \mapsto G(-\sigma^2 + a, 2\sigma^2) < -\sqrt{2}\varphi(-\sigma/\sqrt{2} + a/(\sigma\sqrt{2})) < 0$ and differentiation under the integral sign.

B.2 Proof of Theorem 5.1

Since $\mathrm{Eff}(\mu,\sigma^2) = \frac{\mu^2\,\alpha_{12}(\mu,\sigma)}{\eta+\alpha_1(\mu)/\sigma^2}$, for a fixed value of scaling $\mu > 0$ the efficiency functional goes to zero as $\sigma \to 0$ and $\sigma \to \infty$. Similarly, the fact that the efficiency goes to zero as $\mu \to 0$ for any fixed value of $\sigma > 0$ is straightforward; it remains to verify that the efficiency also converge to zero as $\mu \to \infty$. It suffices to show that $\mu^2\,\alpha_{12}(\mu,\sigma) \to 0$; since for any $x, y \in \mathbb{R}$ we have $\min\left(1, e^x\right)\,\min\left(1, e^y\right) \leq \min\left(1, e^{x+y}\right)$,

$$\alpha_{12} \leq \mathbb{E}\left[F\left(Q^{\infty}_{\Delta} + W_{\Delta}\right)\right] = 2\,\Phi\left\{-\frac{(\mu^2+2\sigma^2)^{1/2}}{2}\right\}$$

and the conclusion readily follows.

B.3 Proof of Equation (6.8)

Equation (3.8) yields that $R \equiv W^* - W$ for $(W^*, W) \sim \pi_{W^*} \otimes \pi_W$ has a density π_R such that the function $r \mapsto e^{r/2}\,\pi_R(r)$ is symmetric i.e. $e^{r/2}\,\pi_R(r) = e^{-r/2}\,\pi_R(-r)$. Similarly, algebra reveals that the joint Gaussian density $\pi_{Q,S}(q,s)$ of the pair $(Q^{\infty}_{\Delta}, S^{\infty}_{\Delta})$ described in Lemma 4.1 is such that

$$e^{q/2}\,\pi_{Q,S}(q,s) = e^{-q/2}\,\pi_{Q,S}(-q,-s).$$

That is because $-\log\pi_{Q,S}(q,s) = a\,q^2 + b\,s^2 + c\,qs - q/2 + \text{(constant)}$ for some coefficients $a, b, c \in \mathbb{R}$. Consequently, since the accept reject function F is such that $e^{-u}F(u) = F(-u)$ for any $u \in \mathbb{R}$, the function

$$\begin{aligned} g(q,r,s) &= e^{(q+s)/2}\,F(r-s)\,\pi_{Q,S}(q,s)\,\pi_R(r) \\ &= e^{-(r-s)/2}\,F(r-s)\left(e^{q/2}\,\pi_{Q,S}(q,s)\right)\left(e^{r/2}\,\pi_R(r)\right) \end{aligned}$$

is such that $g(q,r,s) = g(-q,-r,-s)$. It follows that

$$\begin{aligned}
\mathbb{E}\left[A(W)\right] = \mathbb{E}\left[F'(Q_\Delta^\infty + S_\Delta^\infty) \times F(R - S_\Delta^\infty)\right] &= \iiint_{\mathbb{R}^3} F'(q+s)\, F(r-s)\, \pi_{Q,S}(q,s)\, \pi_R(r)\, dq\, dr\, ds \\
&= \iiint_{\mathbb{R}^3} e^{-(q+s)/2} F'(q+s)\, g(q,r,s)\, dq\, dr\, ds = \iiint_{\mathbb{R}^3} e^{(q+s)/2} F'(-[q+s])\, g(q,r,s)\, dq\, dr\, ds \\
&= \mathbb{E}\left[e^{Q_\Delta^\infty + S_\Delta^\infty} F'(-[Q_\Delta^\infty + S_\Delta^\infty]) F(R - S_\Delta^\infty)\right].
\end{aligned}$$

Consequently, since $F'(u) + e^u\, F'(-u) = F(u)$ for $u \in \mathbb{R}$, it follows that

$$\begin{aligned}
2 \times \mathbb{E}\left[A(W)\right] &= \mathbb{E}\left[F'(Q_\Delta^\infty + S_\Delta^\infty) \times F(R - S_\Delta^\infty) + e^{Q_\Delta^\infty + S_\Delta^\infty} F'(-[Q_\Delta^\infty + S_\Delta^\infty]) F(R - S_\Delta^\infty)\right] \\
&= \mathbb{E}\left[F(Q_\Delta^\infty + S_\Delta^\infty)\right] \equiv \alpha_{12}.
\end{aligned}$$

The proof that $\mathbb{E}\left[B(W)\right] = \alpha_{12}/2$ is similar and thus omitted.

