JYX JYVÄSKYLÄN YLIOPISTO UNIVERSITY OF JYVÄSKYLÄ

This is a self-archived version of an original article. This version may differ from the original in pagination and typographic details.

Author(s): Vihola, Matti; Helske, Jouni; Franks, Jordan

Title: Importance sampling type estimators based on approximate marginal Markov chain Monte Carlo

Year: 2020

Version: Accepted version (Final draft)

Copyright: © Wiley-Blackwell, 2020

Rights: In Copyright

Rights url: http://rightsstatements.org/page/InC/1.0/?language=en

Please cite the original version:

Vihola, M., Helske, J., & Franks, J. (2020). Importance sampling type estimators based on approximate marginal Markov chain Monte Carlo. Scandinavian Journal of Statistics, 47(4), 1339-1376. https://doi.org/10.1111/sjos.12492

Importance sampling type estimators based on approximate marginal MCMC

Matti Vihola[1*] | **Jouni Helske**[1,2*] | **Jordan Franks**[1,3,†]

[1]Department of Mathematics and Statistics, University of Jyvaskyla, Jyvaskyla, Finland

[2]Department of Science and Technology, Linköping University, Norrköping, Sweden

[3]Newcastle University, School of Mathematics, Statistics and Physics, NE1 7RU Newcastle-upon-Tyne, United Kingdom

Correspondence
Matti Vihola, University of Jyvaskyla, Department of Mathematics and Statistics, P.O. Box 35, FI-40014 University of Jyvaskyla, Finland
Email: matti.s.vihola@jyu.fi

Present address
* University of Jyvaskyla, Department of Mathematics and Statistics, P.O. Box 35, FI-40014 University of Jyvaskyla, Finland
† Newcastle University, School of Mathematics, Statistics and Physics, NE1 7RU Newcastle-upon-Tyne, United Kingdom

Funding information
The authors have been supported by the Academy of Finland grants 274740, 284513, 312605 and 315619.

We consider importance sampling (IS) type weighted estimators based on Markov chain Monte Carlo (MCMC) targeting an approximate marginal of the target distribution. In the context of Bayesian latent variable models, the MCMC typically operates on the hyperparameters, and the subsequent weighting may be based on IS or sequential Monte Carlo (SMC), but allows for multilevel techniques as well. The IS approach provides a natural alternative to delayed acceptance (DA) pseudo-marginal/particle MCMC, and has many advantages over DA, including a straightforward parallelisation and additional flexibility in MCMC implementation. We detail minimal conditions which ensure strong consistency of the suggested estimators, and provide central limit theorems with expressions for asymptotic variances. We demonstrate how our method can make use of SMC in the state space models context, using Laplace approximations and time-discretised diffusions. Our experimental results are promising and show that the IS type approach can provide substantial gains relative to an analogous DA scheme, and is often competitive even without parallelisation.

1 | INTRODUCTION

Markov chain Monte Carlo (MCMC) has become a standard tool in Bayesian analysis. The greatest benefit of MCMC is its general applicability — it is guaranteed to be consistent with virtually no assumptions on the underlying model. However, the practical applicability of MCMC generally depends on the dimension of the unknown variables, the number of data, and the computational resources available. Because MCMC is only asymptotically unbiased, and sequential in nature, it can be difficult to implement efficiently with modern parallel and distributed computing facilities (Lee et al., 2010; Green et al., 2015; Wilkinson, 2005).

Accepted Article

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/sjos.12492

This article is protected by copyright. All rights reserved.

We promote a simple two-phase inference approach, based on importance sampling (IS), which is well-suited for parallel implementation. It combines a typically low-dimensional MCMC targeting an approximate marginal distribution with independently calculated estimators, which yield exact inference over the full posterior. The estimator is similar to self-normalised importance sampling, but is more general, allowing for sequential Monte Carlo and multilevel type corrections. The method is naturally applicable in a latent variable models context, where the MCMC operates on the hyperparameter distribution using an approximate marginal likelihood, and re-weighting is based on a sampling scheme on the latent variables. We detail the application of the method with Bayesian state space models, where we use importance sampling and particle filters for correction.

Accepted Article

1.1 | Related work

We consider a framework which combines and generalises upon various previously suggested methods, which, to our knowledge, has not been systematically explored before. Importance sampling correction of MCMC has been suggested early in the MCMC literature (e.g. Hastings, 1970; Glynn & Iglehart, 1989; Doss, 1994), and used, for instance, to estimate Bayes factors using a single MCMC output (Doss, 2010). Related confidence intervals have been suggested based on regeneration (Bhattacharya, 2008) and in case of multiple Markov chains (Tan et al., 2015). Using unbiased estimators of importance weights in this context has been suggested at least by Lin et al. (2000) and Lyne et al. (2015), who consider marginal inference with a generalisation of the pseudo-marginal method, allowing for likelihood estimators that may take negative values, and by Quiroz et al. (2016) with data sub-sampling.

Nested or compound sampling has also appeared in many forms in the Monte Carlo literature. The SMC^2 algorithm (Chopin et al., 2013) is based on an application of nested sequential Monte Carlo steps, which has similarities with our framework, and the IS^2 method (Tran et al., 2014) focuses on the case where the preliminary inference is based on independent sampling. We focus on the MCMC approximation of the marginal distribution, which we believe often to be easily implementable in practice, also when the marginal distribution has a non-standard form. The Markov dependence in the marginal Monte Carlo approximation comes with some extra theoretical issues, which we address in detail.

Our setting highlights explicitly the connection of IS type correction and delayed acceptance (DA) (Fox & Nicholls, 1997; Liu, 2003; Christen & Fox, 2005), and recently developed pseudo-marginal type MCMC (Andrieu & Roberts, 2009; Lin et al., 2000) such as particle MCMC (Andrieu et al., 2010), grouped independence Metropolis-Hastings (Beaumont, 2003), approximate Bayesian computation (ABC) MCMC (Marjoram et al., 2003), the algorithm for estimation of discretely observed diffusions suggested by Beskos et al. (2006), and annealed IS (Karagiannis & Andrieu, 2013; Neal, 2001). Theoretical advances of pseudo-marginal methods (Andrieu & Vihola, 2015, 2016; Doucet et al., 2015; Sherlock et al., 2015; Andrieu et al., 2018; Chopin & Singh, 2015; Lindsten et al., 2015) have already led to more efficient implementation of such methods, but have also revealed fundamental limitations. For instance, the methods may suffer from slow (non-geometric) convergence in practically interesting scenarios (Andrieu & Roberts, 2009; Lee & Łatuszynski, 2014). Adding dependence to the estimators (cf. Andrieu & Vihola, 2016), such as using the recently proposed correlated version of the pseudo-marginal MCMC (Deligiannidis et al., 2018), may help in more efficient implementation in certain scenarios, but a successful implementation of such a method may not always be possible, and the question of efficient parallelisability remains a challenge. The blocked parallelisable particle Gibbs (Singh et al., 2017) has appealing limiting properties, but its implementation still requires synchronisation between every update cycle, which may be costly in some computing environments.

The IS approach which we propose may assuage some of the aforementioned challenges of the pseudo-marginal framework, by replacing a pseudo-marginal MCMC with a fast-mixing but approximate MCMC, and postponing the computationally intensive calculations to parallelisable post-processing; see Section 2.3.

This article is protected by copyright. All rights reserved.

1.2 | Outline

We introduce a generic Bayesian latent variable model in Section 2, detail our approach algorithmically, and compare it with DA. We also discuss practical implications, modifications and possible extensions. After introducing notation in Section 3, we formulate general IS type correction of MCMC and related consistency results in Section 4. We detail the general case (Theorem 7), based on a concept (Definition 5), which we call a 'proper weighting' scheme (following the terminology of Liu, 2003), which is natural and convenient in many contexts. In Section 5, we state central limit theorems and expressions for asymptotic variances. Section 6 focuses on estimators which calculate IS correction once for each accepted state, stemming from a so-called 'jump chain' representation. Section 7 details consistency of our estimators in case the approximate chain is pseudo-marginal.

We then focus on state-space models with linear-Gaussian state dynamics in Section 8, and show how a Laplace approximation can be used both for approximate inference, and for construction of efficient proper weighting schemes. Section 9 describes an instance of our approach in the context of discretely observed diffusions, with an approximate pseudo-marginal chain. We compare empirically several algorithmic variations in Section 10 with Poisson observations, with a stochastic volatility model and with a discretely observed geometric Brownian motion. Section 11 concludes, with discussion.

2 | THE PROPOSED LATENT VARIABLE MODEL INFERENCE METHODOLOGY

A generic Bayesian latent variable model is defined in terms of three random vector, and corresponding conditional densities:

- $\Theta \sim \mathrm{pr}(\,\cdot\,)$ — prior density of (hyper)parameters,
- $X \mid \Theta = \theta \sim \mu^{(\theta)}(\,\cdot\,)$ — prior of latent variables given parameters, and
- $Y \mid (\Theta = \theta, X = x) \sim g^{(\theta)}(\,\cdot \mid x)$ — the observation model.

The aim is inference over the posterior of (Θ, X) given observations $Y = y$, with density $\pi(\theta, x) \propto \mathrm{pr}(\theta)\mu^{(\theta)}(x)g^{(\theta)}(y \mid x)$. Standard MCMC algorithms may, in principle, be applied directly for inference, but the typical high dimension of the latent variable x and the common strong dependency structures often lead to poor performance of generic algorithms.

Our inference approach focuses on the specific structure of the model, based on the factorisation $\pi(\theta, x) = \pi_m(\theta)r(x \mid \theta)$, where the marginal posterior density π_m and the corresponding conditional r are:

$$\pi_m(\theta) := \int \pi(\theta, x)\mathrm{d}x \propto \mathrm{pr}(\theta)L(\theta) \qquad \text{and} \qquad r(x \mid \theta) := \frac{p^{(\theta)}(x, y)}{L(\theta)},$$

with the joint density of the latent and the observed $p^{(\theta)}(x, y)$, and the marginal likelihood $L(\theta)$ given as follows:

$$p^{(\theta)}(x, y) := \mu^{(\theta)}(x)g^{(\theta)}(y \mid x) \qquad \text{and} \qquad L(\theta) := \int p^{(\theta)}(x, y)\mathrm{d}x.$$

Two particularly successful latent variable model inference methods, the integrated nested Laplace approximation (INLA) (Rue et al., 2009) and the particle MCMC methods (PMCMC) (Andrieu et al., 2010), rely on this structure. In essence, the INLA is based on an efficient Laplace approximation $p_a^{(\theta)}(x, y)$ of $p^{(\theta)}(x, y)$, determining an approximate marginal likelihood $L_a(\theta)$ and approximate conditional distribution $r_a(x \mid y)$. Particle MCMC uses a specialised SMC algorithm, which provides an unbiased approximation of expectations with respect to $p^{(\theta)}(x, y)$ allowing for exact inference, and which is particularly efficient in the state space models context.

Accepted Article

This article is protected by copyright. All rights reserved.

2.1 | An algorithmic description

The primary aim of this paper is the efficient use of an approximate marginal likelihood $L_a(\theta)$ within a Monte Carlo framework that leads to efficient, parallelisable and exact inference. For instance, Laplace approximations often lead to a natural choice for $L_a(\theta)$. The inference method which we propose comprises two algorithmic phases, which are summarised below:

Phase 1: Simulate a Markov chain $(\Theta_k)_{k=1,\dots,n}$ targeting an approximate hyperparameter posterior $\pi_a(\theta) \propto \mathrm{pr}(\theta)L_a(\theta)$.

Phase 2: For each Θ_k, sample $(V_k^{(1:m)}, X_k^{(1:m)}) := (V_k^{(1)}, \dots, V_k^{(m)}, X_k^{(1)}, \dots, X_k^{(m)})$ where $V_k^{(i)} \in \mathbb{R}$ and $X_k^{(i)}$ are in the latent variable space, and calculate $W_k^{(i)} := V_k^{(i)}/L_a(\Theta_k)$, which determine a weighted estimator

$$E_n(f) := \frac{\sum_{k=1}^n \sum_{i=1}^m W_k^{(i)} f(\Theta_k, X_k^{(i)})}{\sum_{j=1}^n \sum_{\ell=1}^m W_j^{(\ell)}} \tag{1}$$

of the full posterior expectation $\mathbb{E}_\pi[f(\Theta, X)] = \int f(\theta, x)\pi(\theta, x)\mathrm{d}\theta\mathrm{d}x$.

The following conditions are essential to ensure the consistency of the estimator:

C1: The approximation is consistent, in the sense that $L_a(\theta) > 0$ whenever $L(\theta) > 0$, and $\int \mathrm{pr}(\theta)L_a(\theta)\mathrm{d}\theta < \infty$.

C2: The Markov chain $(\Theta_k)_{k\geq n}$ is Harris ergodic (Definition 1) with respect to π_a.

C3: Denoting $f^*(\theta) := \mathbb{E}_\pi[f(\Theta, X) \mid \Theta = \theta] = \int r(x \mid \theta) f(\theta, x)\mathrm{d}x$, there exists a constant $c_w > 0$ such that

$$\mathbb{E}\left[\sum_{i=1}^m V_k^{(i)} f(\Theta_k, X_k^{(i)}) \,\middle|\, \Theta_k = \theta\right] = c_w L(\theta) f^*(\theta), \tag{2}$$

for all $\theta \in \mathsf{T}$, all functions f of interest, and for $f \equiv 1$ (i.e. (2) holds with $f(\cdot)$ and $f^*(\cdot)$ omitted). The value of c_w need not be known.

Both C1 and C2 are easily satisfied by construction of the approximation, and C3 is satisfied by many schemes. Appendix D reviews how the particle filter leads to such schemes. The conditions C1–C3 are key, but unfortunately not enough to guarantee consistency, which requires also a (mild) integrability condition, which $(W_k^{(1:m)}, X_k^{(1:m)})_{k=1,\dots,n}$ must satisfy. Fortunately, in the common case of non-negative $V_k^{(i)}$, this integrability is guaranteed if also $|f|$ satisfies (2); see Section 4 for details. Further conditions ensure a central limit theorem $\sqrt{n}\{E_n(f) - \mathbb{E}_\pi[f(\Theta, X)]\} \to N(0, \sigma^2)$, as detailed in Section 5.

When Phase 1 is a Metropolis-Hastings algorithm, it is possible to generate only one batch of $(\tilde{V}_k^{(1:m)}, \tilde{X}_k^{(1:m)})$ for each *accepted* state $(\tilde{\Theta}_k)$. If N_k stands for the time spent at $\tilde{\Theta}_k$, then the corresponding weights are determined as $\tilde{W}_k := N_k V_k^{(i)}/L_a(\tilde{\Theta}_k)$; see Section 6 for details about such 'jump chain' estimators.

2.2 | Use with approximate pseudo-marginal MCMC

In many scenarios, such as with time-discretised diffusions, the latent variable prior density $\mu^{(\theta)}$ cannot be evaluated, and exact simulation is impossible or very expensive. Simulation is also expensive with a fine enough time-discretisation.

A coarsely discretised model leads to a natural cheap approximation $\hat{\mu}^{(\theta)}$, but in Phase 1, the Markov chain will often be a pseudo-marginal MCMC (cf. Andrieu & Roberts, 2009), in which case our scheme would have the following form:

Phase 1': Simulate a pseudo-marginal Metropolis-Hastings chain (Θ_k, U_k) for $k = 1, \dots, n$, following

Accepted Article

This article is protected by copyright. All rights reserved.

a. Draw a proposal $\tilde{\Theta}_k$ from $q(\Theta_{k-1}, \cdot)$ and given $\tilde{\Theta}_k$, construct an estimator $\tilde{U}_k \geq 0$ such that $\mathbb{E}[\tilde{U}_k \mid \tilde{\Theta}_k = \theta] = L_a(\theta)$.

b. With probability $\min\left\{1, \frac{\mathrm{pr}(\tilde{\Theta}_k)\tilde{U}_k q(\tilde{\Theta}_k, \Theta_{k-1})}{\mathrm{pr}(\Theta_{k-1})U_{k-1} q(\Theta_{k-1}, \tilde{\Theta}_k)}\right\}$, accept and set $(\Theta_k, U_k) = (\tilde{\Theta}_k, \tilde{U}_k)$; otherwise reject the move.

Phase 2': For each (Θ_k, U_k), sample $(V_k^{(1:m)}, X_k^{(1:m)})$ and set $W_k^{(i)} := V_k^{(i)}/U_k$, which determine the estimator as in (1).

Algorithmically, the pseudo-marginal version above is similar to the method in Section 2.1, with the likelihood $L_a(\Theta_k)$ replaced with its estimator U_k. The requirements for the approximate likelihood C1 and its estimator C3 remain identical, and C2 must hold for the pseudo-marginal chain $(\Theta_k, U_k)_{k\geq 1}$, together with the following condition:

C4: The estimators $\tilde{U}_k$ are strictly positive, almost surely, for all $\tilde{\Theta}_k \in \mathsf{T}$.

These are enough to guarantee consistency; see Section 7, and in particular Proposition 26 for details, which also justifies why C4 is needed for consistency. In practice it may be easily satisfied, because the likelihood estimators $\tilde{U}_k$ may be inflated, if necessary (see Section 11).

Note that the variables $(V_k^{(1:m)}, X_k^{(1:m)})$ may depend on both Θ_k and the related likelihood estimate U_k. The dependency may be useful, if positively correlated $V_k^{(i)}$ and U_k are available, leading to lower variance weights $W_k^{(i)} = V_k^{(i)}/U_k$. This is similar to the correlated pseudo-marginal algorithm (Deligiannidis et al., 2018), which relies on a particular form of $V_k^{(i)}$ and U_k. If positively correlated structure is unavailable, $(V_k^{(1:m)}, X_k^{(1:m)})$ can be constructed independently of U_k.

2.3 | Comparison with delayed acceptance

The key condition, under which we believe our method to be useful, is that the Phase 1 Markov chain is computationally relatively cheap compared to construction of the random variables $(W_k^{(1:m)}, X_k^{(1:m)})$ computed in Phase 2. Similar rationale, and similar building blocks — a π_a-reversible Markov chain and random variables analogous to $(W_k^{(i)}, X_k^{(i)})$ — have been suggested earlier for construction of a delayed acceptance (DA) pseudo-marginal MCMC scheme (cf. Golightly et al., 2015). Such an algorithm defines a Markov chain $(\Theta_k, W_k^{(1:m)}, X_k^{(1:m)})_{k\geq 1}$, with one iteration consisting of the following steps:

DA 1: Draw $\tilde{\Theta}_k \sim P(\Theta_{k-1}, \cdot)$. If $\tilde{\Theta}_k = \Theta_{k-1}$ reject and set $(\Theta_k, W_k^{(1:m)}, X_k^{(1:m)}) = (\Theta_{k-1}, W_{k-1}^{(1:m)}, X_{k-1}^{(1:m)})$; otherwise go to (DA 2).

