JYU DISSERTATIONS 79

Jordan Franks

Markov Chain Monte Carlo Importance Samplers for Bayesian Models with Intractable Likelihoods

UNIVERSITY OF JYVÄSKYLÄ
FACULTY OF MATHEMATICS
AND SCIENCE

JYU DISSERTATIONS 79

Jordan Franks

Markov Chain Monte Carlo Importance Samplers for Bayesian Models with Intractable Likelihoods

Esitetään Jyväskylän yliopiston matemaattis-luonnontieteellisen tiedekunnan suostumuksella
julkisesti tarkastettavaksi yliopiston Historica-rakennuksen salissa H320
toukokuun 4. päivänä 2019 kello 12.

Academic dissertation to be publicly discussed, by permission of
the Faculty of Mathematics and Science of the University of Jyväskylä,
in building Historica, auditorium H320, on May 4, 2019 at 12 o'clock noon.

JYVÄSKYLÄN YLIOPISTO
UNIVERSITY OF JYVÄSKYLÄ

JYVÄSKYLÄ 2019

Editors:

Matti Vihola
Department of Mathematics and Statistics
University of Jyväskylä
Finland

Timo Hautala
Open Science Centre
University of Jyväskylä
Finland

Reviewers:

Marko Laine
Meteorological Research
Finnish Meteorological Institute
Finland

Krzysztof Latuszynski
Department of Statistics
University of Warwick
United Kingdom

Opponent:

Nicolas Chopin
Department of Statistics
ENSAE ParisTech
France

Permanent link to this publication: http://urn.fi/URN:ISBN:978-951-39-7738-2

ISBN 978-951-39-7738-2 (PDF)
URN:ISBN:978-951-39-7738-2
ISSN 2489-9003 Jyväskylä University Printing House, Jyväskylä 2019

Abstract

Markov chain Monte Carlo (MCMC) is an approach to parameter inference in Bayesian models that is based on computing ergodic averages formed from a Markov chain targeting the Bayesian posterior probability. We consider the efficient use of an approximation within the Markov chain, with subsequent importance sampling (IS) correction of the Markov chain inexact output, leading to asymptotically exact inference. We detail convergence and central limit theorems for the resulting MCMC-IS estimators. We also consider the case where the approximate Markov chain is pseudo-marginal, requiring unbiased estimators for its approximate marginal target. Convergence results with asymptotic variance formulae are shown for this case, and for the case where the IS weights based on unbiased estimators are only calculated for distinct output samples of the so-called 'jump' chain, which, with a suitable reweighting, allows for improved efficiency. As the IS type weights may assume negative values, extended classes of unbiased estimators may be used for the IS type correction, such as those obtained from randomised multilevel Monte Carlo. Using Euler approximations and coupling of particle filters, we apply the resulting estimator using randomised weights to the problem of parameter inference for partially observed Itô diffusions. Convergence of the estimator is verified to hold under regularity assumptions which do not require that the diffusion can be simulated exactly. In the context of approximate Bayesian computation (ABC), we suggest an adaptive MCMC approach to deal with the selection of a suitably large tolerance, with IS correction possible to finer tolerance, and with provided approximate confidence intervals. A prominent question is the efficiency of MCMC-IS compared to standard direct MCMC, such as pseudo-marginal, delayed acceptance, and ABC-MCMC. We provide a comparison criterion which generalises the covariance ordering to the IS setting. We give an asymptotic variance bound relating MCMC-IS with the latter chains, as long as the ratio of the true likelihood to the approximate likelihood is bounded. We also perform various experiments in the state space model and ABC context, which confirm the validity and competitiveness of the suggested MCMC-IS estimators in practice.

Contents

Abstract i
List of symbols iii
Foreword iv
List of included articles vi
1. Introduction 1
1.1. Likelihoods 1
1.2. Bayesian inference 2
1.3. Challenges for inference 2
1.4. Approximate families 3
1.5. Overview 3
2. Bayesian inference for state space models on path space 4
2.1. Discretely-observed continuous-time path-dependent process 4
2.2. Model probabilities 5
2.3. Particle filter 6
2.4. Particle marginal Metropolis-Hastings 8
3. Accelerations based on an approximation 9
3.1. Delayed acceptance and importance sampling 9
3.2. The question of relative efficiency 11
3.3. Peskun and covariance orderings of asymptotic variances 11
3.4. Peskun type ordering for importance sampling correction 13
3.5. Comparison results 13
4. Bayesian inference for state space models with diffusion dynamics 15
4.1. Euler approximations 16
4.2. Multilevel Monte Carlo 16
4.3. Coupling of Feynman-Kac models 17
4.4. Debiasing techniques 18
4.5. Joint inference using importance sampling type correction 18
4.6. Computational efficiency and allocations 20
5. Inference via approximate Bayesian computation 21
5.1. Choosing the tolerance in ABC-MCMC 22
5.2. Approximate confidence intervals 22
5.3. Adaptive ABC-MCMC with post-correction 23
5.4. Convergence of the tolerance adaptive ABC-MCMC 24
6. Discussion and directions for future work 25
7. Summary of articles 27
7.1. Article [A]: Importance sampling type estimators... 27
7.2. Article [B]: Importance sampling correction versus... 27
7.3. Article [C]: Unbiased inference for hidden Markov model... 28
7.4. Article [D]: On the use of ABC-MCMC... 28
References 29
Article [A] 35
Article [B] 71
Article [C] 101
Article [D] 133