B.4 Proof of Lemma 6.1

In this section we need to consider asymptotic expansions of the type $\mathbb{E}_{\mathbf{x},w}[\ldots] = \Psi(\mathbf{x},w) + (\text{error term})$, where $(\mathbf{x},w) \in \mathbb{R}^d \times \mathbb{R}$ and $(\text{error term}) = \varepsilon_d(\mathbf{x},w)$ for a function $\varepsilon_d : \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}$. We use the notation $(\text{error term}) = o_{L^2}(1)$ to indicates that, under the equilibrium distribution, the moment of order two of the error term is asymptotically negligible, $\mathbb{E}\left[\varepsilon_d(\mathbf{X}^{(d)}, W)^2\right] \to 0$ as $d \to \infty$ for $(\mathbf{X}^{(d)}, W) \overset{\mathcal{D}}{\sim} \pi^{(d)} \otimes \pi_W$. Since φ is smooth with compact support, a second order Taylor expansion reveals that

$$\mathscr{L}^{(d)} \varphi(\mathbf{x}, w) = (\text{drift term})\, \varphi'(\mathbf{x}) + (1/2)\, (\text{volatility term})\, \varphi''(\mathbf{x}) + o_{L^2}(1) \tag{B.1}$$

where the drift and volatility terms are given by the following conditional expectations,

$$\begin{cases}
(\text{drift term}) & = \; (1/\delta) \times \mathbb{E}_{\mathbf{x},w}\left[\left(X_{1,1}^{(d),*} - x_1\right)\, \alpha_{12}^{(d)}\left(\mathbf{x}, w, \mathbf{X}^{(d),*}, W^{(d),*}\right)\right] \\
(\text{volatility term}) & = \; (1/\delta) \times \mathbb{E}_{\mathbf{x},w}\left[\left(X_{1,1}^{(d),*} - x_1\right)^2\, \alpha_{12}^{(d)}\left(\mathbf{x}, w, \mathbf{X}^{(d),*}, W^{(d),*}\right)\right]
\end{cases} \tag{B.2}$$

with $\mathbf{X}_1^{(d),*} = \mathbf{x} + (\mu/I)\, \delta^{1/2}\, \mathbf{Z}^{(d)}$ and standard centred Gaussian random variable $\mathbf{Z}^{(d)} = (Z_1, \ldots, Z_d)$

- It readily follows from Lemma 4.1 that for π-almost every $\mathbf{x}$ we have

$$(\text{volatility term}) = \alpha_{12} \times (\mu/I)^2 = J(\mu) + o_{L^2}(1). \tag{B.3}$$

- For the drift term, we make use of the following integration-by-part formula, also known as Stein's identity,

$$\mathbb{E}\left[Z \times g(Z)\right] = \mathbb{E}\left[g'(Z)\right] \qquad \text{for} \qquad Z \overset{\mathcal{D}}{\sim} \mathbf{N}\left(0,1\right), \tag{B.4}$$

  which holds for any continuous and piecewise continuously differentiable function $g : \mathbb{R} \to \mathbb{R}$ such that $x \mapsto \max\left(g(x), g'(x)\right)$ is polynomially bounded. In what follows, $F'(u) = e^u\, \mathbb{I}_{u<0}$. The expression for $\alpha_{12}^{(d)}\left(\mathbf{x}, w, \mathbf{X}^{(d),*}, W^{(d),*}\right)$, identity (B.4) and standard algebraic manipulations yield that

$$\begin{aligned}
(\text{drift term}) \;&=\; \delta^{-1/2}\, (\mu/I)\, \mathbb{E}_{\mathbf{x},w}[Z_1\, \alpha_{12}^{(d)}(\mathbf{x}, w, \mathbf{X}^{(d),*}, W^{(d),*})] \\
&=\; (\mu/I)^2\, \mathbb{E}_{\mathbf{x},w}[F'(\mathsf{Q}_\Delta^{(d)} + \mathsf{S}_\Delta^{(d)})\, F(\mathsf{W}_\Delta^{(d)} - \mathsf{S}_\Delta^{(d)})\, \{\ell'(X_{1,1}^{(d),*}) + \partial_x \mathcal{S}(X_{1,1}^{(d),*}, \gamma_1)\}] \\
&\quad\; - (\mu/I)^2\, \mathbb{E}_{\mathbf{x},w}[F(\mathsf{Q}_\Delta^{(d)} + \mathsf{S}_\Delta^{(d)})\, F'(\mathsf{W}_\Delta^{(d)} - \mathsf{S}_\Delta^{(d)})\, \partial_x \mathcal{S}(X_{1,1}^{(d),*}, \gamma_1)] \\
&=\; (\mu/I)^2\, A(w)\, \ell'(x_1) + (\mu/I)^2\, [A(w) - B(w)]\, \partial_x \mathcal{S}(x_1, \gamma_1) + o_{L^2}(1),
\end{aligned} \tag{B.5}$$

  where the functions $A, B : \mathbb{R} \to \mathbb{R}^+$ are defined in Equation (6.7) and the quantities $\mathsf{Q}_\Delta^{(d)}, \mathsf{S}_\Delta^{(d)}$ and $\mathsf{W}_\Delta^{(d)}$ in Equation (6.1).

Plugging (B.5) and (B.3) into (B.1) shows that the limit

$$\lim_{d \to \infty} \mathbb{E}\left[\left|\mathscr{L}^{(d)} \varphi(\mathbf{X}^{(d)}, W) - \mathcal{A}\varphi(X_1^{(d)}, W)\right|^2\right] = 0$$

holds for $\left(\mathbf{X}^{(d)}, W\right) \sim \pi^{(d)} \otimes \pi_W$, as required.