DA 2: Conditional on $\tilde{\Theta}_k$, draw $(\tilde{V}_k^{(1:m)}, \tilde{X}_k^{(1:m)})$ which satisfy (2) with $\tilde{\Theta}_k$ in place of Θ_k, and set $\tilde{W}_k^{(i)} := \tilde{V}_k^{(i)}/L_a(\tilde{\Theta}_k)$. With probability $\min\left\{1, \frac{\sum_{i=1}^m \tilde{W}_k^{(i)}}{\sum_{\ell=1}^m W_{k-1}^{(\ell)}}\right\}$, accept and set $(\Theta_k, W_k^{(1:m)}, X_k^{(1:m)}) = (\tilde{\Theta}_k, \tilde{W}_k^{(1:m)}, \tilde{X}_k^{(1:m)})$; otherwise reject and set $(\Theta_k, W_k^{(1:m)}, X_k^{(1:m)}) = (\Theta_{k-1}, W_{k-1}^{(1:m)}, X_{k-1}^{(1:m)})$

If the pseudo-marginal method is used in DA 1 the value $L_a(\Theta_k)$ is replaced with the related likelihood estimator. Under the same assumptions as required by our scheme, and additionally requiring that $\tilde{W}_k^{(i)} \geq 0$, and that the Markov chain $(\Theta_k, W_k^{(1:m)}, X_k^{(1:m)})_{k\geq 1}$ is Harris, the DA scheme leads to a consistent estimator:

$$\frac{1}{n}\sum_{k=1}^{n}\sum_{i=1}^{m}\left(\frac{W_k^{(i)}}{\sum_{\ell=1}^m W_k^{(\ell)}}\right) f(\Theta_k, X_k^{(i)}) \xrightarrow{n\to\infty} \mathbb{E}_\pi[f(\Theta, X)].$$

Our IS scheme is a natural alternative to such a DA scheme, replacing the independent Metropolis-Hastings type accept-reject step DA 2 with analogous weighting. This relatively small algorithmic change brings many, potentially substantial, benefits over DA, which we note next.

Accepted Article

This article is protected by copyright. All rights reserved.

1. Phase 2 corrections are entirely independent 'post-processing' of Phase 1 MCMC output $(\Theta_k)_{k=1,\dots,n}$, which is easy to implement efficiently using parallel or distributed computing. This is unlike DA 1 and DA 2, which must be iterated sequentially.
2. If Phase 2 correction variables are calculated only once for each accepted Θ_k (so-called 'jump chain' representation, see Section 6), the IS method will typically be computationally less expensive than DA with the same number of iterations, even without parallelisation.
3. The Phase 1 MCMC chain (Θ_k) may be (further) thinned before applying (much more computationally demanding) Phase 2. Thinning of the DA chain is less likely beneficial (cf. Owen, 2017).
4. In case the approximate marginal MCMC (Θ_k) is based on a deterministic likelihood approximation, it is generally 'safer' than (pseudo-marginal) DA using likelihood estimators, because pseudo-marginal MCMC may have issues with mixing (cf. Andrieu & Vihola, 2015). It is also easier to implement efficiently. For instance, popular adaptive MCMC methods which rely on acceptance rate optimisation (Andrieu & Thoms, 2008, and references therein) are directly applicable.
5. Reversibility of the MCMC kernel P in DA 1 is necessary, but not required for the Phase 1 MCMC.
6. The average of the weights, $n^{-1}\sum_{k=1}^{n}\sum_{i=1}^{m} W_k^{(i)}$, provides a consistent estimator of the ratio of the normalising constants of π and π_a, which may be useful in some contexts.
7. Non-negativity of $W_k^{(i)}$ is required in DA 2, but not in Phase 2. This may be useful in certain contexts, where multilevel (Heinrich, 2001; Giles, 2008) or debiasing (McLeish, 2011; Rhee & Glynn, 2015; Vihola, 2017) are applicable. (See also the discussion of Jacob & Thiery (2015) why pseudo-marginal method may not be applicable at all in such a context.)
8. The separation of 'approximate' Phase 1 and 'exact' Phase 2 allows for two-level inference. In statistical practice, preliminary analysis could be based on (fast) purely approximate inference, and the (computationally demanding) exact method could be applied only as a final verification to ensure that the approximation did not affect the findings.

To elaborate the last point, the approximate likelihood $L_a(\theta)$ is usually based on an approximation $p_a^{(\theta)}(x,y)$ of the latent model $p^{(\theta)}(x,y)$. If the approximate model admits tractable expectations of functions f of interest or exact simulation, direct approximate inference is possible, because

$$\frac{1}{n}\sum_{k=1}^{n} f_a^*(\Theta_k) \to \mathbb{E}_{\tilde{\pi}}[f(\Theta,X)], \qquad \text{where} \qquad f_a^*(\theta) := \mathbb{E}_{\tilde{\pi}}[f(\Theta,X) \mid \Theta=\theta],$$

with approximate joint posterior $\tilde{\pi}(\theta,x) \propto \mathrm{pr}(\theta)p_a^{(\theta)}(x,y)$. Then, Phase 2 allows for quantification of the bias $\mathbb{E}_{\tilde{\pi}}[f(\Theta,X)] - \mathbb{E}_{\pi}[f(\Theta,X)]$, and confirmation that both inferences lead to the same conclusions.

Even though IS is likely to bring benefits over DA in many scenarios, there are some situations where DA might perform better. In fact, DA may be used in some scenarios where IS cannot be used at all; namely, P may only be reversible with respect to a positive measure that is not a probability measure, and DA can still be valid. DA may also be more robust with respect to tail behaviour of π_a, where IS needs more care; see the discussion in Section 11. The further theoretical work (Franks & Vihola, to appear) provides upper bounds for ratios of the asymptotic variances of IS and DA in terms of (lower or upper) bounds of the weights $W_k^{(i)}$, and includes examples where DA outperforms IS and vice versa.

3 | NOTATION AND PRELIMINARIES

Throughout the paper, we consider general state spaces while using standard integral notation. If the model at hand is given in terms of standard probability densities, the rest of this paragraph can be skipped. Each space X is assumed to be equipped with a σ-finite dominating measure 'dx' on a σ-algebra denoted with a corresponding calligraphic letter, such as $\mathcal{X}$. Product spaces

Accepted Article

This article is protected by copyright. All rights reserved.

are equipped with the related product σ-algebras and product dominating measures. If X is a subset of an Euclidean space $\mathbb{R}^d$, $\mathrm{d}x$ is taken by default as the Lebesgue measure and $\mathcal{X}$ as the Borel subsets of X. $\mathbb{R}_+$ stands for the non-negative real numbers, and constant unit function is denoted by $\mathbf{1}$.

If ν is a probability density on X, we define the support of ν as $\mathrm{supp}(\nu) := \{x \in \mathsf{X} : \nu(x) > 0\}$, and the probability measure corresponding to ν with the same symbol $\nu(\mathrm{d}x) := \nu(x)\mathrm{d}x$.If $g : \mathsf{X} \to \mathbb{R}$, we denote $\nu(g) := \int g(x)\nu(\mathrm{d}x)$, whenever well-defined. For a probability density or measure ν on X and $p \in [1, \infty)$, we denote by $L^p(\nu)$ the set of measurable $g : \mathsf{X} \to \mathbb{R}$ with $\nu(|g|^p) < \infty$, and by $L^p_0(\nu) := \{g \in L^p(\nu) : \nu(g) = 0\}$ the corresponding set of zero-mean functions. If P is a Markov transition probability, we denote the probability measure $(\nu P)(A) := \int \nu(\mathrm{d}x)P(x, A)$, and the function $(Pg)(x) := \int P(x, \mathrm{d}y)g(y)$. Iterates of transition probabilities are defined recursively through $P^n(x, A) := \int P(x, \mathrm{d}y)P^{n-1}(y, A)$ for $n \geq 1$, where $P^0(y, A) := \mathbb{I}(y \in A)$.

We follow the conventions $0/0 := 0$ and $\mathbb{N} := \{1, 2, \dots\}$. For integers $a \leq b$, we denote by $a{:}b$ the integers within the interval $[a, b]$. We use this notation in indexing, so that $x_{a:b} = (x_a, \dots, x_b)$, $x^{(a:b)} = (x^{(a)}, \dots, x^{(b)})$. If $a > b$, then $x_{a:b}$ or $x^{(a:b)}$ is void, so that for example $g(x, y_{1:0})$ is interpreted as $g(x)$. Similarly, if $i_{1:T}$ is a vector, then $x^{(i_{1:T})} = (x^{(i_1)}, \dots, x^{(i_T)})$ and $x^{(i_{1:T})}_{1:T} = (x^{(i_1)}_1, \dots, x^{(i_T)}_T)$. We also use double-indexing, such as $x_k^{(1:m,1:n)} = (x_k^{(1,1)}, \dots, x_k^{(1,n)}, x_k^{(2,1)}, \dots, x_k^{(m,n)})$.

Throughout the paper, we assume the underlying MCMC scheme to satisfy the following standard condition.

Definition 1 (Harris ergodicity) *A Markov chain is called* Harris ergodic *with respect to ν, if it is ψ-irreducible, Harris recurrent and with invariant probability ν.*

Virtually all MCMC schemes are Harris ergodic (cf. Tierney, 1994; Nummelin, 2002), although in some cases careless implementation could lead to a non-Harris chain (cf. Roberts & Rosenthal, 2006). Thanks to the Harris assumption, all the limit theorems which we give hold for any initial distribution of the related Markov chain.

4 | GENERAL IMPORTANCE SAMPLING TYPE CORRECTION OF MCMC

Hereafter, π_a is a probability density on T and represents an approximation of a probability density π_m of interest. The consistency of IS type correction relies on the following mild assumption.

Assumption 2 *The Markov chain $(\Theta_k)_{k\geq 1}$ and the density π_a satisfy:*

(i) *$(\Theta_k)_{k\geq 1}$ is Harris ergodic with respect to π_a.*
(ii) $\mathrm{supp}(\pi_m) \subset \mathrm{supp}(\pi_a)$.
(iii) *$w_u(\theta) := c_w \pi_m(\theta)/\pi_a(\theta)$, where $c_w > 0$ is a constant.*

If Assumption 2 holds and it is possible to calculate the unnormalised importance weight $w_u(\theta)$ pointwise, the chain $(\Theta_k)_{k\geq 1}$ can be weighted in order to approximate $\pi_m(g)$ for every $g \in L^1(\pi_m)$, using (self-normalised) importance sampling (e.g. Glynn & Iglehart, 1989; Doss, 1994)

$$\frac{\sum_{k=1}^n w_u(\Theta_k)g(\Theta_k)}{\sum_{j=1}^n w_u(\Theta_j)} = \frac{n^{-1}\sum_{k=1}^n w_u(\Theta_k)g(\Theta_k)}{n^{-1}\sum_{j=1}^n w_u(\Theta_j)} \xrightarrow{n\to\infty} \frac{\pi_a(w_u g)}{\pi_a(w_u)} = \pi_m(g) \qquad \text{almost surely,}$$

as Harris ergodicity guarantees the almost sure convergence of both the numerator and the denominator.

In case π_m is a marginal density, which we will focus on, both the ratio $w_u(\theta)$ and the function g (which will be a conditional expectation) are typically intractable. Instead, it is often possible to construct unbiased estimators, which may be used in order

Accepted Article

This article is protected by copyright. All rights reserved.

to estimate the numerator and the denominator, in place of $w_u(\Theta_k)$ and $g(\Theta_k)$, under mild conditions. In order to formalise such a setting, we give the following generic condition for ratio estimators, which resemble the IS correction above.

Assumption 3 *Suppose Assumption 2 holds, and let $(S_k)_{k\geq 1}$, where $S_k = \left(A_k, B_k\right) \in \mathbb{R}^2$, be conditionally independent given $(\Theta_k)_{k\geq 1}$, such that the distribution of S_k depends only on the value of Θ_k, and*

(i) *$f_A(\theta) := \mathbb{E}[A_k \mid \Theta_k = \theta]$ satisfies $\pi_a(f_A) = c_w \pi_m(g)$,*
(ii) *$f_B(\theta) := \mathbb{E}[B_k \mid \Theta_k = \theta]$ satisfies $\pi_a(f_B) = c_w$, and*
(iii) *$\pi_a(m^{(1)}) < \infty$ where $m^{(1)}(\theta) := \mathbb{E}\left[|A_k| + |B_k| \mid \Theta_k = \theta\right]$.*

We record the following simple statement which guarantees consistency under Assumption 3.

Lemma 4 *If Assumption 3 holds for some $g \in L^1(\pi_m)$, then*

$$E_n(g) := \frac{\sum_{k=1}^n A_k}{\sum_{j=1}^n B_j} \xrightarrow{n\to\infty} \pi_m(g) \qquad \textit{almost surely}.$$

The proof of Lemma 4 follows by observing that $(\Theta_k, S_k)_{k\geq 1}$ is Harris ergodic, where $S_k = (A_k, B_k)$, and the functions $h_1(\theta, a, b) = a$ and $h_2(\theta, a, b) = b$ are integrable with respect to its invariant distribution $\breve{\pi}(\mathrm{d}\theta \times \mathrm{d}s) := \pi_a(\mathrm{d}\theta)Q(\theta, \mathrm{d}s)$, where $Q(\theta, A) := \mathbb{P}(S_k \in A \mid \Theta_k = \theta)$; see Lemma 30 in Appendix A.

In the latent variable model discussed in Section 2, the aim is inference over a joint target density $\pi(\theta, x) := \pi_m(\theta) r(x \mid \theta)$ on an extended state space $\mathsf{T} \times \mathsf{X}$. For every function $f \in L^1(\pi)$, we denote by $f^*(\theta) := \int r(x \mid \theta) f(\theta, x)\mathrm{d}x$ the conditional expectation of f given θ, so $\pi(f) = \pi_m(f^*)$. The following formalises a scheme which satisfies Assumption 3 with $g = f^*$ and therefore guarantees consistency for a class of functions $f \in \mathcal{L} \subset L^1(\pi)$.

Definition 5 ($\mathcal{L}$-Proper weighting scheme) *Suppose Assumption 2 holds, and let $(P_k)_{k\geq 1}$ be conditionally independent given $(\Theta_k)_{k\geq 1}$, such that the distribution of each $P_k = (M_k, W_k^{(1:M_k)}, X_k^{(1:M_k)})$ depends only on the value of Θ_k, where $M_k \in \mathbb{N}$, $W_k^{(i)} \in \mathbb{R}$ and $X_k^{(i)} \in \mathsf{X}$. Define for any $f \in L^1(\pi)$,*

$$\xi_k(f) := \sum_{i=1}^{M_k} W_k^{(i)} f(\Theta_k, X_k^{(i)}).$$

Let $\mathcal{L} \subset L^1(\pi)$ be all the functions for which

(i) *$\mu_f(\theta) := \mathbb{E}[\xi_k(f) \mid \Theta_k = \theta]$ satisfies $\pi_a(\mu_f) = c_w \pi(f)$, and*
(ii) *$\pi_a(m_f^{(1)}) < \infty$ where $m_f^{(1)}(\theta) := \mathbb{E}\left[|\xi_k(f)| \mid \Theta_k = \theta\right]$.*

*If $\mathbf{1} \in \mathcal{L}$, then $(W_k^{(1:M_k)}, X_k^{(1:M_k)})_{k\geq 1}$ or equivalently $(\xi_k)_{k\geq 1}$, form a $\mathcal{L}$-*proper weighting scheme.

Remark 6 *Regarding Definition 5:*

(i) *In case of non-negative weights, that is, $W_k^{(i)} \geq 0$ almost surely, we have $|\xi_k(\mathbf{1})| = \xi_k(\mathbf{1})$, so $f \equiv \mathbf{1} \in \mathcal{L}$ if and only if (i) holds for $f \equiv \mathbf{1}$. Further, if (i) holds for both f and $|f|$, then (ii) holds, because $|\xi_k(f)| \leq \xi_k(|f|)$.*
(ii) *When certain multilevel (Heinrich, 2001; Giles, 2008) or debiasing methods (cf. McLeish, 2011; Rhee & Glynn, 2015; Glynn & Rhee, 2014) are applied, $W_k^{(i)}$ generally take also negative values. In such a case, an extra integrability condition is necessary, and we believe (ii) is required for consistency in general.*

Accepted Article

This article is protected by copyright. All rights reserved.

(iii) *Note that $\mathcal{L}$ is closed under linear operations, that is, if $a, b \in \mathbb{R}$ and $f, g \in \mathcal{L}$, then $af + bg \in \mathcal{L}$. This, together with $\mathcal{L}$ containing constant functions, implies that if $f \in \mathcal{L}$, then $\bar{f} := f - \pi(f) \in \mathcal{L}$.*

(iv) *In fact, ξ_k may be interpreted as a* random (signed) measure $\xi_k(\mathrm{d}\theta, \mathrm{d}x) = \sum_{i=1}^{M_k} W_k^{(i)} \delta_{\Theta_k}(\mathrm{d}\theta)\delta_{X_k^{(i)}}(\mathrm{d}x)$. *Our results extend also to a generalisation, where $\delta_{X_k^{(i)}}(\mathrm{d}x)$ is replaced by another measure, as long as (i) and (ii) hold. For instance, in the context of Rao-Blackwellisation, we could have Gaussian distributions in place of $\delta_{X_k^{(i)}}(\mathrm{d}x)$.*

The following consistency result is a direct consequence of Lemma 4.

Theorem 7 *If $(\xi_k)_{k\geq 1}$ form a $\mathcal{L}$-proper weighting scheme, then the IS type estimator is consistent, that is,*

$$E_n(f) := \frac{\sum_{k=1}^n \xi_k(f)}{\sum_{j=1}^n \xi_j(\mathbf{1})} \xrightarrow{n\to\infty} \pi(f), \qquad \textit{almost surely}. \tag{3}$$

Let us next exemplify a 'canonical' setting of a proper weighting scheme, stemming from standard unnormalised importance sampling.

Proposition 8 *Suppose Assumption 2 holds and $q^{(\theta)}(\cdot)$ defines a probability density on X for each $\theta \in \mathsf{T}$ and* $\mathrm{supp}(\pi) \subset \{(\theta, x) : \pi_a(\theta)q^{(\theta)}(x) > 0\}$. *Let*

$$X_k^{(1:m)} \overset{\text{i.i.d.}}{\sim} q^{(\Theta_k)}, \quad V_k^{(i)} := \frac{1}{m} \cdot \frac{c_w \pi(\Theta_k, X_k^{(i)})}{q^{(\Theta_k)}(X_k^{(i)})} \quad \textit{and} \quad W_k^{(i)} := \frac{V_k^{(i)}}{\pi_a(\Theta_k)},$$

where $c_w > 0$ a constant. Then, $(W_k^{(1:m)}, X_k^{(1:m)})_{k\geq 1}$ form a $L^1(\pi)$-proper weighting scheme.