B.5 Proof of Lemma 6.2

The strategy of the proof is as follows. We define three stochastic processes $\left\{W^{(d)}_{\clubsuit,k}\right\}_{k\geq 0}$, $\left\{W^{(d)}_{\spadesuit,k}\right\}_{k\geq 0}$, $\left\{W_{\blacksquare,k}\right\}_{k\geq 0}$ such that

$$\left\{\begin{array}{l} \lim_{d\to\infty}\ \mathbb{P}\left(W^{(d)}_{\clubsuit,k}=W^{(d)}_k\ :\ 0\leq k\leq T^{(d)}\right)\ =\ 1,\\ \left(W^{(d)}_{\clubsuit,k}=W^{(d)}_k\ :\ 0\leq k\leq T^{(d)}\right)\overset{\text{law}}{=}\left(W^{(d)}_{\spadesuit,k}=W^{(d)}_k\ :\ 0\leq k\leq T^{(d)}\right),\\ \lim_{d\to\infty}\ \mathbb{P}\left(W^{(d)}_{\spadesuit,k}=W_{\blacksquare,k}\ :\ 0\leq k\leq T^{(d)}\right)\ =\ 1,\\ \left\{W_{\blacksquare,k}\right\}_{k\geq 0}\text{ is a Markov chain that is ergodic with respect to }\pi_W. \end{array}\right. \tag{B.6}$$

Once (B.6) is proved, Lemma 6.2 immediately follows. Let us now defines these three processes and verify that Equation (B.6) holds. To do so, let us consider i.i.d sequences $\{X_i\}_{i\geq 1}$ and $\{W^*_i\}_{i\geq 1}$ and $\{Z_{i,k}\}_{i,k\geq 1}$ and $\{U_k\}_{k\geq 0}$ respectively marginally distributed as π and π_{W^*} and $\mathbf{N}(0,1)$ and Uniform$([0,1])$. We consider $\{x_i\}_{i\geq 1}$ a realisation of $\{X_i\}_{i\geq 1}$ and for any index $d\geq 1$ we set $\mathbf{X}^{(d)}_0=(x_1,\ldots,x_d)$ and $W^{(d)}_0\overset{\mathcal{D}}{\sim}\pi_W$ and recursively define $\left(\mathbf{X}^{(d)}_{k+1},W^{(d)}_{k+1}\right)=\left(\mathbf{X}^{(d),*}_k,W^*_k\right)$, with $\mathbf{X}^{(d),*}_k=\mathbf{X}^{(d)}_k+(\mu/I)\,\delta^{1/2}\,\mathbf{Z}^{(d)}_k$ and $\mathbf{Z}^{(d)}_k=(Z_{1,k},\ldots,Z_{d,k})$, if

$$U_k\leq F\left(Q^{(d)}_{\Delta,k}+S^{(d)}_{\Delta,k}\right)\times F\left(W^*_k-W^{(d)}_k-S^{(d)}_{\Delta,k}\right) \tag{B.7}$$

and $\left(\mathbf{X}^{(d)}_{k+1},W^{(d)}_{k+1}\right)=\left(\mathbf{X}^{(d)}_k,W^{(d)}_k\right)$ otherwise. In the above

$$\left\{\begin{array}{ll} Q^{(d)}_{\Delta,k} & =\sum_{i=1}^d\ell\left(X^{(d),*}_{k,i}\right)-\ell\left(X^{(d)}_{k,i}\right)\\ S^{(d)}_{\Delta,k} & =\sum_{i=1}^d\mathcal{S}\left(X^{(d),*}_{k,i},\gamma_i\right)-\mathcal{S}\left(X^{(d)}_{k,i},\gamma_i\right). \end{array}\right.$$

Indeed, for any index $d\geq 1$ the process $\left\{\left(\mathbf{X}^{(d)}_k,W^{(d)}_k\right)\right\}_{k\geq 0}$ is a DAPsMRWM Markov chain that targets $\pi^{(d)}\otimes\pi_W$. Let us now define the processes $W_\clubsuit$,$W_\spadesuit$,$W_\blacksquare$.

- We set $W^{(d)}_{\clubsuit,0}=W^{(d)}_0$ and recursively define $W^{(d)}_{\clubsuit,k+1}=W^*_k$ if

$$U_k\leq F\left(Q^{(d)}_{\clubsuit,\Delta,k}+S^{(d)}_{\clubsuit,\Delta,k}\right)\times F\left(W^*_k-W^{(d)}_{\clubsuit,k}-S^{(d)}_{\clubsuit,\Delta,k}\right) \tag{B.8}$$

  and $W^{(d)}_{\clubsuit,k+1}=W^{(d)}_{\clubsuit,k}$ otherwise; we have used the notations

$$\left\{\begin{array}{ll} Q^{(d)}_{\clubsuit,\Delta,k} & =(\mu\delta/I)\ \sum_{i=1}^d\ell'(x_i)Z_{i,k}+(\mu^2\delta^2/2\,I^2)\ \sum_{i=1}^d\ell''(x_i)\\ S^{(d)}_{\clubsuit,\Delta,k} & =(\mu\delta/I)\ \sum_{i=1}^d\mathcal{S}'(x_i,\gamma_i)Z_{i,k}+(\mu^2\delta^2/2\,I^2)\ \sum_{i=1}^d\mathcal{S}''(x_i,\gamma_i). \end{array}\right.$$

- Similarly, we set $W^{(d)}_{\spadesuit,0}=W^{(d)}_0$ and recursively define $W^{(d)}_{\spadesuit,k+1}=W^*_k$ if

$$U_k\leq F\left(Q^{(d)}_{\spadesuit,\Delta,k}+S^{(d)}_{\spadesuit,\Delta,k}\right)\times F\left(W^*_k-W^{(d)}_{\spadesuit,k}-S^{(d)}_{\spadesuit,\Delta,k}\right) \tag{B.9}$$

  and $W^{(d)}_{\spadesuit,k+1}=W^{(d)}_{\spadesuit,k}$ otherwise; we have used the notations $\left(Q^{(d)}_{\spadesuit,\Delta,k},S^{(d)}_{\spadesuit,\Delta,k}\right)$ to designate a Gaussian random variable in $\mathbb{R}^2$, independent from any other source of randomness, with same law as $\left(Q^{(d)}_{\clubsuit,\Delta,k},S^{(d)}_{\clubsuit,\Delta,k}\right)$.

- Finally, we set $W^{(d)}_{\blacksquare,0}=W^{(d)}_0$ and recursively define $W^{(d)}_{\blacksquare,k+1}=W^*_k$ if

$$U_k\leq F\left(Q^{(\infty)}_{\Delta,k}+S^{(\infty)}_{\Delta,k}\right)\times F\left(W^*_k-W^{(d)}_{\blacksquare,k}-S^{(\infty)}_{\Delta,k}\right) \tag{B.10}$$

  and $W^{(d)}_{\blacksquare,k+1}=W^{(d)}_{\blacksquare,k}$ otherwise; in the above $\left\{\left(Q^{(\infty)}_{\Delta,k},S^{(\infty)}_{\Delta,k}\right)\right\}_{k\geq 0}$ is an i.i.d sequence marginally distributed as $\left(Q^{(\infty)}_\Delta,S^{(\infty)}_\Delta\right)$; see Lemma 4.1.

It is obvious that $\left\{W^{(d)}_{\clubsuit,k}\right\}_{k\geq 0}$ and $\left\{W^{(d)}_{\spadesuit,k}\right\}_{k\geq 0}$ have the same law. The fact that $\left\{W^{(d)}_{\blacksquare,k}\right\}_{k\geq 0}$ is a Markov chain ergodic with respect to π_W readily follows from the fact that it is reversible with respect to π_W; it is a standard Gaussian computation. The proof of the first and third equation in (B.6) is based on the following basic remark. For convenience, let us denote by $\mathcal{E}^{(d)}_k$,$\mathcal{E}^{(d)}_{k,\clubsuit}$,$\mathcal{E}^{(d)}_{k,\spadesuit}$,$\mathcal{E}^{(d)}_{k,\infty}$ the Bernoulli random variables indicating whether or not the respective events (B.7),(B.8),(B.9), (B.10) are realised or not. We have

$$1-\mathbb{P}\left(W^{(d)}_{\clubsuit,k}=W^{(d)}_k \;:\; 0\leq k\leq T^{(d)}\right)\leq \sum_{k=0}^{T^{(d)}-1}\mathbb{P}\left(\mathcal{E}^{(d)}_k\neq\mathcal{E}^{(d)}_{k,\clubsuit}\,\middle|\,W^{(d)}_{\clubsuit,k}=W^{(d)}_k\right) \tag{B.11}$$

and the conditional probability $\mathbb{P}\left(\mathcal{E}^{(d)}_k\neq\mathcal{E}^{(d)}_{k,\clubsuit}\,\middle|\,W^{(d)}_{\clubsuit,k}=W^{(d)}_k\right)$ is less than the expectation, conditioned upon the event $\left\{W^{(d)}_{\clubsuit,k}=W^{(d)}_k\right\}$, of the absolute difference

$$\left|F\left(Q^{(d)}_{\Delta,k}+S^{(d)}_{\Delta,k}\right)\,F\left(W^*_k-W^{(d)}_k-S^{(d)}_{\Delta,k}\right)-F\left(Q^{(d)}_{\clubsuit,\Delta,k}+S^{(d)}_{\clubsuit,\Delta,k}\right)\,F\left(W^*_k-W^{(d)}_{\clubsuit,k}-S^{(d)}_{\clubsuit,\Delta,k}\right)\right|. \tag{B.12}$$

Because the $[0,1]$-valued function F is assumed to be Lipschitz, if $W^{(d)}_{\clubsuit,k}=W^{(d)}_k$ the absolute difference in (B.12) is less than $2\times\|F\|_{\text{Lip}}\times\left\{\left|Q^{(d)}_{\Delta,k}-Q^{(d)}_{\clubsuit,\Delta,k}\right|+\left|S^{(d)}_{\Delta,k}-S^{(d)}_{\clubsuit,\Delta,k}\right|\right\}$. Because the second and third derivatives of the log-likelihood function ℓ are globally bounded, a third order Taylor expansion yield that

$$\begin{aligned}
\mathbb{E}\left|Q^{(d)}_{\Delta,k}-Q^{(d)}_{\clubsuit,\Delta,k}\right| &\lesssim d^{-1/2}\,\mathbb{E}\left|\sum_{i=1}^{d}\left(\ell'(X^{(d)}_{k,i})-\ell'(x_i)\right)\,Z_{i,k}\right|+d^{-1}\,\mathbb{E}\left|\sum_{i=1}^{d}\left(\ell''(X^{(d)}_{k,i})-\ell''(x_i)\right)\,Z^2_{i,k}\right|+\mathcal{O}(d^{-1/2})\\
&\lesssim d^{-1/2}\left\{\sum_{i=1}^{d}\mathbb{E}\left[\left(\ell'(X^{(d)}_{k,i})-\ell'(x_i)\right)^2\right]\right\}^{1/2}+d^{-1}\left\{\sum_{i=1}^{d}\mathbb{E}\left[\left(\ell''(X^{(d)}_{k,i})-\ell''(x_i)\right)^2\right]\right\}^{1/2}+\mathcal{O}(d^{-1/2})\\
&=\mathcal{O}(k\,d^{-1/2}).
\end{aligned}$$