When the weights are all positive, sub-sampling may be used in order to save memory.

Proposition 9 *Suppose that $(W_k^{(1:M_k)}, X_k^{(1:M_k)})_{k\geq 1}$ forms a $\mathcal{L}$-proper weighting scheme with non-negative $W_k^{(1:M_k)} \geq 0$ (a.s.). Let $W_k := \sum_{i=1}^{M_k} W_k^{(i)}$ and let (I_k) be random variables conditionally independent of $(\Theta_k, X_k^{(i)})$ such that $\mathbb{P}(I_k = i) = W_k^{(i)}/W_k$ (and let $I_k = 1$ if $W_k = 0$). Then, $(W_k, X_k^{(I_k)})_{k\geq 1}$ forms a $\mathcal{L}$-proper weighting scheme.*

The sub-sampling estimator simplifies to

$$E_n(f) = \frac{\sum_{k=1}^n W_k f(\Theta_k, X_k^{(I_k)})}{\sum_{k=1}^n W_k}.$$

We conclude by a complementary statement about convex combinations of multiple proper sampling schemes.

Proposition 10 *Suppose $(\xi_{k,j})_{k\geq 1}$ forms a $\mathcal{L}$-proper weighting scheme for each $j \in \{1:N\}$, then, for any constants $\beta_1, \dots, \beta_N \geq 0$ with $\sum_{j=1}^N \beta_j = 1$, the convex combinations $\xi_k(f) := \sum_{j=1}^N \beta_j \xi_{k,j}(f)$ form a $\mathcal{L}$-proper sampling scheme.*

5 | ASYMPTOTIC VARIANCE AND A CENTRAL LIMIT THEOREM

The asymptotic variance is a common efficiency measure for Markov chain estimators, because it coincides with the limiting variance of a central limit theorem (CLT), under general conditions (Kipnis & Varadhan, 1986; Maxwell & Woodroofe, 2000).

Accepted Article

This article is protected by copyright. All rights reserved.

Definition 11 *Suppose the Markov chain $(\Theta_k)_{k\geq 1}$ on T has transition probability P which is Harris ergodic with respect to invariant probability π_a. For $f \in L^2(\pi_a)$, the asymptotic variance of f with respect to P is*

$$\mathrm{Var}(f, P) := \lim_{n\to\infty} \mathbb{E}\left(\frac{1}{\sqrt{n}} \sum_{k=1}^{n} \left[f(\Theta_k^{(s)}) - \pi_a(f) \right] \right)^2,$$

whenever the limit exists in $[0, \infty]$, where $(\Theta_k^{(s)})_{k\geq 1}$ stands for the stationary Markov chain *with transition probability P, that is, with $\Theta_1^{(s)} \sim \pi_a$.*

In what follows, we denote by $\bar{f}(\theta, x) = f(\theta, x) - \pi(f)$ the centred version of any $f \in L^1(\pi)$, and recall that if $f \in \mathcal{L}$, then $\bar{f} \in \mathcal{L}$. We also denote $m_f^{(2)}(\theta) := \mathbb{E}[|\xi_k(f)|^2 \mid \Theta_k = \theta]$ for any $f \in \mathcal{L}$. The proof of the following CLT is given in Appendix B.

Theorem 12 *Suppose that the conditions of Theorem 7 are satisfied, and $(\Theta_k)_{k\geq 1}$ is aperiodic. Let $f \in \mathcal{L} \cap L^2(\pi)$ and denote $\bar{f}(\theta, x) := f(\theta, x) - \pi(f)$. If $\pi_a(m_{\bar{f}}^{(2)}) < \infty$ and either of the following hold:*

(i) *$(\Theta_k)_{k\geq 1}$ is reversible and $\mathrm{Var}(\mu_{\bar{f}}, P) < \infty$, or*
(ii) *$\sum_{n=1}^{\infty} n^{-3/2} \{\pi_m([\sum_{k=0}^{n-1} P^k \mu_{\bar{f}}]^2)\}^{1/2} < \infty$,*

then, the estimator $E_n(f)$ defined in (3) *satisfies a CLT:*

$$\sqrt{n}[E_n(f) - \pi(f)] \xrightarrow{n\to\infty} N\left(0, \sigma_f^2\right), \qquad \textit{in distribution, where} \qquad \sigma_f^2 := \frac{\mathrm{Var}(\mu_{\bar{f}}, P) + \pi_a(v)}{c_w^2}, \tag{4}$$

and where $v(\theta) := \mathrm{Var}\left(\xi_k(\bar{f}) \mid \Theta_k = \theta\right)$.

Remark 13 *In case of reversible chains, the condition in Theorem 12 (i) is essentially optimal, and the CLT relies on a result due to Kipnis & Varadhan (1986). The condition always holds when $(\Theta_k)_{k\geq 1}$ is geometrically ergodic, for instance $(\Theta_k)_{k\geq 1}$ is a random-walk Metropolis algorithm and π_a is light-tailed (Jarner & Hansen, 2000; Roberts & Tweedie, 1996). In case $(\Theta_k)_{k\geq 1}$ is sub-geometric, such as polynomial, extra conditions are required; see for instance (Jarner & Roberts, 2007). The condition (ii) applies for non-reversible chains, and relies on a result due to Maxwell & Woodroofe (2000). If there exists $g \in L^2(\pi_a)$ which solves the Poisson equation $g - Pg = \mu_{\bar{f}}$, then (ii) holds, but this is not necessary. See also the review on Markov chain CLTs by Jones (2004).*

When Theorem 12 holds, we recall how a consistent confidence interval may be constructed.

Corollary 14 *Suppose that $\hat{a}_n$ is an estimator of the integrated autocovariance of the sequence $(\xi_1(f), \ldots, \xi_n(f))$. If the conditions of Theorem 12 hold and $\hat{a}_n$ is consistent (see below), then for any $z_q > 0$,*

$$\mathbb{P}\left(\pi(f) \in \left[\mathbb{E}_n(f) \pm z_q \left(\frac{\hat{s}_n}{n} \right)^{1/2} \right] \right) \xrightarrow{n\to\infty} 1 - 2\Phi(z_q), \qquad \textit{where} \qquad \hat{s}_n := \frac{\hat{a}_n}{\left(n^{-1} \sum_{k=1}^{n} \xi_k(1)\right)^2}$$

where Φ is the standard Gaussian distribution function.

The proof of Corollary 14 is given in Appendix B.

Accepted Article

This article is protected by copyright. All rights reserved.

By consistency of $\hat{a}_n$ we mean that it converges to $\gamma_0 + 2\sum_{k\geq 1}\gamma_k$ in probability, which is assumed to exist and be finite, where $(\gamma_k)_{k\geq 1}$ is the stationary lag-k autocovariance of the sequence $(\xi_k(f))_{k\geq 1}$. We refer the reader to consult, for instance, (Flegal & Jones, 2010) for details about consistent integrated autocovariance estimators.

Note that the latter term $\pi_a(v)$ in (4) contains the contribution of the 'noise' in the IS estimates. If the estimators $\xi_k(f)$ are made increasingly accurate, in the sense that $\pi_a(v)$ becomes negligible, the limiting case corresponds to an IS corrected approximate MCMC and calculating averages over conditional expectations $\mu_{\bar{f}}(\theta)$. We conclude by relating the asymptotic variance with a straightforward estimator.

Theorem 15 *Suppose $f \in \mathcal{L} \cap L^2(\pi)$ and $\pi_a(v) < \infty$ where v is defined in Theorem 12, and also $\pi_a(m_{\mathbf{1}}^{(2)}) < \infty$. Then, the estimator*

$$\hat{v}_n := \frac{\sum_{k=1}^n \left(\xi_k(f) - \xi_k(\mathbf{1})E_n(f)\right)^2}{\left(\sum_{j=1}^n \xi_j(\mathbf{1})\right)^2}$$

satisfies $n\hat{v}_n \to \pi_a(v + \mu_{\bar{f}}^2)/c_w^2$ almost surely as $n \to \infty$.

Proof of Theorem 15 is given in Appendix B.

Remark 16 *When P corresponds to i.i.d. sampling from π_a, the estimator $n\hat{v}_n$ in Theorem 15 provides a consistent estimate for the CLT variance $\sigma_{\bar{f}}^2$. In most practical cases, $\mathrm{Var}(\mu_{\bar{f}}, P) \geq \pi_a(\mu_{\bar{f}}^2)$ (which is always true when P is a positive operator (cf. Andrieu & Vihola, 2015)), and then $n\hat{v}_n$ provides an empirical lower bound of the asymptotic variance. Furthermore, it may be useful to inspect the 'decomposition' of the asymptotic variance into $n\hat{v}_n$ and the residual $\hat{s}_n - n\hat{v}_n$. The former may be regarded as the 'independent IS variance' and the residual may be interpreted as 'excess marginal MCMC variance,' as it converges to $2c_w^{-2}\pi_a(\mu_{\bar{f}}^2)\sum_{k\geq 1}\rho_k$, where ρ_k are the stationary autocorrelations of $(\xi_k(f))_{k\geq 1}$. If the residual term is small relative to $\hat{s}_n$, this suggests that either $\mu_{\bar{f}}^2 \ll v$, and/or that ρ_k are small. These inspections could provide insight to choosing the parameters of the underlying marginal MCMC and the proper weighting.*

6 | JUMP CHAIN ESTIMATORS

Many MCMC algorithms such as the Metropolis-Hastings include an accept-reject mechanism, which results in blocks of repeated values $\Theta_k = \ldots = \Theta_{k+b}$. In the context of IS type correction, and when the computational cost of each estimate ξ_k is high, it may be desirable to construct only one estimator per each *accepted* state. To formalise such an algorithm we consider the 'jump chain' representation of the approximate marginal chain (cf. Douc & Robert, 2011; Doucet et al., 2015; Deligiannidis & Lee, 2018).

Definition 17 (Jump chain) *Suppose that $(\Theta_k)_{k\geq 1}$ is Harris ergodic with respect to π_a. The corresponding jump chain $(\tilde{\Theta}_k)_{k\geq 1}$ with holding times $(N_k)_{k\geq 1}$ is defined as follows:*

$$\tilde{\Theta}_k := \Theta_{\bar{N}_{k-1}+1} \qquad \textit{and} \qquad N_k := \inf\left\{j \geq 1 \,:\, \Theta_{\bar{N}_{k-1}+j+1} \neq \tilde{\Theta}_k\right\},$$

where $\bar{N}_k := \sum_{j=1}^k N_j$, and with $\bar{N}_0 \equiv 0$.

Remark 18 *If $(\Theta_k)_{k\geq 1}$ corresponds to a Metropolis-Hastings chain, with non-diagonal proposal distributions q (that is, $q(\theta, \{\theta\}) = 0$ for every $\theta \in \mathsf{T}$), then the jump chain $(\tilde{\Theta}_k)$ consists of the accepted states, and $N_k - 1$ is the number of rejections occurred at state $(\tilde{\Theta}_k)$.*

Accepted Article

This article is protected by copyright. All rights reserved.

Hereafter, we denote by $\alpha(\theta) := \mathbb{P}(\Theta_{k+1} \neq \Theta_k \mid \Theta_k = \theta)$ the overall acceptance probability at θ. We consider next the practically important 'jump IS' estimator, involving a proper weighting for each accepted state.

Assumption 19 *Suppose that Assumption 2 holds, and let $(\tilde{\Theta}_k, N_k)_{k\geq 1}$ denote the corresponding jump chain (Definition 17). Let $(\xi_k)_{k\geq 1}$ be a $\mathcal{L}$-proper weighting scheme, where the variables $(M_k, W_k^{(1:M_k)}, X_k^{(1:M_k)})$ in the scheme are now allowed to depend on both $\tilde{\Theta}_k$ and N_k, and the conditions (i) and (ii). in Definition 5 are replaced with the following:*

(i) $\mathbb{E}[\xi_k(f) \mid \Theta_k = \theta, N_k = n] = \mu_f(\theta)$ *for all* $n \in \mathbb{N}$ *and* $\pi_a(\mu_f) = c_w \pi(f)$,
(ii) $\pi_a(\bar{m}^{(1)}) < \infty$ *where* $\bar{m}^{(1)}(\theta) := \sup_{n\in\mathbb{N}} \mathbb{E}\big[|\xi_k(f)| \mid \Theta_k = \theta, N_k = n\big]$.

Theorem 20 *Suppose Assumption 19 holds, then,*

$$E_n(f) := \frac{\sum_{k=1}^n N_k \xi_k(f)}{\sum_{j=1}^n N_j \xi_j(\mathbf{1})} \xrightarrow{n\to\infty} \pi(f) \qquad \textit{almost surely}. \tag{5}$$

The proof follows from Lemma 4 because $(\tilde{\Theta}_k)$ is Harris ergodic with invariant probability $\tilde{\pi}_a(\theta) \propto \pi_a(\theta)\alpha(\theta)$; see Proposition 33 in Appendix C. Furthermore, the holding times $N_k \geq 0$ are, conditional on $(\tilde{\Theta}_k)$, independent geometric random variables with parameter $\alpha(\tilde{\Theta}_k)$ (Proposition 33), and therefore $\mathbb{E}[N_k \mid \tilde{\Theta}_k = \theta] = 1/\alpha(\theta)$.

Remark 21 *Regarding Assumption 19:*

(i) *Condition (ii) in Assumption 19 is practically convenient, because ξ_k are usually chosen either as independent of N_k, or increasingly accurate in N_k (often taking M_k proportional to N_k); see the discussion below. However, (ii) is not optimal: it is not hard to find examples where the estimator is strongly consistent, even though $\bar{m}^{(1)}(\theta) = \infty$ for some $\theta \in \mathsf{T}$.*
(ii) *In case each ξ_k is constructed as a mean of independent $(\xi_{k,1}, \dots, \xi_{k,N_k})$ (cf. Proposition 10), the jump chain estimator coincides with the simple estimator discussed in Section 5 (at jump times). However, the jump chain estimator offers more flexibility, which may allow for variance reduction, for instance by using a single mN_k particle filter (cf. Appendix D) instead of an average of N_k independent m-particle filters, or by stratification or control variates.*
(iii) *Even though we believe that the estimators of the form* (5) *are often appropriate, we note that in some cases Rao-Blackwellised lower-variance estimators of $1/\alpha(\tilde{\Theta}_j)$ may be used instead of N_k, as suggested by Douc & Robert (2011).*

Let us finally consider a central limit theorem corresponding to the estimator in Theorem 20, whose proof is given in Appendix C.

Theorem 22 *Suppose Assumption 19 holds, $(\tilde{\Theta}_k)_{k\geq 1}$ is aperiodic, $f \in \mathcal{L} \cap L^2(\pi)$,*

$$\pi_a\big(\alpha \tilde{m}^{(2)}\big) < \infty, \qquad \textit{where} \qquad \tilde{m}^{(2)}(\theta) := \mathbb{E}\big[N_k^2 |\xi_k(\bar{f})|^2 \mid \tilde{\Theta}_k = \theta\big], \tag{6}$$

and one of the following holds:

(i) $(\Theta_k)_{k\geq 1}$ *is reversible and* $\mathrm{Var}(\mu_{\bar{f}}, P) < \infty$.
(ii) *There exists* $g \in L^2(\pi_a)$ *satisfying the Poisson equation* $g - Pg = \mu_{\bar{f}}$.

Accepted Article

This article is protected by copyright. All rights reserved.

Then, the estimator $E_n(f)$ in (5) satisfies

$$\sqrt{n}\big[E_n(f) - \pi(f)\big] \xrightarrow{n\to\infty} N(0,\sigma^2) \qquad \textit{in distribution, where} \qquad \sigma^2 = \frac{\pi_a(\alpha)}{c_w^2}\Big[\mathrm{Var}\big(\mu_{\bar f}, P\big) + \pi_a(\alpha\tilde v)\Big],$$

and where $\tilde v(\theta) := \mathbb{E}\big[N_k^2 \mathrm{Var}\big(\xi_k(\bar f) \mid \tilde\Theta_k = \theta, N_k\big) \mid \tilde\Theta_k = \theta\big]$.

Let us briefly discuss the conditions and implications of Theorem 22 under certain specific cases. When the acceptance probability is bounded from below, $\inf_\theta \alpha(\theta) > 0$, using a proper weighting ξ_k independent of N_k is 'safe', because

$$\tilde v(\theta) \le \tilde m^{(2)}(\theta) \le \frac{2-\alpha(\theta)}{\alpha^2(\theta)} b(\theta); \quad b(\theta) := \sup_{n\ge 1} \mathbb{E}\big[|\xi_k(\bar f)|^2 \mid \tilde\Theta_k = \theta, N_k = n\big],$$

and so $\pi_a(b) < \infty$ guarantees (6). For example, if $(\Theta_k)_{k\ge1}$ is L^2-geometrically ergodic, then the acceptance probability is (essentially) bounded away from zero (Roberts & Tweedie, 1996), and $g := \sum_{k\ge0} P^k \mu_{\bar f} \in L^2(\pi_a)$ satisfies $g - Pg = \mu_{\bar f}$, so that (ii) is satisfied.

When ξ_k corresponds to an average of i.i.d. $\xi_{k,1}, \ldots, \xi_{k,N_k}$ (cf. Proposition 10) which do not depend on N_k,

$$\mathrm{Var}\big(\xi_k(\bar f) \mid \tilde\Theta_k = \theta, N_k\big) = \hat v(\theta)/N_k; \quad \hat v(\theta) := \mathrm{Var}\big(\xi_{k,1}(\bar f) \mid \tilde\Theta_k = \theta\big).$$

Then, $\pi_a(\alpha\tilde v) = \pi_a(\hat v)$, which leads to an asymptotic variance that coincides with simple IS correction (cf. Theorem 12).