We have used the fact that for any exponent $p\geq 1$ we have $\mathbb{E}\left[\left|X^{(d)}_{k,i}-x_i\right|^p\right]^{1/p}\lesssim k\,d^{-1/2}$, which readily follows from the triangular inequality. Similarly, we have that $\mathbb{E}\left|S^{(d)}_{\Delta,k}-S^{(d)}_{\clubsuit,\Delta,k}\right|\lesssim k\,d^{-1/2}$. Plugging these estimates in (B.11) shows that

$$1-\mathbb{P}\left(W^{(d)}_{\clubsuit,k}=W^{(d)}_k \;:\; 0\leq k\leq T^{(d)}\right)\lesssim d^{-1/2}\sum_{k=0}^{T^{(d)}-1}k\;\to\;0$$

since $T^{(d)}=d^{\gamma}$ for some exponent $\gamma\in(0,1/4)$; we have thus proved that $\mathbb{P}\left(W^{(d)}_{\clubsuit,k}=W^{(d)}_k \;:\; 0\leq k\leq T^{(d)}\right)$ converges to one as $d\to\infty$. The proof of the estimate $\mathbb{P}\left(W^{(d)}_{\spadesuit,k}=W^{(d)}_{\blacksquare,k} \;:\; 0\leq k\leq T^{(d)}\right)\to 1$ uses the same ingredients and is thus omitted.

C DARWM simulation study on Gaussian target with logistic approximation

Table 2 lists the values of φ_1 and φ_2 used for the thirteen different logistic approximations, π_a, together with the relevant acceptance rates.

D Marginal likelihood under the linear noise approximation

For simplicity of exposition we assume an observation regime of the form $\mathbf{Y}_t=\mathbf{U}_t+\boldsymbol{\varepsilon}_t$ with $\boldsymbol{\varepsilon}_t\sim\mathbf{N}\left(0,\boldsymbol{\Sigma}\right)$ where $\boldsymbol{\varepsilon}_t$ is a length-d_x Gaussian random vector. Suppose that $\mathbf{U}_1$ is fixed at some value $\mathbf{u}_1$. The marginal likelihood (and hence the posterior up to proportionality) under the LNA, $\pi_a(\mathbf{y}_{1:n}|\mathbf{x})$ can be obtained as follows.

Algorithm	φ_1	φ_2	β_1	β_2	α_1	$\alpha_{2\|1}$
RWM					0.2616	
DA	0.0	0.6	0.834	0.834	0.261	0.128
DA	0.0	1.2	0.441	0.449	0.069	0.533
DA	0.0	1.8	-0.042	0.262	0.041	0.738
DA	0.0	2.3	-0.467	0.649	0.034	0.595
DA	0.0	2.7	-0.810	1.025	0.032	0.492
DA	0.5	1.2	0.466	0.552	0.370	0.547
DA	1.0	1.2	0.535	0.763	0.140	0.151
DA	1.5	1.2	0.630	0.979	0.482	0.276
DA	0.6	1.8	0.056	0.681	0.0650	0.279
DA	0.5	2.3	-0.351	0.941	0.049	0.289
DA	0.0	1.5–2.0			0.248	0.772
DA	0.0	1.2–2.7			0.238	0.609
DA	0.0–1.0	1.2			0.377	0.517

Table 2: Values of φ_1 and φ_2 used in (7.1), and the corresponding values of β_1 and β_2 (where calculable), $\alpha_1(\widehat{\lambda}_{rwm})$ and $\alpha_{2|1}(\widehat{\lambda}_{rwm})$.

1. Initialisation. Compute $\pi_a(\mathbf{y}_1|\mathbf{x}) = \varphi\left(\mathbf{y}_1\,;\,\mathbf{u}_1\,,\,\mathbf{\Sigma}\right)$ where $\varphi\left(\mathbf{y}_1\,;\,\mathbf{u}_1\,,\,\mathbf{\Sigma}\right)$ denotes the Gaussian density with mean vector $\mathbf{u}_1$ and variance matrix $\mathbf{\Sigma}$. Set $\mathbf{a}_1 = \mathbf{u}_1$ and $\mathbf{C}$ to be the $d_x \times d_x$ matrix of zeros.

2. For times $t = 1, 2, \ldots, n-1$,

   (a) Prior at $t+1$. Initialise the LNA with $\mathbf{z}_t = \mathbf{a}_t$, $\mathbf{m}_t = 0$ and $\mathbf{V}_t = C_t$. Note that $\mathbf{m}_s = \mathbf{0}$ for all $s > t$. Integrate the ODE system (7.2) forward to $t+1$ to obtain $\mathbf{z}_{t+1}$ and $\mathbf{V}_{t+1}$. Hence $\mathbf{X}_{t+1}|\mathbf{y}_{1:t} \sim \mathbf{N}\left(\mathbf{z}_{t+1}, \mathbf{V}_{t+1}\right)$.

   (b) One-step forecast. Using the observation equation, we have that $\mathbf{Y}_{t+1}|\mathbf{y}_{1:t} \sim \mathbf{N}\left(\mathbf{z}_{t+1}, \mathbf{V}_{t+1} + \mathbf{\Sigma}\right)$. Compute $\pi_a(\mathbf{y}_{1:t+1}|\mathbf{x}) = \pi_a(\mathbf{y}_{1:t}|\mathbf{x})\, \varphi\left(\mathbf{y}_{t+1}\,;\,\mathbf{z}_{t+1}\,,\,\mathbf{V}_{t+1} + \mathbf{\Sigma}\right)$.

   (c) Posterior at $t+1$. Combining the distributions in (a) and (b) gives $\mathbf{U}_{t+1}|\mathbf{y}_{1:t+1} \sim \mathbf{N}\left(\mathbf{a}_{t+1}, \mathbf{C}_{t+1}\right)$ where $\mathbf{a}_{t+1} = \mathbf{z}_{t+1} + \mathbf{V}_{t+1}(\mathbf{V}_{t+1} + \mathbf{\Sigma})^{-1}\left(\mathbf{y}_{t+1} - \mathbf{z}_{t+1}\right)$ and $\mathbf{C}_{t+1} = \mathbf{V}_{t+1} - \mathbf{V}_{t+1}((\mathbf{V}_{t+1} + \mathbf{\Sigma})^{-1}\mathbf{V}_{t+1}$.

Acknowledgements

AHT acknowledges support from the Singapore Ministry of Education Tier 2 (MOE2016-T2-2-135) and a Young Investigator Award Grant (NUSYIA FY16 P16; R-155-000-180-133).

References

[AB01] Siu-Kui Au and James L. Beck. Estimation of small failure probabilities in high dimensions by subset simulation. *Probabilistic Engineering Mechanics*, 16(4):263 – 277, 2001.

[ADH10] C. Andrieu, A. Doucet, and R. Holenstein. Particle Markov chain Monte Carlo methods. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 72(3):269–342, 2010.

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

[AV15] C. Andrieu and M. Vihola. Convergence properties of pseudo marginal Markov chain Monte Carlo algorithms. *Ann. Appl. Probab.*, 25(2):1030–1077, 2015.

[BC14] Simon Barthelmé and Nicolas Chopin. Expectation propagation for likelihood-free inference. *Journal of the American Statistical Association*, 109(505):315–333, 2014.

[BDM12] M. Bédard, R. Douc, and E. Moulines. Scaling analysis of multiple-try MCMC methods. *Stoch. Proc. Appl.*, 122(3):758–786, 2012.

[BDMD13] J. Bérard, P. Del Moral, and A. Doucet. A lognormal central limit theorem for particle approximations of normalizing constants. *arXiv preprint arXiv:1307.0181*, 2013.

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

[Béd07] M. Bédard. Weak convergence of Metropolis algorithms for non-i.i.d. target distributions. *Ann. Appl. Probab.*, 17(4):1222–1244, 2007.

[BGJM11] S. Brooks, A. Gelman, G. L. Jones, and X.-L. Meng, editors. *Handbook of Markov chain Monte Carlo*. Chapman & Hall/CRC Handbooks of Modern Statistical Methods. CRC Press, Boca Raton, FL, 2011.

[BGLR15] Marco Banterle, Clara Grazian, Anthony Lee, and Christian P. Robert. Accelerating Metropolis-Hastings algorithms by Delayed Acceptance. *arXiv e-prints*, page arXiv:1503.00996, Mar 2015.

[BPS04] L. A. Breyer, M. Piccioni, and S. Scarlatti. Optimal scaling of MALA for nonlinear regression. *Ann. Appl. Probab.*, 14(3):1479–1505, 2004.

[BR08] Mylène Bédard and Jeffrey S. Rosenthal. Optimal scaling of Metropolis algorithms: heading toward general target distributions. *Canad. J. Stat.*, 36:483–503, 2008.

[BRS09] A. Beskos, G. O. Roberts, and A. Stuart. Optimal scalings for local Metropolis-Hastings chains on nonproduct targets in high dimensions. *Ann. Appl. Probab.*, 19(3):863–898, 2009.

[BWK08] R. J. Boys, D. J. Wilkinson, and T. B. L. Kirkwood. Bayesian inference for a discretely observed stochastic-kinetic model. *Stat. Comput.*, 18:125–135, 2008.

[CB18] Thomas A. Catanach and James L. Beck. Bayesian Updating and Uncertainty Quantification using Sequential Tempered MCMC with the Rank-One Modified Metropolis Algorithm. *arXiv e-prints*, page arXiv:1804.08738, Apr 2018.

[CF05] J. A. Christen and C. Fox. Markov chain Monte Carlo using an approximation. *J. Comp. Graph. Stat.*, 14(4):795–810, 2005.

[CFO11] T Cui, C Fox, and MJ O'Sullivan. Bayesian calibration of a large-scale geothermal reservoir model by a new adaptive delayed acceptance metropolis hastings algorithm. *Water Resources Research*, 47(10), 2011.