Remark 23 *Our condition in Theorem 22 (ii), requires the solution to the Poisson equation, and is therefore more stringent than the Maxwell-Woodroofe condition in Theorem 12 (ii). We believe that the result holds more generally, but this would require knowing properties of the jump chain $(\tilde\Theta_k)_{k\ge1}$, which may be unavailable. Theorem 12 (ii) may be applied instead, if such properties are available.*

7 | PSEUDO-MARGINAL APPROXIMATE CHAIN

We next discuss how our limiting results still apply, in case the approximate chain is a pseudo-marginal MCMC, as discussed in Section 2.2. Let us first formalise a pseudo-marginal Markov chain $(\boldsymbol\Theta_k^\circ)_{k\ge1}$, where $\boldsymbol\Theta_k^\circ = (\Theta_k, \Phi_k)$ takes values on $\mathsf{T}\times\mathsf{S}_\Phi$. Let $\boldsymbol\Theta_0^\circ \in \mathsf{T}\times\mathsf{S}_\Phi$ such that $U(\Phi_0) > 0$, and for $k \ge 1$, iterate

PM1 Generate $\tilde\Theta_k \sim q(\Theta_{k-1}, \cdot)$ and $\tilde\Phi_k \sim Q_a(\tilde\Theta_k, \cdot)$.
PM2 With probability $\min\big\{1, \frac{U(\tilde\Phi_k)q(\tilde\Theta_k,\Theta_{k-1})}{U(\Phi_{k-1})q(\Theta_{k-1},\tilde\Theta_k)}\big\}$, accept and set $(\Theta_k,\Phi_k) = (\tilde\Theta_k, \tilde\Phi_k)$; otherwise reject and set $(\Theta_k,\Phi_k) = (\Theta_{k-1},\Phi_{k-1})$.

Above, $Q_a(\theta, \cdot)$ defines a (regular conditional) distribution on (a measurable space) S_Φ, and $U : \mathsf{S}_\Phi \to \mathbb{R}_+$ is a (measurable) function. Under the following condition, the pseudo-marginal Markov chain $(\boldsymbol\Theta_k^\circ)_{k\ge1}$ is reversible with respect to the probability measure $\pi_a^\circ(\mathrm{d}\theta, \mathrm{d}\phi) := \mathrm{d}\theta Q_a(\theta, \mathrm{d}\phi) U(\phi)/c_a$, which admits the marginal $\pi_a(\theta)$ (e.g. Andrieu & Vihola, 2015):

Assumption 24 *There exists a constant $c_a > 0$ such that for each θ, the random variable $\Phi_\theta \sim Q_a(\theta, \cdot)$ satisfies $\mathbb{E}[U(\Phi_\theta)] = c_a \pi_a(\theta)$.*

In addition, $(\boldsymbol\Theta_k^\circ)_{k\ge1}$ is easily shown to be Harris ergodic under minimal conditions.

Accepted Article

This article is protected by copyright. All rights reserved.

Hereafter, when we refer to the results in Sections 4–6, we take the chain $(\boldsymbol{\Theta}_k^\circ)_{k\geq 1}$ as the approximate chain (in place of $(\Theta_k)_{k\geq 1}$), with approximate marginal distribution $\pi_a^\circ(\mathrm{d}\theta, \mathrm{d}\phi)$ (in place of $\pi_a(\theta)\mathrm{d}\theta$). The following abstract minimal condition ensures consistency of an IS type estimator. We discuss practically relevant sufficient conditions later in Proposition 28.

Assumption 25 *Suppose Assumption 2 holds, $(\boldsymbol{\Theta}_k^\circ)_{k\geq 1}$ is Harris ergodic, $c_m > 0$ is a constant, and let $(P_k)_{k\geq 1}$ be conditionally independent given $(\boldsymbol{\Theta}_k^\circ)_{k\geq 1}$, such that the distribution of each $P_k = (M_k, V_k^{(1:M_k)}, X_k^{(1:M_k)})$ depends only on $\boldsymbol{\Theta}_k^\circ$, where $M_k \in \mathbb{N}$, $V_k^{(i)} \in \mathbb{R}$ and $X_k^{(i)} \in \mathsf{X}$. Define for any $f \in L^1(\pi)$, $\zeta_k(f) := \sum_{i=1}^{M_k} V_k^{(i)} f(\Theta_k, X_k^{(i)})$, and let $\mathcal{L} \subset L^1(\pi)$ stand for all the functions for which*

(i) $\iint Q_a(\theta, \mathrm{d}\phi)\mathbb{I}\,(U(\phi) > 0)\,\mathbb{E}[\zeta_k(f) \mid \Theta_k = \theta, \Phi_k = \phi]\mathrm{d}\theta = c_m \pi(f)$, *and*
(ii) $\iint Q_a(\theta, \mathrm{d}\phi)\mathbb{I}\,(U(\phi) > 0)\,\mathbb{E}\big[|\zeta_k(f)| \mid \Theta_k = \theta, \Phi_k = \phi\big]\mathrm{d}\theta < \infty$.

Proposition 26 *Suppose Assumption 24 and 25 hold, and $\mathbf{1} \in \mathcal{L}$. Then, Theorem 7 holds with*

$$\xi_k(f) := \sum_{i=1}^{M_k} W_k^{(i)} f(\Theta_k, X_k^{(i)}) \qquad \textit{where} \qquad W_k^{(i)} = \frac{V_k^{(i)}}{U(\Phi_k)}.$$

The proof of Proposition 26 follows by noting a proper weighting scheme involving the augmented approximate marginal distribution π_a° and target distribution π° (Lemma 27), and Theorem 7.

Lemma 27 *Suppose the conditions of Proposition 26 hold. Then, ξ_k form a $\mathcal{L}^\circ$-proper weighting scheme, with $\mathcal{L}^\circ := \{f^\circ(\theta, \phi, x) = f(\theta, x) \,:\, f \in \mathcal{L}\}$, in the sense of Proposition 5, corresponding to*

(i) *approximate marginal $\pi_a^\circ(\mathrm{d}\theta, \mathrm{d}\phi) = \mathrm{d}\theta Q_a(\theta, \mathrm{d}\phi)U(\phi)/c_a$,*
(ii) *target $\pi^\circ\big((\mathrm{d}\theta, \mathrm{d}\phi), \mathrm{d}x\big)$ which admits the marginal $\pi(\theta, x)\mathrm{d}\theta\mathrm{d}x$.*

Proof For any $f^\circ \in L^\circ$ and $\phi \in \mathsf{S}_\Phi$, let $\nu_f(\theta, \phi) := \mathbb{E}[\zeta_k(f) \mid \Theta_k = \theta, \Phi_k = \phi]$. Whenever $U(\phi) > 0$, define

$$\mu_{f^\circ}^\circ(\theta, \phi) := \mathbb{E}[\xi_k(f^\circ) \mid \Theta_k = \theta, \Phi_k = \phi] = \nu_f(\theta, \phi)/U(\phi),$$

and $\mu_{f^\circ}^\circ(\theta, \phi) := 0$ otherwise. We have

$$\pi_a^\circ(\mu_{f^\circ}^\circ) = c_a^{-1} \iint Q_a(\theta, \mathrm{d}\phi)\mathbb{I}\,(U(\phi) > 0)\,\nu_f(\theta, \phi)\mathrm{d}\theta = c_w \pi(f),$$

by Assumption 25 (i), where $c_w = c_m/c_a$. We also have

$$m_{f^\circ}^{\circ(1)}(\theta, \phi) := \mathbb{E}[\xi_k(f^\circ) \mid \Theta_k = \theta, \phi_k = \phi] = |\nu_f(\theta, \phi)|/U(\phi),$$

so $\pi_a^\circ(m_{f^\circ}^{\circ(1)}) < \infty$ by Assumption 25 (ii).

Let us finally consider different conditions, which guarantee Assumption 25 (i); the integrability Assumption 25 (ii) may be shown similarly.

Proposition 28 *Assumption 25 (i) holds if one of the following hold:*

Accepted Article

This article is protected by copyright. All rights reserved.

(i) *For π_a-a.e. $\theta \in \mathsf{T}$, $U(\Phi_\theta) > 0$ a.s. and*

$$\mathbb{E}[\zeta_k(f) \mid \Theta_k = \theta] = c_m \pi_m(\theta) f^*(\theta), \tag{7}$$

where $\mathbb{E}[\zeta_k(f) \mid \Theta_k = \theta] = \int Q_a(\theta, \mathrm{d}\phi)\mathbb{E}[\zeta_k(f) \mid \Theta_k = \theta, \Phi_k = \phi]$.

(ii) *ζ_k only depend on Θ_k, and for π_a-a.e. $\theta \in \mathsf{T}$,*

$$\mathbb{E}[\zeta_k(f) \mid \Theta_k = \theta] = c_m \pi_m(\theta) f^*(\theta)/p(\theta),$$

where $p(\theta) := \mathbb{P}(U(\Phi_\theta) > 0)$ with $\Phi_\theta \sim Q_a(\theta, \cdot)$.

(iii) *For π_a-a.e. $\theta \in \mathsf{T}$ (7) holds, and $U(\phi) = 0$ implies $\mathbb{E}[\zeta_k(f) \mid \Theta_k = \theta, \Phi_k = \phi] = 0$.*

Proof Note that (i) implies (iii), under which

$$\iint Q_a(\theta, \mathrm{d}\phi)\mathbb{I}(U(\phi) > 0)\, v_f(\theta, \phi)\mathrm{d}\theta = c_m \int \pi_m(\theta) f^*(\theta)\mathrm{d}\theta = c_m \pi(f),$$

where $v_f(\theta, \phi) = \mathbb{E}[\zeta_k(f) \mid \Theta_k = \theta, \Phi_k = \phi]$.

In case of (ii), we have $v_f(\theta, \phi) = \mathbb{E}[\zeta_k(f) \mid \Theta_k = \theta]$ and so

$$\int Q_a(\theta, \mathrm{d}\phi)\mathbb{I}(U(\phi) > 0)\, v_f(\theta, \phi) = c_m \pi_m(\theta) f^*(\theta).$$

Remark 29 *Proposition 28 (i) is the most straightforward in the latent variable context, and often sufficient, since we may choose a positive $U(\phi)$ (e.g. by considering inflated $\tilde{U}(\phi) = U(\phi) + \epsilon$ instead). Proposition 28 (ii) may be used directly to verify the validity of an MCMC version of the lazy ABC algorithm (Prangle, 2016). It also demonstrates why positivity plays a key role: if only (7) is assumed and $p(\theta)$ is non-constant, then $p(\theta)$ must be accounted for, or else we end up with biased estimators targeting a marginal proportional to $\pi_m(\theta)p(\theta)$. Proposition 28 (iii) demonstrates that strict positivity is not necessary, but in this case a delicate dependency structure is required.*

8 | STATE SPACE MODELS AND LINEAR-GAUSSIAN STATE DYNAMICS

State space models (SSM) are latent variable models which are commonly used in time series analysis (cf. Cappé et al., 2005). In the setting of Section 2, SSMs are parametrised by $\theta \in \mathsf{T}$, and $x = z_{1:T} \in \mathsf{X} = \mathsf{S}_z^T$ and $y = y_{1:T} \in \mathsf{Y} = \mathsf{S}_y^T$, and

$$\mu^{(\theta)}(x) = \prod_{t=1}^{T} \mu_t^{(\theta)}(z_t \mid z_{t-1}) \qquad \text{and} \qquad g^{(\theta)}(y \mid x) = \prod_{t=1}^{T} g_t^{(\theta)}(y_t \mid z_t),$$

where, by convention, $\mu_1^{(\theta)}(z_1 \mid z_0) := \mu_1^{(\theta)}(z_1)$. That is, the latent states $Z_{1:T}$ form a Markov chain with initial density $\mu_1^{(\theta)}$ and state transition densities $\mu_t^{(\theta)}$, and $g_t^{(\theta)}$ define the model for the observations $Y_{1:T}$ given $Z_{1:T}$.

Appendix D reviews general techniques to construct $V_\theta^{(1:m)}$ and $X_\theta^{(1:m)}$ for which $\zeta_\theta(h) := \sum_{i=1}^m V_\theta^{(i)} h(X^{(i)})$ satisfy:

$$\mathbb{E}[\zeta_\theta(h)] = \int p^{(\theta)}(z_{1:T}, y_{1:T}) h(z_{1:T}) \mathrm{d}z_{1:T}, \tag{8}$$

Accepted Article

This article is protected by copyright. All rights reserved.

for any θ and for some class of functions $h : \mathsf{S}_z^T \to \mathbb{R}$. These random variables lead directly to a proper weighting; see Corollary 39 in Appendix D.

We focus next on a special case of the general SSM, where both S_z and S_y are Euclidean and $\mu_t^{(\theta)}$ are linear-Gaussian, but the observation models $g_t^{(\theta)}$ may be non-linear and/or non-Gaussian, taking the form

$$g_t^{(\theta)}(y_t \mid z_t) = \eta_t^{(\theta)}(y_t \mid H_t^{(\theta)} z_t).$$

Our setting covers exponential family observation models with Gaussian, Poisson, binomial, negative binomial, and Gamma distributions, and a stochastic volatility model. This class contains a large number of commonly used models, such as structural time series models, cubic splines, generalised linear mixed models, and classical autoregressive integrated moving average models.

8.1 | Marginal approximation

The scheme we consider here is based on (Shephard & Pitt, 1997; Durbin & Koopman, 1997), and relies on a Laplace approximation $p_a^{(\theta)}(z_{1:T}, \tilde{y}_{1:T}^{(\theta)}) = \mu^{(\theta)}(z_{1:T})\tilde{g}^{(\theta)}(\tilde{y}_{1:T}^{(\theta)} \mid z_{1:T})$, where $\tilde{g}^{(\theta)}(\tilde{y}_{1:T}^{(\theta)} \mid z_{1:T}) := \prod_{t=1}^T \tilde{g}_t^{(\theta)}(\tilde{y}_t^{(\theta)} \mid z_t)$. The linear-Gaussian terms $\tilde{g}_t$ approximate g_t in terms of pseudo-observations $\tilde{y}_t^{(\theta)}$ and pseudo-covariances $R_t^{(\theta)}$, which are found by an iterative process, which we detail next for a fixed θ. Denote $D_t^{(n)}(z_t) := \frac{\partial^n}{\partial^n z_t} \log \eta_t^{(\theta)}(y_t \mid z_t)$, and assume that $\tilde{z}_{1:T}$ is an initial estimate for the mode $\hat{z}_{1:T}^{(\theta)}$ of $p^{(\theta)}(z_{1:T} \mid y_{1:T})$ following:

1. $R_t^{(\theta)} = -[D_t^{(2)}(H_t^{(\theta)}\tilde{z}_t)]^{-1}$ and $\tilde{y}_t^{(\theta)} = H_t^{(\theta)}\tilde{z}_t + R_t^{(\theta)} D_t^{(1)}(H_t^{(\theta)}\tilde{z}_t)$
2. Run the Kalman filter and smoother for the model with $g_t^{(\theta)}(y_t \mid z_t)$ replaced by $\tilde{g}_t^{(\theta)}(\tilde{y}_t^{(\theta)} \mid z_t) = N(\tilde{y}_t^{(\theta)}; H_t^{(\theta)} z_t, R_t^{(\theta)})$ and set $\tilde{z}_{1:T}$ to the smoothed mean.

These steps are then repeated until convergence, which is typically quick: often less than 10 iterations are enough (Durbin & Koopman, 2000).

Consider the following decomposition of the marginal likelihood:

$$L(\theta) = \tilde{L}_a(\theta) \frac{g^{(\theta)}(y_{1:T} \mid \hat{z}_{1:T}^{(\theta)})}{\tilde{g}^{(\theta)}(\tilde{y}_{1:T}^{(\theta)} \mid \hat{z}_{1:T}^{(\theta)})} \mathbb{E}\left[\frac{g^{(\theta)}(y_{1:T} \mid Z_{1:T})/g^{(\theta)}(y_{1:T} \mid \hat{z}_{1:T}^{(\theta)})}{\tilde{g}^{(\theta)}(\tilde{y}_{1:T}^{(\theta)} \mid Z_{1:T})/\tilde{g}^{(\theta)}(\tilde{y}_{1:T}^{(\theta)} \mid \hat{z}_{1:T}^{(\theta)})}\right], \tag{9}$$

where $\tilde{L}_a(\theta) := \int p_a^{(\theta)}(z_{1:T}, \tilde{y}_{1:T}^{(\theta)}) \mathrm{d}z_{1:T}$ is the marginal likelihood (from the Kalman filter), and the expectation is taken with respect to the approximate smoothing distribution $p_a^{(\theta)}(z_{1:T} \mid \tilde{y}_{1:T}^{(\theta)}) = p_a^{(\theta)}(z_{1:T}, \tilde{y}_{1:T}^{(\theta)})/\tilde{L}_a(\theta)$. If the pseudo-likelihoods $\tilde{g}_t^{(\theta)}$ are nearly proportional to the true likelihoods $g_t^{(\theta)}$ around the mode of $p_a^{(\theta)}(z_{1:T} \mid y_{1:T})$, the expectation in (9) is close to one. Our approximation is based on dropping the expectation in (9): $L_a(\theta) := \tilde{L}_a(\theta) g^{(\theta)}(y_{1:T} \mid \hat{z}_{1:T})/\tilde{g}^{(\theta)}(\tilde{y}_{1:T}^{(\theta)} \mid \hat{z}_{1:T})$. The same approximate likelihood $L_a(\theta)$ was also used in a maximum likelihood setting by Durbin & Koopman (2012) as an initial objective function before more expensive importance sampling based maximisation was done.

The evaluation of the approximation $L_a(\theta)$ above requires a reconstruction of the Laplace approximation for each value of θ. We call this *local approximation*, and consider also a faster *global approximation* variant, where the pseudo-observations and covariances are constructed only once, at the maximum likelihood estimate of θ.

Accepted Article

This article is protected by copyright. All rights reserved.

8.2 | Proper weighting schemes

The simplest approach to construct a proper weighting scheme based on the Laplace approximations is to use the approximate smoothing distribution $p_a^{(\theta)}(z_{1:T} \mid y_{1:T})$ as IS proposal. We consider such scheme using the simulation smoother (Durbin & Koopman, 2002) with one antithetic variable, which we call **SPDK**, following Shephard & Pitt (1997).

We consider also several variants of M_t and G_t in the particle filter discussed in Appendix D. The bootstrap filter (Gordon et al., 1993), abbreviated as **BSF**, uses $M_t = \mu_t^{(\theta)}$ and $G_t = g_t^{(\theta)}(y_t \mid \cdot)$, and hence does not rely on an approximation. Inspired by the developments in (Whiteley & Lee, 2014; Guarniero et al., 2017), we consider also the choice

$$M_t(z_t \mid z_{1:t-1}) = p_a^{(\theta)}(z_t \mid z_{t-1}, y_{1:T}), \qquad \text{and} \qquad G_t(z_{1:t}) = g_t^{(\theta)}(y_t \mid z_t)/\tilde{g}_t^{(\theta)}(\tilde{y}_t \mid z_t),$$

where $p_a^{(\theta)}(z_t \mid z_{t-1}, y_{1:T}) = p_a^{(\theta)}(z_t \mid z_{1:t-1}, y_{1:T})$ are conditionals of $p_a^{(\theta)}(z_{1:T} \mid y_{1:T})$. This would be optimal in our setting if the G_t were constants (Guarniero et al., 2017). As they are often approximately so, we believe that this choice, which we call **ψ-APF** following Guarniero et al. (2017), can provide substantial benefits over BSF.