[DLCMR17] Alain Durmus, Sylvain Le Corff, Eric Moulines, and Gareth O Roberts. Optimal scaling of the random walk metropolis algorithm under l p mean differentiability. *Journal of Applied Probability*, 54(4):1233–1260, 2017.

[DM04] P. Del Moral. *Feynman-Kac formulae*. Probability and its Applications (New York). Springer-Verlag, New York, 2004. Genealogical and interacting particle systems with applications.

[DPDK15] A. Doucet, M. K. Pitt, G. Deligiannidis, and R. Kohn. Efficient implementation of Markov chain Monte Carlo when using an unbiased likelihood estimator. *Biometrika*, 102(2):295–313, 03 2015.

[DS19] Johan Dahlin and Thomas B. Schön. Getting Started with Particle Metropolis-Hastings for Inference in Nonlinear Dynamical Models. *arXiv e-prints*, page arXiv:1511.01707, Nov 2019.

[EDGG$^+$05] Y Efendiev, A Datta-Gupta, V Ginting, X Ma, and B Mallick. An efficient two-stage markov chain monte carlo method for dynamic data integration. *Water Resources Research*, 41(12), 2005.

[EHL06] Y. Efendiev, T. Hou, and W. Luo. Preconditioning markov chain monte carlo simulations using coarse-scale models. *SIAM Journal on Scientific Computing*, 28(2):776–803, 2006.

[EK86] S. N. Ethier and T. G. Kurtz. *Markov processes: Characterization and convergence*, volume 6. Wiley New York, 1986.

[ER17] Richard G. Everitt and Paulina A. Rowińska. Delayed acceptance ABC-SMC. *arXiv e-prints*, page arXiv:1708.02230, Aug 2017.

[FG14] Maurizio Filippone and Mark Girolami. Pseudo-marginal Bayesian inference for Gaussian processes. *IEEE Tran. Pattern Anal. Mach. Intell.*, 36(11):2214–2226, 2014.

[FGS14] P. Fearnhead, V. Giagos, and C. Sherlock. Inference for reaction networks using the Linear Noise Approximation. *Biometrics*, 70:457–466, 2014.

[FNL11] M. Fielding, D. J. Nott, and S.-Y. Liong. Efficient MCMC schemes for computationally expensive posterior distributions. *Technometrics*, 53(1), 2011.

[FS11] T. Flury and N. Shephard. Bayesian inference based only on simulated likelihood: particle filter analysis of dynamic economic models. *Econometric Theory*, 27(05):933–956, 2011.

[FV17] Jordan Franks and Matti Vihola. Importance sampling correction versus standard averages of reversible MCMCs in terms of the asymptotic variance. *arXiv e-prints*, page arXiv:1706.09873, Jun 2017.

[GHS15] Andrew Golightly, Daniel A. Henderson, and Chris Sherlock. Delayed acceptance particle mcmc for exact inference in stochastic kinetic models. *Statistics and Computing*, 25(5):1039–1055, Sep 2015.

[Gil77] D. T. Gillespie. Exact stochastic simulation of coupled chemical reactions. *J. Phys. Chem.*, 81:2340–2361, 1977.

[GRS96] W. R. Gilks, S. Richardson, and D. J. Spiegelhalter. *Markov Chain Monte Carlo in practice.* Chapman and Hall, London, UK, 1996.

[GW11] A. Golightly and D. J. Wilkinson. Bayesian parameter inference for stochastic biochemical network models using particle Markov chain Monte Carlo. *Interface Focus*, 1(6):807–820, 2011.

[HRM$^+$11] D. C. Higdon, S. J. Reese, D. Moulton, J. A. Vrugt, and C. Fox. Posterior exploration for computationally intensive forward models. In S. Brooks, A. Gelman, G. L. Jones, and X.-L. Meng, editors, *Handbook of Markov chain Monte Carlo*, chapter 16, pages 401–418. CRC Press, Boca Raton, FL, 2011.

[KdV12] J. Knape and P. de Valpine. Fitting complex population models by combining particle filters with Markov chain Monte Carlo. *Ecology*, 93(2):256–263, 2012.

[KS06] Jari Kaipio and Erkki Somersalo. *Statistical and computational inverse problems*, volume 160. Springer Science & Business Media, 2006.

[Liu01] J. S. Liu. *Monte Carlo Strategies In Scientific Computing.* Springer, 2001.

[MFS08] J. D. Moulton, C. Fox, and D. Svyatskiy. Multilevel approximations in sample-based inversion from the Dirichlet-to-Neumann map. *J. Phys.: Conf. Ser.*, 124(1), 2008.

[MPS12] J. C. Mattingly, N. S. Pillai, and A. M. Stuart. Diffusion limits of the random walk Metropolis algorithm in high dimensions. *Ann. Appl. Probab.*, 22(3):881–930, 2012.

[PBCV06] Martyn Plummer, Nicky Best, Kate Cowles, and Karen Vines. Coda: Convergence diagnosis and output analysis for mcmc. *R News*, 6(1):7–11, 2006.

[PDS11] G. Poyiadjis, A. Doucet, and S. S. Singh. Particle approximations of the score and observed information matrix in state space models with application to parameter estimation. *Biometrika*, 98(1):65–80, 2011.

[PdSSGK12] M. K. Pitt, R. dos Santos Silva, P. Giordani, and R. Kohn. On some properties of Markov chain Monte Carlo simulation methods based on the particle filter. *Journal of Econometrics*, 171(2):134 – 151, 2012.