9 | DISCRETELY OBSERVED DIFFUSIONS

In many applications, for instance in finance or physical systems modelling, the SSM state transitions arise naturally from a continuous time diffusion model, such as

$$\mathrm{d}\tilde{Z}_t = m^{(\theta)}(t, \tilde{Z}_t)\mathrm{d}t + \sigma^{(\theta)}(t, \tilde{Z}_t)\mathrm{d}B_t,$$

where B_t is a (vector valued) Brownian motion and where $m^{(\theta)}$ and $\sigma^{(\theta)}$ are functions (vector and matrix valued, respectively). The latent variables $X = (Z_1, \dots, Z_T)$ are assumed to follow the law of $(\tilde{Z}_{t_1}, \dots, \tilde{Z}_{t_T})$, so $\mu_k^{(\theta)}$ would ideally be the transition densities of $\tilde{Z}_{t_k}$ given $\tilde{Z}_{t_{k-1}}$. These transition densities are generally unavailable (for non-linear diffusions), but standard time-discretisation schemes allow for straightforward approximate simulation (cf. Kloeden & Platen, 1992). The denser the time-discretisation the less bias, but the computational complexity of the simulation is higher — generally proportional to the size of the mesh.

The MCMC-IS may be applied to speed up the inference of discretely observed diffusions by the following simple two-level approach. The 'true' state transition $\mu_t^{(\theta)}$ are based on 'fine enough' discretisations, which are assumed to ensure a negligible bias, but which are expensive to simulate. Cheaper 'coarse' discretisation corresponds to transitions $\hat{\mu}_t^{(\theta)}$.

Because neither of the models admit exact calculations, we may only use a pseudo-marginal approximate chain as discussed in Sections 2.2 and 7). More specifically, we may use the bootstrap filter (Appendix D) with SSM $(\hat{\mu}_t^{(\tilde{\Theta}_k)}, g_t^{(\tilde{\Theta}_k)})$ to generate the likelihood estimators $\tilde{U}_k$ in Phase 1', and in Phase 2', we may use bootstrap filters for SSM $(\mu_t^{(\Theta_k)}, g_t^{(\Theta_k)})$ to generate $(V_k^{(i)}, X_k^{(i)})$.

Assuming that the observation model satisfies $g_t^{(\theta)} > 0$ guarantees the validity of this scheme, because then $\tilde{U}_k > 0$ (see Proposition 28 (i)). It is most straightforward to simulate the bootstrap filters in Phases 1' and 2' independent of each other, but they may be made dependent as well, by using a coupling strategy (cf. Sen et al., 2018). The correction phase could be also based on exact sampling for diffusions (Beskos et al., 2006), which allow for elimination of the discretisation bias entirely.

The recent work (Franks et al., 2018) details how unbiased inference is also possible with IS type correction, using randomised multilevel Monte Carlo.

Accepted Article

This article is protected by copyright. All rights reserved.

10 | EXPERIMENTS

We did experiments for our generic framework with SSMs, using Laplace approximations (Section 8) and an approximation based on coarsely discretised diffusions (Section 9). We compared several approaches in our experiments:

AI Approximate inference with MCMC targeting $\pi_a(\theta)$, and for each accepted $\tilde{\Theta}_k$, sampling one realisation from $\tilde{p}^{(\tilde{\Theta}_k)}(z_{1:T} \mid y_{1:T})$.

PM Pseudo-marginal MCMC with m samples targeting directly $\pi(\theta, x)$.

DA Two-level delayed acceptance pseudo-marginal MCMC with first stage acceptance based on $\pi_a(\theta)$ and with target $\pi(\theta, x)$.

IS1 Jump chain IS correction with mN_k samples for each accepted $\tilde{\Theta}_k$.

IS2 Jump chain IS correction with m samples for each accepted $\tilde{\Theta}_k$.

The IS1 algorithm is similar to simple IS estimator (1), but is expected to be generally safer; see Remark 21 (ii) Except for AI, all the algorithms are asymptotically exact. Ignoring the effects of parallel implementation, the average computational complexity, or cost, of DA and IS2 are roughly comparable, and we have similar pairing between PM and IS1. However, as the weighting in IS methods is based only on the post-burn-in chain, the IS methods are generally somewhat faster.

We used a random walk Metropolis algorithm for π_a with a Gaussian proposal distribution, whose covariance was adapted during burn-in following (Vihola, 2012), targeting the acceptance rate 0.234. In DA, the adaptation was based on the first stage acceptance probability, which was also used in PM for SPDK and ψ-APF variants as this led to more robust adaptation.

All the experiments were conducted in `R` (R Core Team, 2016) using the `bssm` package which is available online (Helske & Vihola, 2017). The experiments were run on a Linux server with four 24-core Intel Xeon E7-8890 2.6GHz processors with total 2.1TB of RAM.

In each experiment, we calculated the Monte Carlo estimates several times independently, and the inverse relative efficiency (IRE) was reported. The IRE, defined as the mean square error (MSE) of the estimate multiplied by the average computation time, provides a justified way to compare Monte Carlo algorithms with different costs (Glynn & Whitt, 1992).

Further details and results of the experiments may be found in the preprint version of our article (Vihola et al., 2017).

10.1 | Laplace approximations

In case of Laplace approximations, the maximum likelihood estimate of θ was always used as the starting value of MCMC. We used sub-sampling as in Proposition 9, and sampled one trajectory $Z_{1:T}$ per each accepted state. We tested the exact methods with three different IS correction schemes, SPDK, BSF and ψ-APF, described in Section 8.2. For BSF and ψ-APF, the filter-smoother estimates as in Proposition 38 (i) were used. When calculating the MSE, we used the average over all estimates from all unbiased algorithms as the ground truth.

For all the exact methods, we chose the IS accuracy parameter m based on a pilot experiment, following the guidelines for optimal tuning of pseudo-marginal MCMC in (Doucet et al., 2015). More specifically, m was set so that the standard deviation of the logarithm of the likelihood estimate, denoted with δ, was around 1.2 in the neighbourhood of the posterior mean of θ. We kept the same m for all methods, for comparability, even though in some cases optimal choice might differ (Sherlock et al., 2017).

Accepted Article

This article is protected by copyright. All rights reserved.

10.1.1 | Poisson observations

Our first model is of the following form:

$$g_t^{(\theta)}(y_t \mid z_t) = \text{Poisson}(y_t; e^{u_t}), \quad \text{and} \quad \begin{pmatrix} u_{t+1} \\ v_{t+1} \end{pmatrix} = \begin{pmatrix} u_t + v_t + \sigma_\eta \eta_t \\ v_t + \sigma_\xi \xi_t \end{pmatrix},$$

with $Z_1 = (U_1, V_1) \sim N(0, 0.1I)$, where $\xi_t, \eta_t \sim N(0,1)$. For testing our algorithms, we simulated a single set of observations $y_{1:100}$ from this model with $Z_1 = (0,0)$ and $\theta = (\sigma_\eta, \sigma_\xi) = (0.1, 0.01)$. We used a uniform prior $U(0, 2s)$ for the parameters, where the cut-off parameter s was set to 1.6 based on the sample standard deviation of $\log(y_{1:T})$, where zeros were replaced with 0.1. Results were not sensitive to this upper bound.

Based on a pilot optimisation, we set $m = 10$ for SPDK and ψ-APF, and $m = 200$ for BSF, leading to $\delta \approx 0.05$ for SPDK, $\delta \approx 0.1$ for ψ-APF, and $\delta \approx 1.2$ for BSF. For all algorithms, we used 100,000 MCMC iterations with the first half discarded as burn-in. We ran all the algorithms independently 1000 times.

Table 1 shows the IREs, which are re-scaled such that all IREs of PM-BSF equal one. The overall acceptance rate of DA-BSF was around 0.101, and 0.218-0.234 for all others. All exact methods led to essentially the same overall mean estimate $(0.093, 0.016, -0.075, 2.618)$ for $(\sigma_\eta,\ \sigma_\xi,\ u_1,\ u_{100})$, in contrast with AI showing some bias on (u_1, u_{100}), with overall mean estimates $(-0.063, 2.629)$ and $(-0.064, 2.631)$ with local and global approximation, respectively. IS2-BSF clearly outperformed DA-BSF in terms of IRE, because of the burn-in benefit. Similarly, IS1-BSF outperformed PM-BSF by a clear margin. With SPDK and ψ-APF, the IS1 and IS2 outperformed the PM and DA alternatives, but with a smaller margin because of smaller overall execution times. There were no significant differences between the SEs of local and global variants, but the global one was faster leading to smaller IREs. The execution times of SPDK were slightly less than ψ-APF due to the use of antithetic variable which increased the relative efficiency of SPDK algorithm.

Using Corollary 14, we constructed 95% confidence intervals for our IS-MCMC estimators, and computed the average coverage of these intervals over all replications. The coverages were between 0.93 and 0.99, averaging the nominal 0.95 over all methods, with no clear differences between them. In addition, we computed the 'variance decomposition' as suggested in Remark 16. That is, we calculated the estimator $n\hat{v}_n$ of Theorem 15, and calculated the proportion of it with respect to the overall asymptotic variance estimated with $\hat{s}_n$ of Corollary 14, using the integrated autocovariance estimator suggested by Sokal (1996). For the hyperparameters σ_η and σ_ξ, these proportions were around 0.5 and 0.6, respectively, except for IS2-BSF which resulted proportions 0.7 and 0.8. This suggests that the MCMC autocorrelations and the IS correction contribute roughly equally to the overall uncertainty For the latent states (u_1, u_{100}) the proportion was close to 1 in all cases. This is expected, because in the case of the latent variables, the centred conditional expectation terms $\mu_{\bar{f}}^2$ are expected to be small relative to the conditional variance v.

10.1.2 | Stochastic volatility model

Our second illustration is more challenging, involving analysis of real time series: the daily log-returns for the S&P index from 4/1/1995 to 28/9/2016, with total number of observations $T = 5473$. The data was analysed using the following stochastic volatility (SV) model:

$$Y_t \mid Z_t \sim N(0, e^{Z_t}), \qquad Z_{t+1} \mid Z_t \sim N(\nu + \phi(Z_t - \nu), \sigma_\eta^2),$$

with $Z_1 \sim N(\nu, \sigma_\eta^2/(1-\phi^2))$. We used a uniform prior on $[-0.9999, 0.9999]$ for ϕ, a half-Gaussian prior with standard deviation 5 for σ_η, and a zero-mean Gaussian prior with standard deviation 5 for ν. SPDK was expected to be problematic, due to its well-

Accepted Article

This article is protected by copyright. All rights reserved.

TABLE 1 IREs for the Poisson model, with local (top) and global (bottom) approximations. Times are in seconds. For PM-BSF, IREs are one and time 811s.

	BSF				SPDK				ψ-APF			
	AI	DA	IS1	IS2	PM	DA	IS1	IS2	PM	DA	IS1	IS2
Time	82	273	540	178	137	93	102	86	206	110	111	94
σ_η	0.031	0.354	0.240	0.134	0.051	0.035	0.037	0.033	0.077	0.044	0.043	0.034
σ_ξ	0.037	0.388	0.290	0.159	0.063	0.043	0.049	0.042	0.091	0.054	0.056	0.042
u_1	0.964	0.795	0.544	0.414	0.122	0.076	0.090	0.040	0.176	0.095	0.093	0.072
u_{100}	1.713	0.825	0.523	0.382	0.113	0.079	0.082	0.042	0.172	0.101	0.095	0.067
Time	23	214	482	120	78	35	44	28	147	51	53	36
σ_η	0.012	0.284	0.220	0.091	0.033	0.015	0.019	0.013	0.065	0.024	0.022	0.017
σ_ξ	0.065	0.340	0.268	0.118	0.041	0.019	0.024	0.015	0.074	0.029	0.027	0.019
u_1	0.211	0.601	0.440	0.271	0.061	0.028	0.032	0.013	0.122	0.043	0.039	0.027
u_{100}	0.660	0.592	0.433	0.240	0.067	0.031	0.038	0.015	0.126	0.048	0.042	0.023

known exponential deterioration in T, unlike the particle filter which often scales much better in T (Whiteley, 2013). In addition, it is known that for this particular model, the importance weights may have large variability (Pitt et al., 2013; Koopman et al., 2009). While in principle ψ-APF may also be affected by such fluctuations, we did not observe any problems with it in our experiments.

Based on our pilot experiment, we chose $m = 10$ for ψ-APF, $m = 70$ for SPDK and $m = 3400$ for BSF, which all led to $\delta \approx 1.1$. We used 100,000 MCMC iterations with the first half discarded as burn-in, and 100 independent replications. the IREs re-scaled here with respect to DA-BSF are shown in Table 2. The PM and IS1 were not tested because of their high costs. The results with global approximation are shown only for AI, and indicate significant computational savings. The parallelisation with 8 cores dropped the execution time nearly ideally. The total acceptance rates were 0.1 for DA-BSF, PM-SPDK and DA-ψ-APF, 0.06 for DA-SPDK, and 0.15 for PM-ψ-APF.

Like in the Poisson experiment, the overall means of the exact methods were close to each other, but AI had some bias, this time also with some of the hyperparameters (σ_η and ν). The IS1 and IS2 methods outperformed the PM and DA methods similarly as in the Poisson experiment. Due to a much smaller m, the DA-SPDK and DA-ψ-APF were an order of magnitude faster than DA-BSF. Diagnostics from the individual runs of PM-SPDK and DA-SPDK sometimes showed poor mixing, and despite the large reductions in execution time, the IREs were worse than PM-BSF. We observed also cases with a few very large correction weights in IS1-SPDK and IS2-SPDK, which had some impact also on their efficiencies. The SEs of DA-ψ-APF were comparable with the DA-BSF. We did not experience problems with mixing or overly large weights with ψ-APF, which suggests ψ-APF being more robust than SPDK. There were no significant differences in the SEs between the exact methods when using the local and global approximation schemes.

We calculated the asymptotic variances, the 95% confidence intervals and the variance decomposition as in Section 10.1.1. The coverages of 95% CIs for IS-MCMC estimators were between 0.91–1 for all variants and variables, except for σ_η for which the average coverages were 0.77, 0.85, and 0.88 for IS1-SPDK, IS2-BSF, and IS2-SPDK, but this is likely due to random fluctuation, because there were only 100 replications. For the latent states, the proportions $n\hat{v}_n/\hat{s}_n$ were close to 1 in all cases, as in Section 10.1.1. For the hyperparameters, the proportion was around 0.9 for IS1-SPDK, 0.6 for IS1-ψ-APF, 1 for IS2-SPDK,

Accepted Article

This article is protected by copyright. All rights reserved.

TABLE 2 IREs for SV model. Times are in hours. AI^{G} is with global approximation and $\mathrm{IS2}^8$ is with 8 parallel cores. For DA-BSF, IREs are one and time 37h.

	BSF				SPDK				ψ-APF			
	AI	AI^{G}	IS2	$\mathrm{IS2}^8$	PM	DA	IS1	IS2	PM	DA	IS1	IS2
Time	0.9	0.2	19.6	3.1	3.8	1.6	2.4	1.2	2.1	1.2	1.2	1
ϕ	0.114	0.114	0.328	0.068	0.634	2.901	0.346	0.645	0.035	0.036	0.012	0.017
σ_η	0.466	0.259	0.493	0.068	3.014	1.190	0.456	0.545	0.035	0.035	0.013	0.019
ν	0.008	1.023	0.388	0.080	2.346	1.867	0.287	0.606	0.041	0.033	0.013	0.019
Z_1	0.498	0.184	0.389	0.096	1.834	1.081	1.448	0.201	0.067	0.033	0.013	0.019
Z_{5473}	0.571	0.152	0.385	0.053	2.611	1.224	1.021	0.396	0.035	0.031	0.009	0.014

and 0.8 for both IS2-ψ-APF and IS2-BSF. These proportions suggest that SPDK is less efficient than ψ-APF, which is in line with the observed IREs in Table 2.

10.2 | Discretely observed Geometric Brownian motion

Our last experiment was about a discretely observed diffusion as discussed in Section 9. The model was a geometric Brownian motion, with noisy log-observations:

$$\mathrm{d}\tilde{Z}_t = \nu \tilde{Z}_t \mathrm{d}t + \sigma_z \tilde{Z}_t \mathrm{d}B_t, \qquad Y_k \mid (Z_k = z) \sim N(\log(z), \sigma_y^2),$$

with $\tilde{Z}_0 \equiv 1$, where $(B_t)_{t\geq 1}$ stands for the standard Brownian motion, and where $Z_k = \tilde{Z}_k$. The discretisations $\mu_t^{(\theta)}$ and $\hat{\mu}_t^{(\theta)}$ were based on a Milstein discretisation with uniform meshes of sizes 2^{L_F} and 2^{L_C}, respectively, with $L_C = 4$ and $L_F = 16$, reflected to positive values. We did not consider optimising L_C and L_F, but rather aimed for illustrating the potential gains for the IS2 algorithm from parallelisation. The data was a single simulated realisation of length 50 from the exact model, with $\nu = 0.05$, $\sigma_x = 0.3$, and $\sigma_y = 1$. We used a half-Gaussian prior with s.d. 0.1 for ν, a half-Gaussian prior with s.d. 0.5 for σ_x, and $N(1.5, 0.5^2)$ prior truncated to > 0.5 for σ_y. For both IS2 and DA, and both levels, we used $m = 50$ which led to $\delta \approx 0.6$.

Assuming a unit cost for each step in the BSF, the total average cost of a parallel IS2 run is $n2^{L_C} + \alpha(n - n_b)2^{L_F}/M$, where α is the mean acceptance rate of the approximate MCMC, n_b is the length of burn-in and M is the number of parallel cores used for the weighting. We chose $n = 5000$, $n_b = 2500$, $M = 48$, and the target acceptance rate $\alpha = 0.234$, leading to an expected 43-fold speed-up due to the parallelisation of IS2.

Single run of DA cannot be easily parallelised, but we ran instead multiple independent DA chains in parallel, and averaged their outputs for inference. While such parallelisation may not be optimal, it allows for utilisation of all of the available computational resources. The running time of each DA chain was constrained to be similar to the time required by IS2, leading to $n = 100$ with $n_b = 50$. Because of the short runs, we suspected that initial bias could play a significant role, which was explored by running two experiments, with MCMC initialised either to the prior mean $\theta_0 = (0.08, 0.4, 1.5)$, or to an independent sample from the prior. We experimented also with further thinning, by forming the IS2 correction based on every other accepted state.

Table 3 summarises the results from 100 replications. The run time of the parallel DA algorithms was defined as the maximum run time of all parallel chains. The parallelisation speedup of IS2 was nearly optimal, as well as the further speedup from thinning. The SEs with prior mean initialisation were similar between DA and IS2, but DA produced slightly biased results, leading to 9.5 to 13.0 times higher IREs. The efficiency gains of thinning were inconclusive, indicating some gains for

Accepted Article

This article is protected by copyright. All rights reserved.

TABLE 3 Results for the geometric Brownian motion experiment using 48 cores. IS2^t is with thinning, and time is in minutes. Ground truth (GT) was calculated with MCMC using exact latent inference.

		Mean					IRE				
Init.		Prior mean			Prior sample		Prior mean			Prior sample	
	GT	DA	IS2	IS2^t	DA	IS2	DA	IS2	IS2^t	DA	IS2
Time	—	12.3	3.4	1.9	14.0	3.3	12.3	3.4	1.9	14.0	3.3
ν	0.053	0.061	0.053	0.053	0.064	0.053	0.069	0.004	0.002	0.135	0.004
σ_x	0.253	0.278	0.253	0.253	0.251	0.252	0.576	0.029	0.019	0.336	0.022
σ_y	1.058	1.054	1.058	1.058	1.083	1.058	0.088	0.020	0.014	1.010	0.022
Z_1	1.254	1.273	1.254	1.246	1.243	1.252	0.670	0.109	0.119	0.805	0.103
Z_{50}	2.960	2.953	2.966	2.935	20.773	2.971	12.605	1.880	2.074	4×10^6	2.308

the hyperparameters θ, but not for the state variables. The smaller memory requirements and smaller absolute time requirements for the thinning make it still appealing. With prior sample initialisation, DA behaved sometimes poorly, in contrast with IS2 which behaved similarly with both initialisation strategies.

10.3 | Summary of results

In our experiments with Laplace approximations, IS1 and IS2 were competitive alternatives to PM and DA, respectively, even without parallelisation. The differences were more emphasised when the cost of correction (number of samples m) was higher. The ψ-APF was generally preferable over SPDK, and BSF was the least efficient. The global approximation gave additional performance boost in our experiments, without compromising the accuracy of the estimates, but we stress that it may not be stable in all scenarios.

As noted earlier, the use of the guidelines by Doucet et al. (2015) were not necessarily optimal in our setting. We did an additional experiment to inspect how the choice of m affects the IRE with BSF in the Poisson model, and with ψ-APF in the SV model. Figure 1 shows the average IREs as a function of m. Both IS2 and DA behaved similarly, and IS2 was less than DA uniformly in terms of IRE. In the Poisson-BSF case, the choice $m = 200$ based on (Doucet et al., 2015) appears nearly optimal. In case of the SV-ψ-APF, the optimal m for DA and IS2 was around 50, which was higher than $m = 10$ based on (Doucet et al., 2015). This is likely because of the initial overhead cost of the approximation.

The discretely observed geometric Brownian motion example illustrated the potential gains which may be achieved by using the IS2 method in a parallel environment. While we admit that our experiment is academic, we believe that it is indicative, and shows that IS2 can provide substantial gains, and makes reliable inference possible in a much shorter time than DA. The IS framework is less prone to issues with burn-in bias, which can be problematic with naive MCMC parallelisation based on independent chains.

11 | DISCUSSION

Our framework of IS type estimators based on approximate marginal MCMC provides a general way to construct consistent estimators. Our experiments demonstrate that the IS estimator can provide substantial speedup relative to a delayed acceptance (DA) analogue with parallel computing, and appears to be competitive to DA even without parallelisation. We believe that IS is

Accepted Article

This article is protected by copyright. All rights reserved.

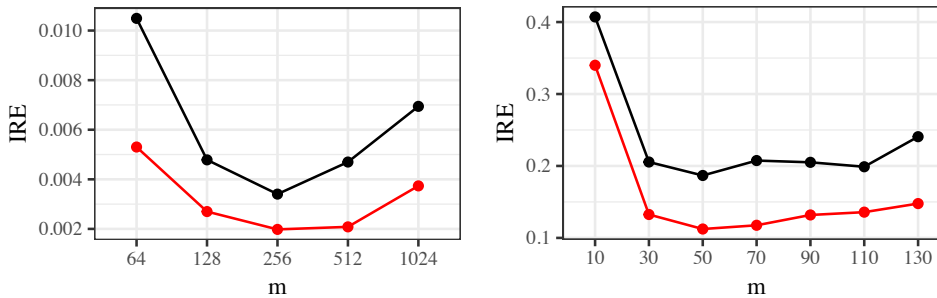

FIGURE 1 Average IRE of $(\sigma_\eta, \sigma_\xi, Z_1, Z_{100})$ in the Poisson model with BSF (left) and of $(\phi, \sigma_\eta, \nu, Z_1, Z_{5473})$ in the SV model with ψ-APF (right). DA is shown in black and IS2 in red.

often better than DA in practice, but it is not hard to find simple examples where DA can be arbitrarily better than IS (and vice versa) (Franks & Vihola, to appear). Our followup work (Franks & Vihola, to appear) complements our findings by theoretical considerations, with guaranteed asymptotic variance bounds between IS and DA.

IS is known to be difficult to implement efficiently in high dimensions, but this is not a major concern in most latent variable models, where the hyperparameters are low-dimensional. It is also generally desirable to design the approximate marginal π_a to have heavier tails than the desired marginal π_m, in order to guarantee bounded (expected) weights. When the method of Section 2 is used, the IS weight may be directly regularised by inflating the (estimated) approximate likelihood, for instance with $L_a(\theta) + \epsilon$, with some $\epsilon > 0$. If the likelihood L is bounded, then $w_u(\theta) \propto L(\theta)/(L_a(\theta) + \epsilon)$ is bounded as well. The latter approach can be seen as an instance of defensive importance sampling (Hesterberg, 1995), using the prior as a proposal component. Other generic safe IS schemes may also be useful (cf. Owen & Zhou, 2000), and tempering may be applied for the likelihood as well.

We used adaptive MCMC in order to construct the marginal chain $(\Theta_k)_{k\geq 1}$ in our experiments, and believe that it is often useful (cf. Andrieu & Thoms, 2008). Note, however, that our theoretical results do not apply directly with adaptive MCMC, unless the adaptation is stopped after suitable burn-in. Our results could be extended to hold with continuous adaptation, under certain technical conditions. We detailed proper weighting schemes based on standard IS and particle filters. We note that various PF variations, such as Rao-Blackwellisation, alternative resampling strategies (Cappé et al., 2005), or quasi-Monte Carlo updates (Gerber & Chopin, 2015), apply directly. PFs can also be useful beyond the state space models context (Del Moral et al., 2006). Twisted particle filters (Whiteley & Lee, 2014; Ala-Luhtala et al., 2016) could also be applied, instead of the ψ-APF.

In a diffusion context, a proper weighting can be constructed based on randomised multilevel Monte Carlo, as recently described in (Franks et al., 2018), and ABC post-correction may be seen as IS-type correction (Vihola & Franks, 2020). Laplace approximations are available for a wider class of Gaussian latent variable models beyond SSMs (cf. Rue et al., 2009). Variational approximations (Beal, 2003; Jordan, 2004) and expectation propagation (Minka, 2001) have been found useful in a wide variety of models. In the SSM context, various non-linear filters could also be applied (cf. Särkkä, 2013). Our framework provides a generic validation mechanism for approximate inference, where assessment of bias is difficult in general (cf. Ogden, 2017). Contrary to purely approximate inference, our approach only requires moderately accurate approximations, as demonstrated by our experiment with global Laplace approximations. Debiased MCMC, as suggested by Glynn & Rhee (2014) and further explored in (Jacob et al., 2020a,b), may also lead to useful proper weighting schemes.

Accepted Article

This article is protected by copyright. All rights reserved.

ACKNOWLEDGEMENTS

We thank Christophe Andrieu, Arnaud Doucet, Anthony Lee and Chris Sherlock for many insightful remarks.

REFERENCES

Ala-Luhtala, J., Whiteley, N., Heine, K. & Piché, R. (2016) An introduction to twisted particle filters and parameter estimation in non-linear state-space models. *IEEE Trans. Signal Process.*, **64**, 4875–4890.

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

Andrieu, C., Lee, A. & Vihola, M. (2018) Uniform ergodicity of the iterated conditional SMC and geometric ergodicity of particle Gibbs samplers. *Bernoulli*, **24**, 842–872.

Andrieu, C. & Roberts, G. O. (2009) The pseudo-marginal approach for efficient Monte Carlo computations. *Ann. Statist.*, **37**, 697–725.

Andrieu, C. & Thoms, J. (2008) A tutorial on adaptive MCMC. *Statist. Comput.*, **18**, 343–373.

Andrieu, C. & Vihola, M. (2015) Convergence properties of pseudo-marginal Markov chain Monte Carlo algorithms. *Ann. Appl. Probab.*, **25**, 1030–1077.

— (2016) Establishing some order amongst exact approximations of MCMCs. *Ann. Appl. Probab.*, **26**, 2661–2696.

Beal, M. J. (2003) *Variational algorithms for approximate Bayesian inference*. Ph.D. thesis, University College London.

Beaumont, M. A. (2003) Estimation of population growth or decline in genetically monitored populations. *Genetics*, **164**, 1139–1160.

Beskos, A., Papaspiliopoulos, O., Roberts, G. O. & Fearnhead, P. (2006) Exact and computationally efficient likelihood-based estimation for discretely observed diffusion processes. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, **68**, 333–382.

Bhattacharya, S. (2008) Consistent estimation of the accuracy of importance sampling using regenerative simulation. *Statist. Probab. Lett.*, **78**, 2522–2527.

Cappé, O., Moulines, E. & Rydén, T. (2005) *Inference in Hidden Markov Models*. Springer.

Chopin, N., Jacob, P. & Papaspiliopoulos, O. (2013) SMC^2: A sequential Monte Carlo algorithm with particle Markov chain Monte Carlo updates. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, **75**, 397–426.

Chopin, N. & Singh, S. S. (2015) On particle Gibbs sampling. *Bernoulli*, **21**, 1855–1883.

Christen, J. A. & Fox, C. (2005) Markov chain Monte Carlo using an approximation. *J. Comput. Graph. Statist.*, **14**.

Del Moral, P. (2004) *Feynman-Kac Formulae*. Springer.

Del Moral, P., Doucet, A. & Jasra, A. (2006) Sequential Monte Carlo samplers. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, **68**, 411–436.

Deligiannidis, G., Doucet, A. & Pitt, M. K. (2018) The correlated pseudomarginal method. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, **80**, 839–870.

Deligiannidis, G. & Lee, A. (2018) Which ergodic averages have finite asymptotic variance? *Ann. Appl. Probab.*, **28**, 2309–2334.

Doss, H. (1994) Discussion: Markov chains for exploring posterior distributions. *Ann. Statist.*, **22**, 1728–1734.

— (2010) Estimation of large families of Bayes factors from Markov chain output. *Statist. Sinica*, **20**, 537–560.

Accepted Article

This article is protected by copyright. All rights reserved.

Douc, R., Cappé, O. & Moulines, E. (2005) Comparison of resampling schemes for particle filtering. In *Proc. Image and Signal Processing and Analysis, 2005*, 64–69.

Douc, R., Moulines, E., Priouret, P. & Soulier, P. (2018) *Markov chains*. Springer.

Douc, R. & Robert, C. P. (2011) A vanilla Rao-Blackwellization of Metropolis-Hastings algorithms. *Ann. Statist.*, **39**, 261–277.

Doucet, A., de Freitas, N. & Gordon, N. (2001) *Sequential Monte Carlo Methods in Practice*. New York: Springer-Verlag.

Doucet, A., Godsill, S. & Andrieu, C. (2000) On sequential Monte Carlo sampling methods for Bayesian filtering. *Statist. Comput.*, **10**, 197–208.

Doucet, A. & Lee, A. (2019) Sequential Monte Carlo methods. In *Handbook of Graphical Models* (eds. M. Matthuis, M. Drton, S. Lauritzen, & M. Wainwright), 165–188. CRC press.

Doucet, A., Pitt, M., Deligiannidis, G. & Kohn, R. (2015) Efficient implementation of Markov chain Monte Carlo when using an unbiased likelihood estimator. *Biometrika*, **102**, 295–313.

Durbin, J. & Koopman, S. J. (1997) Monte Carlo maximum likelihood estimation for non-Gaussian state space models. *Biometrika*, **84**, 669–684.

— (2000) Time series analysis of non-Gaussian observations based on state space models from both classical and Bayesian perspectives. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, **62**, 3–56.

— (2002) A simple and efficient simulation smoother for state space time series analysis. *Biometrika*, **89**, 603–615.

— (2012) *Time series analysis by state space methods*. New York: Oxford University Press, 2nd edn.

Flegal, J. M. & Jones, G. L. (2010) Batch means and spectral variance estimators in Markov chain Monte Carlo. *Ann. Statist.*, **38**, 1034–1070.

Fox, C. & Nicholls, G. (1997) Sampling conductivity images via MCMC. In *Proceedings in The Art and Science of Bayesian Image Analysis* (eds. K. V. Mardia, C. A. Gill & R. G. Aykroyd), 91–100. Leeds University Press.

Franks, J., Jasra, A., Law, K. & Vihola, M. (2018) Unbiased inference for discretely observed hidden markov model diffusions. *Preprint*, arXiv:1807.10259.

Franks, J. & Vihola, M. (to appear) Importance sampling and delayed acceptance via a Peskun type ordering. *Stochastic Process. Appl.*

Gerber, M. & Chopin, N. (2015) Sequential quasi-Monte Carlo. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, **77**, 509–579.

Giles, M. B. (2008) Multilevel Monte Carlo path simulation. *Oper. Res.*, **56**, 607–617.

Glynn, P. W. & Iglehart, D. L. (1989) Importance sampling for stochastic simulations. *Management Science*, **35**, 1367–1392.

Glynn, P. W. & Rhee, C.-H. (2014) Exact estimation for Markov chain equilibrium expectations. *J. Appl. Probab.*, **51**, 377–389.

Glynn, P. W. & Whitt, W. (1992) The asymptotic efficiency of simulation estimators. *Oper. Res.*, **40**, 505–520.

Godsill, S. J., Doucet, A. & West, M. (2004) Monte Carlo smoothing for nonlinear time series. *J. Amer. Statist. Assoc.*, **99**, 156–168.

Golightly, A., Henderson, D. A. & Sherlock, C. (2015) Delayed acceptance particle MCMC for exact inference in stochastic kinetic models. *Statist. Comput.*, **25**, 1039–1055.

Gordon, N. J., Salmond, D. J. & Smith, A. F. M. (1993) Novel approach to nonlinear/non-Gaussian Bayesian state estimation. *IEE Proceedings-F*, **140**, 107–113.

Accepted Article

This article is protected by copyright. All rights reserved.

Green, P. J., Łatuszyński, K., Pereyra, M. & Robert, C. P. (2015) Bayesian computation: a summary of the current state, and samples backwards and forwards. *Statist. Comput.*, **25**, 835–862.

Guarniero, P., Johansen, A. M. & Lee, A. (2017) The iterated auxiliary particle filter. *J. Amer. Statist. Assoc.*, **112**, 1636–1647.

Hastings, W. K. (1970) Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, **57**, 97–109.

Heinrich, S. (2001) Multilevel Monte Carlo methods. In *Large-scale scientific computing*, 58–67. Springer.

Helske, J. & Vihola, M. (2017) *bssm: Bayesian inference of non-linear and non-Gaussian state space models in R*. `https://CRAN.R-project.org/package=bssm`.

Hesterberg, T. (1995) Weighted average importance sampling and defensive mixture distributions. *Technometrics*, **37**, 185–194.

Jacob, P. E., Lindsten, F. & Schön, T. B. (2020a) Smoothing with couplings of conditional particle filters. *J. Amer. Statist. Assoc.*, **115**, 721–729.

Jacob, P. E., O'Leary, J. & Atchadé, Y. F. (2020b) Unbiased Markov chain Monte Carlo with couplings. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, **82**, 543–600.

Jacob, P. E. & Thiery, A. H. (2015) On nonnegative unbiased estimators. *Ann. Statist.*, **43**, 769–784.

Jarner, S. F. & Hansen, E. (2000) Geometric ergodicity of Metropolis algorithms. *Stochastic Process. Appl.*, **85**, 341–361.

Jarner, S. F. & Roberts, G. O. (2007) Convergence of heavy-tailed Monte Carlo Markov chain algorithms. *Scand. J. Stat.*, **34**, 781–815.

Jones, G. L. (2004) On the Markov chain central limit theorem. *Probab. Surv.*, **1**, 299–320.

Jordan, M. I. (2004) Graphical models. *Statist. Sci.*, 140–155.

Karagiannis, G. & Andrieu, C. (2013) Annealed importance sampling reversible jump MCMC algorithms. *J. Comput. Graph. Statist.*, **22**, 623–648.

Kipnis, C. & Varadhan, S. S. (1986) Central limit theorem for additive functionals of reversible Markov processes and applications to simple exclusions. *Comm. Math. Phys.*, **104**, 1–19.

Kitagawa, G. (1996) Monte Carlo filter and smoother for non-Gaussian nonlinear state space models. *J. Comput. Graph. Statist.*, **5**, 1–25.

Kloeden, P. E. & Platen, E. (1992) *Numerical solution of stochastic differential equations*. Springer.

Koopman, S. J., Shephard, N. & Creal, D. (2009) Testing the assumptions behind importance sampling. *J. Econometrics*, **149**, 2 – 11.

Lee, A. & Łatuszynski, K. (2014) Variance bounding and geometric ergodicity of Markov chain Monte Carlo kernels for approximate Bayesian computation. *Biometrika*, **101**, 655–671.

Lee, A., Yau, C., Giles, M. B., Doucet, A. & Holmes, C. C. (2010) On the utility of graphics cards to perform massively parallel simulation of advanced Monte Carlo methods. *J. Comput. Graph. Statist.*, **19**, 769–789.

Lin, L., Liu, K. & Sloan, J. (2000) A noisy Monte Carlo algorithm. *Phys. Rev. D*, **61**.

Lindsten, F., Douc, R. & Moulines, E. (2015) Uniform ergodicity of the Particle Gibbs sampler. *Scand. J. Stat.*, **42**, 775–797.

Liu, J. S. (2003) *Monte Carlo Strategies in Scientific Computing*. New York: Springer-Verlag.