[PG10] C. Pasarica and A. Gelman. Adaptively scaling the Metropolis algorithm using expected squared jumped distance. *Statistica Sinica*, 20(1):343, 2010.

[PST12] N. S. Pillai, A. M. Stuart, and A. H. Thiery. Optimal scaling and diffusion limits for the Langevin algorithm in high dimensions. *Ann. Appl. Probab.*, 22(6):2320–2356, 2012.

[PST14] N. S. Pillai, A. M. Stuart, and A. H. Thiery. Gradient flow from a random walk in Hilbert space. *Stochastic Partial Differential Equations: Analysis and Computations*, 2(2):196–232, 2014.

[QTVK18] Matias Quiroz, Minh-Ngoc Tran, Mattias Villani, and Robert Kohn. Speeding up mcmc by delayed acceptance and data subsampling. *Journal of Computational and Graphical Statistics*, 27(1):12–22, 2018.

[Ras03] C. E. Rasmussen. Gaussian processes to speed up hybrid Monte Carlo for expensive Bayesian integrals. In *Bayesian Statistics 7: Proceedings of the 7th Valencia International Meeting*, pages 651–659. Oxford University Press, 2003.

[RGG97] G. O. Roberts, A. Gelman, and W. R. Gilks. Weak convergence and optimal scaling of random walk Metropolis algorithms. *Ann. Appl. Probab.*, 7:110–120, 1997.

[RR98] G. O. Roberts and J. S. Rosenthal. Optimal scaling of discrete approximations to Langevin diffusions. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 60(1):255–268, 1998.

[RR01] G. O. Roberts and J. S. Rosenthal. Optimal scaling for various Metropolis-Hastings algorithms. *Statistical Science*, 16:351–367, 2001.

[RR14a] G. O. Roberts and J. S. Rosenthal. Complexity bounds for MCMC via diffusion limits. *arXiv preprint arXiv:1411.0712*, 2014.

[RR14b] G. O. Roberts and J. S. Rosenthal. Minimising MCMC variance via diffusion limits, with an application to simulated tempering. *Ann. Appl. Probab.*, 24(1):131–149, 2014.

[SDDP18] Sebastian M. Schmon, George Deligiannidis, Arnaud Doucet, and Michael K. Pitt. Large Sample Asymptotics of the Pseudo-Marginal Method. *arXiv e-prints*, page arXiv:1806.10060, Jun 2018.

[SFR10] C. Sherlock, P. Fearnhead, and G. O. Roberts. The random walk Metropolis: linking theory and practice through a case study. *Statist. Sci.*, 25(2):172–190, 2010.

[SGH17] Chris Sherlock, Andrew Golightly, and Daniel A. Henderson. Adaptive, delayed-acceptance mcmc for targets with expensive likelihoods. *Journal of Computational and Graphical Statistics*, 26(2):434–444, 2017.

[She13] C. Sherlock. Optimal scaling of the random walk Metropolis: general criteria for the 0.234 acceptance rule. *J. App. Prob.*, 50(1):1–15, 2013.

[She16] Chris Sherlock. Optimal scaling for the pseudo-marginal random walk metropolis: Insensitivity to the noise generating mechanism. *Methodology and Computing in Applied Probability*, 18(3):869–884, Sep 2016.

[SL17] Chris Sherlock and Anthony Lee. Variance bounding of delayed-acceptance kernels. *arXiv e-prints*, page arXiv:1706.02142, Jun 2017.

[Smi11] M. E. Smith. Estimating nonlinear economic models using surrogate transitions. Available from `https://files.nyu.edu/mes473/public/Smith_Surrogate.pdf`, 2011.

[SR09] C. Sherlock and G. O. Roberts. Optimal scaling of the random walk Metropolis on elliptically symmetric unimodal targets. *Bernoulli*, 15(3):774–798, 2009.

[STL17] Chris Sherlock, Alexandre H Thiery, and Anthony Lee. Pseudo-marginal metropolis–hastings sampling using averages of unbiased estimators. *Biometrika*, 104(3):727–734, 2017.

[STRR15] Chris Sherlock, Alexandre H. Thiery, Gareth O. Roberts, and Jeffrey S. Rosenthal. On the efficiency of pseudo-marginal random walk metropolis algorithms. *Ann. Statist.*, 43(1):238–275, 02 2015.

[Stu10] Andrew M Stuart. Inverse problems: a bayesian perspective. *Acta Numerica*, 19:451–559, 2010.

[Tho13] James William Thomas. *Numerical partial differential equations: finite difference methods*, volume 22. Springer Science & Business Media, 2013.

[VHF16] Matti Vihola, Jouni Helske, and Jordan Franks. Importance sampling type estimators based on approximate marginal MCMC. *arXiv e-prints*, page arXiv:1609.02541, Sep 2016.

[vK01] N. G. van Kampen. *Stochastic Processes in Physics and Chemistry.* North-Holland, 2001.

[YRR19] Jun Yang, Gareth O Roberts, and Jeffrey S Rosenthal. Optimal scaling of metropolis algorithms on general target distributions. *arXiv preprint arXiv:1904.12157*, 2019.

[ZBK17] Giacomo Zanella, Mylène Bédard, and Wilfrid S Kendall. A dirichlet form approach to mcmc optimal scaling. *Stochastic Processes and their Applications*, 127(12):4053–4082, 2017.