Lyne, A.-M., Girolami, M., Atchade, Y., Strathmann, H. & Simpson, D. (2015) On Russian roulette estimates for Bayesian inference with doubly-intractable likelihoods. *Statist. Sci.*, **30**, 443–467.

Accepted Article

This article is protected by copyright. All rights reserved.

Marjoram, P., Molitor, J., Plagnol, V. & Tavaré, S. (2003) Markov chain Monte Carlo without likelihoods. *Proc. Natl. Acad. Sci. USA*, **100**, 15324–15328.

Maxwell, M. & Woodroofe, M. (2000) Central limit theorems for additive functionals of Markov chains. *Ann. Probab.*, **28**, 713–724.

McLeish, D. (2011) A general method for debiasing a Monte Carlo estimator. *Monte Carlo Methods Appl.*, **17**, 301–315.

Meyn, S. & Tweedie, R. L. (2009) *Markov Chains and Stochastic Stability*. Cambridge University Press, 2nd edn.

Minka, T. P. (2001) Expectation propagation for approximate Bayesian inference. In *Proceedings of the Seventeenth conference on Uncertainty in artificial intelligence*, 362–369.

Neal, R. M. (2001) Annealed importance sampling. *Statist. Comput.*, **11**, 125–139.

Nummelin, E. (2002) MC's for MCMC'ists. *Int. Statist. Rev.*, **70**, 215–240.

Ogden, H. E. (2017) On asymptotic validity of naive inference with an approximate likelihood. *Biometrika*, **104**, 153–164.

Owen, A. & Zhou, Y. (2000) Safe and effective importance sampling. *J. Amer. Statist. Assoc.*, **95**, 135–143.

Owen, A. B. (2017) Statistically efficient thinning of a Markov chain sampler. *J. Comput. Graph. Statist.*, **26**, 738–744.

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

Pitt, M. K., Tran, M.-N., Scharth, M. & Kohn, R. (2013) On the existence of moments for high dimensional importance sampling. *Preprint*, arXiv:1307.7975.

Prangle, D. (2016) Lazy ABC. *Statist. Comput.*, **26**, 171–185.

Quiroz, M., Villani, M. & Kohn, R. (2016) Exact subsampling MCMC. *Preprint*, arXiv:1603.08232v2.

R Core Team (2016) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL: `https://www.R-project.org/`.

Rhee, C.-H. & Glynn, P. W. (2015) Unbiased estimation with square root convergence for SDE models. *Oper. Res.*, **63**, 1026–1043.

Roberts, G. O. & Rosenthal, J. S. (2006) Harris recurrence of Metropolis-within-Gibbs and trans-dimensional Markov chains. *Ann. Appl. Probab.*, **16**, 2123–2139.

Roberts, G. O. & Tweedie, R. L. (1996) Geometric convergence and central limit theorems for multidimensional Hastings and Metropolis algorithms. *Biometrika*, **83**, 95–110.

Rue, H., Martino, S. & Chopin, N. (2009) Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, **71**, 319–392.

Särkkä, S. (2013) *Bayesian filtering and smoothing*. Cambridge University Press.

Sen, D., Thiery, A. H. & Jasra, A. (2018) On coupling particle filter trajectories. *Statist. Comput.*, **28**, 461–475.

Shephard, N. & Pitt, M. K. (1997) Likelihood analysis of non-Gaussian measurement time series. *Biometrika*, **84**, 653–667.

Sherlock, C., Thiery, A. H. & Lee, A. (2017) Pseudo-marginal Metropolis–Hastings sampling using averages of unbiased estimators. *Biometrika*, **104**, 727–734.

Sherlock, C., Thiery, A. H., Roberts, G. O. & Rosenthal, J. S. (2015) On the efficiency of pseudo-marginal random walk Metropolis algorithms. *Ann. Statist.*, **43**, 238–275.

Accepted Article

This article is protected by copyright. All rights reserved.

Singh, S. S., Lindsten, F. & Moulines, E. (2017) Blocking strategies and stability of particle Gibbs samplers. *Biometrika*, **104**, 953–969.

Sokal, A. D. (1996) Monte Carlo methods in statistical mechanics: Foundations and new algorithms. *Lecture notes*.

Tan, A., Doss, H. & Hobert, J. P. (2015) Honest importance sampling with multiple Markov chains. *J. Comput. Graph. Statist.*, **24**, 792–826.

Tierney, L. (1994) Markov chains for exploring posterior distributions. *Ann. Statist.*, **22**, 1701–1728.

Tran, M.-N., Scharth, M., Pitt, M. K. & Kohn, R. (2014) Importance sampling squared for Bayesian inference in latent variable models. *Preprint*, arXiv:1309.3339v3.

Vihola, M. (2012) Robust adaptive Metropolis algorithm with coerced acceptance rate. *Statist. Comput.*, **22**, 997–1008.

— (2017) Unbiased estimators and multilevel Monte Carlo. *Oper. Res.*, **66**, 448–462.

Vihola, M. & Franks, J. (2020) On the use of approximate Bayesian computation Markov chain Monte Carlo with inflated tolerance and post-correction. *Biometrika*, **107**, 381–395.

Vihola, M., Helske, J. & Franks, J. (2017) Importance sampling type estimators based on approximate marginal Markov chain Monte Carlo. *Preprint*, arXiv:1609.02541v3.

Whiteley, N. (2013) Stability properties of some particle filters. *Ann. Appl. Probab.*, **23**, 2500–2537.

Whiteley, N. & Lee, A. (2014) Twisted particle filters. *Ann. Statist.*, **42**, 115–141.

Wilkinson, D. J. (2005) Parallel Bayesian computation. In *Handbook of Parallel Computing and Statistics* (ed. E. J. Kontoghiorghes), 481–512. Chapman & Hall/CRC.

Accepted Article

A | PROPERTIES OF AUGMENTED MARKOV CHAINS

Throughout this section, suppose that K is a Markov kernel on X and Q is a kernel from X to a space S. We consider here properties of an augmented Markov kernel $\check{K}$ defined on $\mathsf{X} \times \mathsf{S}$ as follows:

$$\check{K}\big((x,s), \mathrm{d}x' \times \mathrm{d}s'\big) := K(x, \mathrm{d}x')Q(x', \mathrm{d}s').$$

We first state the following basic result.

Lemma 30 *The properties of K and the augmented chain $\check{K}$ are related as follows:*

(i) *Let* $\mathrm{irr}(K)$ *denote the set of ϕ-irreducibility measures of a Markov kernel K, then*
- $\varphi_K \in \mathrm{irr}(K) \implies \varphi_{\check{K}}(\mathrm{d}x \times \mathrm{d}s) := \varphi_K(\mathrm{d}x)Q(x, \mathrm{d}s) \in \mathrm{irr}(\check{K})$,
- $\varphi_{\check{K}} \in \mathrm{irr}(\check{K}) \implies \varphi_K(\mathrm{d}x) := \varphi_{\check{K}}(\mathrm{d}x \times \mathsf{S}) \in \mathrm{irr}(K)$.

(ii) *The implications in* ((i)) *hold when* $\mathrm{irr}(K)$ *and* $\mathrm{irr}(\check{K})$ *are replaced with sets of maximal irreducibility measures of K and $\check{K}$, respectively.*

(iii) *The invariant probabilities of K and $\check{K}$ satisfy:*
- $\nu K = \nu \implies \check{\nu}\check{K} = \check{\nu}$ *where* $\check{\nu}(\mathrm{d}x \times \mathrm{d}y) := \nu(\mathrm{d}x)Q(x, \mathrm{d}y)$,
- $\check{\nu}\check{K} = \check{\nu} \implies \nu K = \nu$ *where* $\nu(\mathrm{d}x) := \check{\nu}(\mathrm{d}x \times \mathsf{S})$.

These implications hold also with invariance replaced by reversibility.

This article is protected by copyright. All rights reserved.

(iv) *K is Harris recurrent if and only if $\check{K}$ is Harris recurrent.*

(v) *Suppose $h : \mathsf{X} \times \mathsf{S} \to \mathbb{R}$ is measurable and such that $m_h(x) := \int Q(x, \mathrm{d}s)h(x, s)$ and $(K^n m_h)(x)$ are well-defined. Then, for any $n \geq 1$ and $s \in \mathsf{S}$, $(\check{K}^n h)(x, s) = (K^n m_h)(x)$.*

Proof The inheritance of irreducibility measures (i), maximal irreducibility measures (ii), invariant measures (iii), and reversibility is straightforward.

For Harris recurrence (iv), let the probability ϕ_K be a maximal irreducibility measure for K, then $\phi_{\check{K}}(\mathrm{d}x \times \mathrm{d}s) := \phi_K(\mathrm{d}x)Q(x, \mathrm{d}s)$ is the maximal irreducibility measure for $\check{K}$. Let $C \in \mathcal{X} \otimes \mathcal{S}$ with $\phi_{\check{K}}(C) > 0$, and choose $\epsilon > 0$ such that $\phi_K(C(\epsilon)) > 0$, where $C(\epsilon) := \{x \in \mathsf{X} : Q(x, C_x) > \epsilon\}$ with $C_x := \{s \in \mathsf{S} : (x, s) \in C\}$. Notice that

$$\mathbb{P}\left(\sum_{k=1}^{\infty} \mathbb{I}\left((X_k, S_k) \in C\right) = \infty\right) \geq \mathbb{P}\left(\sum_{k=1}^{\infty} \mathbb{I}\left(S_{\tau_k} \in C_{X_{\tau_k}}\right) = \infty\right),$$

where τ_k are the hitting times of (X_k) to $C(\epsilon)$. This concludes the proof because $\mathbb{I}\left(S_{\tau_k} \in C_{X_{\tau_k}}\right)$ are independent Bernoulli random variables with success probability at least ϵ. The converse statement is similar.

For (v), it is enough to notice that for any $(x, s) \in \mathsf{X} \times \mathsf{S}$ and $n \geq 1$, it holds that $\check{K}^n\left((x, s), \mathrm{d}x' \times \mathrm{d}s'\right) = K^n(x, \mathrm{d}x')Q(x', \mathrm{d}s)$.

We next state the following generic results about the asymptotic variance and the central limit theorem of an augmented Markov chain. For $h \in L_0^2(\check{\nu})$, we denote as above the conditional mean $m_h(x) := \int Q(x, \mathrm{d}s)h(x, s)$ and the conditional variance $v_h(x) := \int Q(x, \mathrm{d}s)h^2(x, s) - m_h^2(x)$.

Lemma 31 *Let $h \in L_0^2(\check{\nu})$. The asymptotic variance of an augmented Markov chain satisfies*

$$\mathrm{Var}(h, \check{K}) = \mathrm{Var}(m_h, K) + \nu(v_h),$$

whenever $\mathrm{Var}(m_h, K)$ *is well-defined.*

Proof Let (X_k, S_k) be a stationary Markov chain with transition probability $\check{K}$.

$$\mathrm{Var}\left(\frac{1}{\sqrt{n}} \sum_{k=1}^{n} h(X_k, S_k)\right) = \check{\nu}(h^2) + \frac{2}{n} \sum_{i=1}^{n-1} \sum_{\ell=1}^{n-i} \mathbb{E}[h(X_0, S_0)h(X_\ell, S_\ell)],$$

by stationarity. For $\ell \geq 1$, Lemma 30 (v) implies

$$\mathbb{E}[h(X_0, S_0)h(X_\ell, S_\ell)] = \mathbb{E}[m_h(X_0)m_h(X_\ell)].$$

We deduce for any $n \geq 1$

$$\mathrm{Var}\left(\frac{1}{\sqrt{n}} \sum_{k=1}^{n} h(X_k, S_k)\right) = \mathrm{Var}\left(\frac{1}{\sqrt{n}} \sum_{k=1}^{n} m_h(X_k)\right) + \nu(v_h),$$

because $\check{\nu}(h^2) - \nu(m_h^2) = \nu(v_h)$. The claim follows by taking limit $n \to \infty$.

Accepted Article

This article is protected by copyright. All rights reserved.

Lemma 32 *Suppose K is Harris ergodic and aperiodic, and $h \in L^2_0(\check{\nu})$. The CLT*

$$Z_n := \frac{1}{\sqrt{n}} \sum_{k=1}^{n} h(X_k, S_k) \xrightarrow{n\to\infty} N\left(0, \sigma_h^2\right), \qquad \textit{where} \qquad \sigma_h^2 := \mathrm{Var}(m_h, K) + \nu(v_h) \tag{10}$$

holds for every initial distribution, if one of the following holds:

(i) *K is reversible and* $\mathrm{Var}(m_h, K) < \infty$.
(ii) $\sum_{n=1}^{\infty} n^{-3/2} \left\{ \nu\left(\left[\sum_{i=0}^{n-1} K^i m_h \right]^2 \right) \right\}^{1/2} < \infty$.
(iii) *There exists $g \in L^2(\nu)$ which solves the Poisson equation $g - Kg = m_h$. In this case,* $\mathrm{Var}(m_h, K) = \nu(g^2 - (Kg)^2)$.

Proof The case (i) follows from Lemma 31 and the CLT of Kipnis & Varadhan (1986), which implies (10) for the initial distribution $\check{\nu}$. Because the jump chain is Harris by Lemma 30 (iv), (10) holds for every initial distribution (Douc et al., 2018, Corollary 21.1.6).

The case (ii) follows similarly, but relies on a result due to Maxwell & Woodroofe (2000), which implies (10) for the initial distribution $\check{\nu}$, if $\sum_{n=1}^{\infty} n^{-3/2} \left\{ \check{\nu}\left(\left[\sum_{i=0}^{n-1} \check{K}^i h \right]^2 \right) \right\}^{1/2} < \infty$. Notice that for $n \geq 2$ by Lemma 30 (v),

$$\check{\nu}\left(\left[\sum_{i=0}^{n-1} \check{K}^i h \right]^2 \right) = \check{\nu}\left(\left[(h - m_h) + \sum_{i=0}^{n-1} K^i m_h \right]^2 \right) = \nu(v_h) + \nu\left(\left[\sum_{i=0}^{n-1} K^i m_h \right]^2 \right).$$

Because $(a+b)^{1/2} \leq a^{1/2} + b^{1/2}$ for $a, b \geq 0$ and $\nu(v_h) < \infty$, the claim follows.

For (iii), we first observe that

$$\check{g} - \check{K}\check{g} = h \qquad \text{where} \qquad \check{g}(x, s) := g(x) + h(x, s) - m_h(x) \in L^2(\check{\nu}).$$

Indeed, it is clear that $\check{g} \in L^2(\check{\nu})$ and because $(\check{K}\check{g})(x, s) = (Kg)(x)$,

$$\check{g}(x, s) - (\check{K}\check{g})(x, s) = g(x) - (Kg)(x) + h(x, s) - m_h(x) = h(x, s).$$

The CLT and asymptotic variance follow from (Meyn & Tweedie, 2009, Theorem 17.4.4).

B | PROOFS ABOUT CLT AND ASYMPTOTIC VARIANCE

Proof of Theorem 12 Whenever $\sum_{i=1}^{n} \xi_i(\mathbf{1}) > 0$, we may write

$$\sqrt{n}\left[E_n(f) - \pi(f)\right] = \frac{n^{-1/2} \sum_{k=1}^{n} \xi_k(\bar{f})}{n^{-1} \sum_{j=1}^{n} \xi_j(\mathbf{1})}. \tag{11}$$

The denominator converges to $c_w > 0$ almost surely, so by Slutsky's lemma, it is enough to show that the numerator converges in distribution to $N\left(0, \mathrm{Var}(\nu_{\bar{f}}, P) + \pi_a(v)\right)$. This follows from Lemma 32 (i) and (ii), under conditions (i) and (ii), respectively.

Proof of Corollary 14 If $\sqrt{n}\left[E_n(f) - \pi(f)\right] \to N(0, \sigma_f^2)$, Slutsky's lemma applied to (11) implies that $n^{-1/2} \sum_{k=1}^{n} \xi_k(\bar{f}) \to N(0, c_w^2 \sigma_f^2)$. Consistency of the integrated autocovariance estimator implies $\hat{a}_n \to \sum_{k=-\infty}^{\infty} \gamma_k = c_w^2 \sigma_f^2$, and therefore $\hat{s}_n \to \sigma_f^2$. The conclusion follows by a standard continuity argument.

Accepted Article

This article is protected by copyright. All rights reserved.

Proof of Theorem 15 For n large enough such that $\sum_{j=1}^{n} \xi_j(\mathbf{1}) > 0$, we may write

$$n\hat{v}_n = \frac{\frac{1}{n}\sum_{k=1}^{n}\left(\xi_k(f) - \xi_k(\mathbf{1})E_n(f)\right)^2}{\left(\frac{1}{n}\sum_{j=1}^{n}\xi_j(\mathbf{1})\right)^2}.$$

The denominator converges to c_w^2, and the numerator can be written as

$$\frac{1}{n}\sum_{k=1}^{n}\left[\xi_k^2(\bar{f}) + \xi_k^2(\mathbf{1})D_n^2 + 2\xi_k(\mathbf{1})\xi_k(\bar{f})D_n\right] \qquad \text{with} \qquad D_n := \pi(f) - E_n(f).$$

The term $n^{-1}\sum_{k=1}^{n}\xi_k^2(\bar{f}) \to \pi_a(v + \mu_{\bar{f}}^2)$, and because $D_n \to 0$, the remainder terms tend to zero.

C | PROOFS ABOUT JUMP CHAIN ESTIMATORS

In this section, K is assumed to be a Markov kernel on X which is non-degenerate, that is, $a(x) := K(x, \mathsf{X} \setminus \{x\}) > 0$ for all $x \in \mathsf{X}$. The following proposition complements (Douc & Robert, 2011, Lemma 1) and (Deligiannidis & Lee, 2018), which are stated for more specific cases.

Proposition 33 *Suppose (X_k) is a Markov chain with kernel K and $(\tilde{X}_k)$ the corresponding jump chain with holding times (N_k) (Definition 17). Then, the following hold:*

(i) *$(\tilde{X}_k)$ is Markov with transition kernel $\tilde{K}(x, A) = K(x, A \setminus \{x\})/a(x)$.*
(ii) *The holding times (N_k) are conditionally independent given $(\tilde{X}_k)$, and each N_k has geometric distribution with parameter $a(\tilde{X}_k)$.*
(iii) *If K admits invariant probability $\nu(\mathrm{d}x)$, then $\tilde{K}$ admits invariant probability $\tilde{\nu}(\mathrm{d}x) := \nu(\mathrm{d}x)a(x)/\nu(a)$. In addition, if K is reversible with respect to ν, then $\tilde{K}$ is reversible with respect to $\tilde{\nu}$.*
(iv) *(X_k) is ψ-irreducible if and only if $(\tilde{X}_k)$ is ψ-irreducible, with the same maximal irreducibility measure.*
(v) *(X_k) is Harris recurrent if and only if $(\tilde{X}_k)$ is Harris recurrent.*

Proof The expression of the kernel (i) is due to straightforward conditioning, and (ii) was observed by Douc & Robert (2011). The invariance (iii) follows from

$$\int \tilde{\nu}(\mathrm{d}x)\tilde{K}(x, A) = \frac{1}{\nu(a)}\int \nu(\mathrm{d}x)\left[K(x, A) - \mathbb{I}(x \in A)\,K(x, \{x\})\right] = \frac{1}{\nu(a)}\left[\nu(A) - \int_A \nu(\mathrm{d}x)\big(1 - a(x)\big)\right] = \tilde{\nu}(A),$$

and the reversibility is shown in (Douc & Robert, 2011). For (iv) it is sufficient to observe that

$$\forall x \in \mathsf{X} : \sum_{n\geq 1}\mathbb{P}_x(X_n \in A) > 0 \iff \forall x \in \mathsf{X} : \sum_{n\geq 1}\mathbb{P}_x(\tilde{X}_n \in A) > 0,$$

where $\mathbb{P}_x(\,\cdot\,) = \mathbb{P}(\,\cdot \mid X_0 = \tilde{X}_0 = x)$, which holds because the sets $\{X_k\}_{k\geq 0}$ and $\{\tilde{X}_k\}_{k\geq 0}$ coincide. Similarly, (v) holds because

$$\forall x \in \mathsf{X} : \mathbb{P}_x(\eta_A = \infty) = 1 \iff \forall x \in \mathsf{X} : \mathbb{P}_x(\tilde{\eta}_A = \infty) = 1,$$

Accepted Article

This article is protected by copyright. All rights reserved.

where $\eta_A := \sum_{k=1}^{\infty} \mathbb{1}\left(X_k \in A\right)$ and $\tilde{\eta}_A := \sum_{k=1}^{\infty} \mathbb{1}(\tilde{X}_k \in A)$.

We now state results about the asymptotic variance of the jump chain, complementing the reversible case characterisation of Deligiannidis & Lee (2018) and Doucet et al. (2015).

Proposition 34 *Let $f \in L_0^2(\tilde{\nu})$. With the notation of Proposition 33,*

(i) *If K is reversible, then $\mathrm{Var}(f, \tilde{K}) < \infty$ iff $af \in L^2(\nu)$ and $\mathrm{Var}(af, K) < \infty$, and*

$$\mathrm{Var}(f, \tilde{K}) = \nu(a)^{-1}\left[\mathrm{Var}(af, K) - \nu\left(a(1-a)f^2\right)\right]. \tag{12}$$

(ii) *If there exists a function $g \in L^2(\nu)$ which satisfies $g - Kg = af$, then $\mathrm{Var}(f, \tilde{K}) < \infty$, $\mathrm{Var}(af, K) < \infty$, (12) holds, $g - \tilde{K}g = f$ and $g \in L^2(\tilde{\nu})$.*

Proof The reversible case (i) is a restatement of (Deligiannidis & Lee, 2018, Theorem 1).

Consider then (ii). By Proposition 33 (i), we obtain for any $h : \mathsf{X} \to \mathbb{R}$ with Kh well-defined,

$$(\tilde{K}h)(x) = \frac{(Kh)(x) - \left(1 - a(x)\right)h(x)}{a(x)} = \frac{(Kh)(x) - h(x)}{a(x)} + h(x).$$

Consequently, we observe that $g - \tilde{K}g = a^{-1}\left(g - Kg\right) = f$ implying (ii). Because $g \in L^2(\tilde{\nu})$, Lemma 32 (iii) and a straightforward calculation yield

$$\mathrm{Var}(f, \tilde{K}) = \tilde{\nu}\left(g^2 - (\tilde{K}g)^2\right) = 2\langle g, g - \tilde{K}g\rangle_{\tilde{\nu}} - \langle g - \tilde{K}g, g - \tilde{K}g\rangle_{\tilde{\nu}} = \nu(a)^{-1}\left[2\langle g, g - Kg\rangle_{\nu} - \nu(af^2)\right],$$

where $\langle f, g\rangle_{\nu} := \int f(x)g(x)\nu(\mathrm{d}x)$. Similarly, by Lemma 32 (iii)

$$\mathrm{Var}(af, K) = \nu\left(g^2 - (Kg)^2\right) = 2\langle g, g - Kg\rangle_{\nu} - \nu(a^2 f^2),$$

which allows us to conclude.

Proof of Theorem 22 Whenever $\sum_{j=1}^{n} \xi_j(\mathbf{1}) > 0$, we may write

$$\sqrt{n}\left[E_n(f) - \pi(f)\right] = \frac{n^{-1/2}\sum_{k=1}^{n} N_k \xi_k(\bar{f})}{n^{-1}\sum_{j=1}^{n} N_j \xi_j(\mathbf{1})}.$$

We shall show below that the CLT holds for the numerator, with asymptotic variance $\sigma^2 := \left[\mathrm{Var}(\mu_{\bar{f}}, P) + \pi_a(\alpha\tilde{v})\right]/\pi_a(\alpha)$. This implies the claim by Slutsky's lemma, as the denominator converges to $c_w/\pi_a(\alpha)$. For the rest of the proof, let $\tilde{P}$ and $\check{P}$ be the Markov kernels of $(\tilde{\Theta}_k)_{k\geq 1}$ and $(\tilde{\Theta}_k, N_k, \xi_k(\bar{f}))_{k\geq 1}$, respectively, and let $\tilde{\pi}$ and $\check{\pi}$ be the corresponding invariant probabilities. Note that the function $h(\theta, n, \xi) := n\xi$ is in $L^2(\check{\pi})$ by assumption (6).

In case (i) holds, also $\tilde{P}$ and $\check{P}$ are reversible by Proposition 33 (iii) and Lemma 30 (iii). Lemma 32 (i) with $K = \tilde{P}$, $\check{K} = \check{P}$, $\nu = \tilde{\pi}$ and $\check{\nu} = \check{\pi}$ implies that a CLT holds for h whenever the asymptotic variance is finite:

$$\mathrm{Var}(h, \check{P}) = \mathrm{Var}\left(\mu_{\bar{f}}/\alpha, \tilde{P}\right) + \pi_a(\alpha\tilde{v}_{N\xi})/\pi_a(\alpha),$$

Accepted Article

This article is protected by copyright. All rights reserved.

where, by the variance decomposition formula,

$$\tilde{v}_{N\xi}(\theta) := \mathrm{Var}(N_k \xi_k(\bar{f}) \mid \tilde{\Theta}_k = \theta) = \tilde{v}(\theta) + \mathrm{Var}\big(N_k \mathbb{E}[\xi_k(\bar{f}) \mid \tilde{\Theta}_k = \theta, N_k] \mid \tilde{\Theta}_k = \theta\big) = \tilde{v}(\theta) + \mu_{\bar{f}}^2(\theta)\big(1 - \alpha(\theta)\big)/\alpha^2(\theta).$$

Proposition 34 (i) implies that

$$\mathrm{Var}\big(\mu_{\bar{f}}/\alpha, \tilde{P}\big) = \pi_a(\alpha)^{-1}\big[\mathrm{Var}(\mu_{\bar{f}}, P) - \pi_a\big((1-\alpha)\mu_{\bar{f}}^2/\alpha\big)\big],$$

which implies $\mathrm{Var}(h, \check{P}) = \sigma^2$.

Consider then (ii). Proposition 34 (ii) implies that $g - \tilde{P}g = \mu_{\bar{f}}/\alpha$, and $g \in L^2(\tilde{\pi})$. Lemma 32 (iii) implies the CLT, and together with Proposition 34 (ii) leads to $\mathrm{Var}(h, \check{P}) = \sigma^2$.

D | PROPER WEIGHTINGS FOR GENERAL STATE SPACE MODELS

We review some techniques to construct proper weightings in case of general state-space models introduced in Section 8. First, note that the simple IS correction may be applied directly (see Proposition 8). Note that (8) is satisfied for all integrable h, so $\mathcal{L} = L^1(\pi)$. It is often useful to combine such schemes as in Proposition 10, allowing for instance variance reduction by using pairs of antithetic variables (Durbin & Koopman, 2000).

For the rest of the section, we focus on the particle filter (PF) algorithm (Gordon et al., 1993); see also the monographs (Doucet et al., 2001; Del Moral, 2004; Cappé et al., 2005). We consider a generic version of the algorithm, with the following components (cf. Del Moral, 2004):

1. Proposal distributions: M_1 is a probability density on S_z and $M_t(\cdot \mid z_{1:t-1})$ defines conditional densities on S_z given $z_{1:t-1} \in \mathsf{S}_z^{t-1}$.
2. Potential functions: $G_t : \mathsf{S}_z^t \to \mathbb{R}_+$.
3. Resampling laws: $\mathrm{Res}(\cdot \mid \bar{\omega}^{(1:m)})$ defines a probability distribution on $\{1:m\}^m$ for every discrete probability mass $\bar{\omega}^{(1:m)}$.

The following well-known two conditions are minimal to ensure unbiasedness, which is required for proper weighting:

Assumption 35 *Suppose that the following hold:*

(i) $\prod_{t=1}^T M_t(z_t \mid z_{1:t-1}) G_t(z_{1:t}) = p^{(\theta)}(z_{1:T}, y_{1:T})$ *for all* $z_{1:T} \in \mathsf{S}_z^T$.
(ii) $\mathbb{E}\big[\sum_{i=1}^m \mathbb{I}(A^{(i)} = j)\big] = m\bar{\omega}^{(j)}$, *where* $A^{(1:m)} \sim \mathrm{Res}(\cdot \mid \bar{\omega}^{(1:m)})$, *for any* $j \in \{1:m\}$ *and any probability mass vector* $\bar{\omega}^{(1:m)}$.

Assumption 35 (i) holds with traditionally used 'filtering' potentials $G_t(z_{1:t}) := g_t^{(\theta)}(y_t \mid z_t)\mu_t^{(\theta)}(z_t \mid z_{t-1})/M_t(z_t \mid z_{1:t-1})$, assuming a suitable support condition. We discuss another choice of M_t and G_t in Section 8, inspired by the 'twisted SSM' approach of Guarniero et al. (2017). It allows a 'look-ahead' strategy based on approximations of the full smoothing distributions $q^{(\theta)}(z_{1:T} \mid y_{1:T})$. Assumption 35 (ii) allows for multinomial resampling, where $A_t^{(i)}$ are independent draws from $\bar{\omega}_t^{(1:m)}$, but also for lower variance schemes, including stratified, residual and systematic resampling methods (cf. Douc et al., 2005).

Below, whenever the index 'i' appears, it takes values $i = 1, \dots, m$.

Algorithm 36 (Particle filter) *Initial state:*

(i) *Sample* $Z_1^{(i)} \sim M_1$ *and set* $\bar{Z}_1^{(i)} = Z_1^{(i)}$.

Accepted Article

This article is protected by copyright. All rights reserved.

(ii) *Calculate* $\omega_1^{(i)} := G_1(Z_1^{(i)})$ *and set* $\bar{\omega}_1^{(i)} := \omega_1^{(i)}/\omega_1^*$ *where* $\omega_1^* = \sum_{j=1}^m \omega_1^{(j)}$.

For $t = 2, \dots, T$*, do:*

(iii) *Sample* $A_{t-1}^{(1:m)} \sim \mathrm{Res}(\cdot \mid \bar{\omega}_{t-1}^{(1:m)})$.

(iv) *Sample* $Z_t^{(i)} \sim M_t(\cdot \mid \bar{Z}_{t-1}^{(A_{t-1}^{(i)})})$ *and set* $\bar{Z}_t^{(i)} = (\bar{Z}_{t-1}^{(A_{t-1}^{(i)})}, Z_t^{(i)})$.

(v) *Calculate* $\omega_t^{(i)} := G_t(\bar{Z}_{t-1}^{(A_{t-1}^{(i)})}, Z_t)$ *and set* $\bar{\omega}_t^{(i)} := \omega_t^{(i)}/\omega_t^*$ *where* $\omega_t^* = \sum_{j=1}^m \omega_t^{(j)}$.

Remark 37 *If all weights are zero,* $\omega_t^* = 0$*, then Algorithm 36 may be terminated immediately (cf. Proposition 38).*

The following result summarises alternative ways how the random variables $(V_\theta^{(1:m)}, X_\theta^{(1:m)})$ may be constructed from the PF output, in order to satisfy (8). The results stated below are gathered from the literature (e.g. Del Moral, 2004; Pitt et al., 2012), and some may be stated under slightly more stringent conditions; a self-contained proof of Proposition 38 may be found, for instance, in (Vihola et al., 2017).

Proposition 38 *Let* $\theta \in \mathsf{T}$ *be fixed, assume* Res*,* M_t *and* G_t *satisfy Assumption 35, and let* $h : \mathsf{S}_z^T \to \mathbb{R}$ *be such that the integral in* (8) *is well-defined and finite. Consider the random variables generated by Algorithm 36, and let* $U := \prod_{t=1}^T \left(\frac{1}{m}\omega_t^*\right)$*. Then,*

(i) *the random variables* $(V_\theta^{(1:m)}, X_\theta^{(1:m)})$ *where* $V_\theta^{(i)} = U\bar{\omega}_T^{(i)}$ *and* $X_\theta^{(i)} = \bar{Z}_T^{(i)}$ *satisfy* (8).

Suppose in addition that $M_t(z_t \mid z_{1:t-1})G_t(z_{1:t}) = C_t(z_{t-1:t})$ *for all* $t \in \{1:T\}$ *and all* $z_{1:T} \in \mathsf{S}_z^T$*. Define for* $t \in \{2:T\}$*, and any* $i_t, i_{t-1} \in \{1:m\}$*, the backwards sampling probabilities*

$$b_{t-1}(i_{t-1} \mid i_t) := \frac{\bar{\omega}_{t-1}^{(i_{t-1})} C_t(Z_{t-1}^{(i_{t-1})}, Z_t^{(i_t)})}{\sum_{\ell=1}^m \bar{\omega}_{t-1}^{(\ell)} C_t(Z_{t-1}^{(\ell)}, Z_t^{(i_t)})}, \quad \text{and} \quad b_T(i_T \mid i_{T+1}) = \bar{\omega}_T^{(i_T)}.$$

(ii) *Let* $I_{1:T}$ *be random indices generated recursively backwards by* $I_T \sim b_T$ *and* $I_t \sim b_t(\cdot \mid I_{t+1})$*. The random variables* $(V_\theta^{(1)}, X_\theta^{(1)})$ *satisfy* (8)*, where* $V_\theta^{(1)} = U$ *and* $X_\theta^{(1)} = Z_{1:T}^{(I_{1:T})}$.

(iii) *If* $h(z_{1:T}) = \hat{h}(z_{t-1}, z_t)$ *for some* $t \in \{2:T\}$*, that is,* h *is constant in all coordinates except* $t-1$ *and* t*, then, the random variables* $(V_\theta^{(1:m,1:m)}, X_\theta^{(1:m,1:m)})$ *satisfy* (8) *(with* $\hat{h}$ *on the left), where*

a. $X_\theta^{(i,j)} := (Z_{t-1}^{(i)}, Z_t^{(j)})$,

b. $V_\theta^{(i,j)} := U b_{t-1}(i \mid j)\hat{\omega}_t^{(j)}$*, and where*

c. $\hat{\omega}_T^{(i)} := \bar{\omega}_T^{(i)}$ *and* $\hat{\omega}_t^{(i)} := \sum_{k=1}^m \hat{\omega}_{t+1}^{(k)} b_t(i \mid k)$ *for* $t = T-1, \dots, t$.

(iv) *If* $h(z_{1:T}) = \hat{h}(z_t)$ *for some* $t \in \{1:T\}$*, then the random variables* $(V_\theta^{(1:m)}, X_\theta^{(1:m)})$ *satisfy* (8) *(with* $\hat{h}$ *on the left), where* $X_\theta^{(i)} = Z_t^{(i)}$ *and* $V_\theta^{(i)} = U\hat{\omega}_t^{(i)}$ *are defined in c..*

The estimator in Proposition 38 (i) was called the filter-smoother by Kitagawa (1996). This property was shown in (Del Moral, 2004, Theorem 7.4.2) in case of multinomial resampling, and extended later (cf. Andrieu et al., 2010). Proposition 38 (ii) corresponds to backwards simulation smoothing (Godsill et al., 2004). Drawing a single backward trajectory does not improve on the filter-smoother (cf. Doucet & Lee, 2019), but drawing several $I_{1:T}$ independently may lead to lower variance estimators. Proposition 38 (iii) and its special case (iv) correspond to the forward-backward smoother (Doucet et al., 2000); see also (Cappé et al., 2005). It is a Rao-Blackwellised version of (ii), but applicable only when considering estimates of a single marginal (pair). This scheme can lead to lower variance, but suffers from $O(m^2)$ complexity.

Accepted Article

This article is protected by copyright. All rights reserved.

We next formally state how Proposition 38 allows to use Algorithm 36 to derive a proper weighting scheme.

Corollary 39 *Let $(\Theta_k)_{k\geq 1}$ be a Markov chain which is Harris ergodic with respect to π_a. Suppose each $(V_k^{(1:m)}, X_k^{(1:m)})$ corresponds to an independent run of Algorithm 36 with $\theta = \Theta_k$, as defined in Proposition 38 (i), (ii), (iii) or (iv). Then, $(W_k^{(1:m)}, X_k^{(1:m)})_{k\geq 1}$ with $W_k^{(i)} := \mathrm{pr}(\theta_k)V_k^{(i)}/\pi_a(\theta_k)$ provide a proper weighting scheme for target distribution $\pi(\theta, x_{1:T}) = p(\theta, x_{1:T} \mid y_{1:T})$ (Definition 5), for the following classes of functions, respectively:*

$$(i) \quad \mathcal{L} = L^1(\pi), \qquad (iii) \quad \mathcal{L} = \left\{ f \in L^1(\pi) \,:\, f(\theta, x_{1:T}) = f(\theta, x_{t-1:t}), \text{ for some } t \in \{2{:}T\} \right\},$$

$$(ii) \quad \mathcal{L} = L^1(\pi), \qquad (iv) \quad \mathcal{L} = \left\{ f \in L^1(\pi) \,:\, f(\theta, x_{1:T}) = f(\theta, x_t), \text{ for some } t \in \{1{:}T\} \right\}.$$

In case $(\Theta_k, U_k)_{k\geq 1}$ is a pseudo-marginal algorithm, $W_k := \mathrm{pr}(\theta_k)V_k^{(i)}/U_k$.

Accepted Article

This article is protected by copyright. All rights reserved.