List of symbols

Symbol	Page	Meaning
$(\mathsf{Y}, \mathcal{Y}, \mathscr{P})$	1	statistical model
$\mathscr{P}$	1	family of probability distributions
$p^{(\theta)}$	1	data (probability) distribution
$L(\theta)$	2	likelihood of parameter θ
$\mathscr{P}_\ell$	3	ℓ-approximate family of data (probability) distributions
$\mathscr{P}_\infty$	3	ideal family of data (probability) distributions
(M_p, G_p)	5	Feynman-Kac model with transition M_p and potential G_p
$p^{(\theta,\ell)}$	3	data (probability) distribution in $\mathscr{P}_\ell$
	5	ℓ-smoother for Feynman-Kac model $(M_p^{(\theta,\ell)}, G_p^{(\theta)})$
$p_u^{(\theta,\ell)}$	5	unnormalised ℓ-smoother on latent states for model $(M_p^{(\theta,\ell)}, G_p^{(\theta)})$
$\pi^{(\ell)}$	5	joint ℓ-posterior probability on parameters and latent states
$\pi_m^{(\ell)}$	6	marginal ℓ-posterior probability on parameters
p_u	6	unnormalised smoother for model (M_p, G_p)
$\hat{p}_u$	6	unbiased estimator for p_u for model (M_p, G_p)
$\hat{L}(\theta)$	9	unbiased estimator for $L(\theta)$
σ_K^2	11	asymptotic variance for Markov kernel K
$\mathcal{E}_K$	12	Dirichlet form for Markov kernel K
$\Pi^{(0)}$	13	stationary probability of 0-model Markov chain
$\Pi^{(\infty)}$	13	stationary probability of ∞-model Markov chain
w	13	importance sampling weight from $\Pi^{(0)}$ to $\Pi^{(\infty)}$
σ_{IS}^2	14	asymptotic variance of MCMC-IS estimator
$\Delta^{(\theta,\ell)}$	17	delta particle filter unbiased estimator
$(p_\ell)_{\ell\in\mathbb{N}}$	18	probability mass function on $\ell = 1, 2, 3, \ldots$
$\tilde{\Delta}^{(\theta,\ell)}(\phi)$	18	unbiased estimator for $p_u^{(\theta,\ell)}(\phi)$
$\mathscr{C}(m)$	20	total computational cost for m iterations
$\mathscr{M}(\kappa)$	20	realised length of the chain with budget κ
$p^{(\theta,1/\epsilon)}$	21	data (probability) distribution, with ABC tolerance ϵ (or precision $1/\epsilon$)
$L^{(1/\epsilon)}$	22	approximate Bayesian computation (ABC) likelihood with ABC tolerance ϵ (or precision $1/\epsilon$)

Foreword

The modern reliance on probability theory to model the universe and various aspects of life reveals on the one hand our tremendous lack of knowledge and ability to understand and hence predict the workings of the universe with Newtonian precision. On the other hand, the success of probability theory reveals the hidden order of the universe, as well as the significant deductive reasoning capacities of humankind, where from the disorder of incomplete knowledge arises the order of probabilistic laws. Statistics allows us to ascertain to what extent our deductive reasoning is justified by real observation. Statistics acts as the intermediary, allowing dialogue to proceed between our perceived knowledge of the (mechanistic and probabilistic) laws of the universe and of the universe as she presents herself to us in actual fact.

Of utmost importance for the development of statistics has been the increasing computational ability in the computer age [cf. 30]. As the speed of computers increases, so does the potential complexity of problems increase which statistical methods can handle with precision. Considerable interest therefore lies in the development of computational methods which are efficient and able to perform the demanding computational tasks of modern statistics. It is the scientific and humanistic hope for this thesis, that the work will serve to the advancement of human knowledge, and that it will be solely useful to the commendable pursuits of humankind.

As the fields of probability and statistics are intellectually challenging, any small progress in this field is dependent upon a stable, friendly, and stimulating working and living environment. First of all, I would like to thank Dr. Matti Vihola, for being a wonderful adviser, scientist, and person. In the beginning, I knew very little about computational statistics and Monte Carlo methods, but due greatly to Matti's tremendous help and patience, my knowledge and skills in mathematics and statistics has grown considerably. This thesis would not have been possible without his help. The enclosed introductory text has also benefited greatly from his insightful remarks. As his first sole doctoral student, I have one of the early claims to be able to call him mathematico-statistical father.

As for a stable, friendly, and stimulating working and living environment, I would like to thank the University of Jyväskylä and its employees, for being able to pursue my doctoral studies here. The last three years have been enjoyable as a place to work, study, and live. Financial support is gratefully acknowledged from the Academy of Finland ('Exact approximate Monte Carlo methods for complex Bayesian inference,' grants 284513 and 312605, PI: Dr. Matti Vihola).

Sincere thanks to the reviewers, Prof. Marko Laine (Finnish Meteorological Institute) and Prof. Krzysztof Łatuszyński (University of Warwick).

There are many other individual persons whom I should thank for being a help these last few years. As I drew up an account of all the people whom I would like to thank, it became ever-expanding, touching every aspect and time of my life. I simply could not do proper justice to those who have helped me, and I would run the danger of leaving somebody unintentionally out. I therefore would simply like to thank the many precious people who have been a positive impact on me, without going into all the details here. They and there deeds are simply too many to be entrusted to these few pages.

I think the saying is true, and hope it is true: when someone has stayed somewhere long enough, the place becomes forever a part of the person. I wish to thank the

many people in Jyväskylä whom I have enjoyed getting to know during these last few years. Language has not always been an insurmountable barrier. I will miss you, and I will miss Jyväskylä—the snow, the sauna, the summer, the lakes, the festivals, the food, the people—all of which make Finland a special place to live.

I mention regards to the researchers everywhere with whom I have had the privilege to meet. Statistics, like other scientific disciplines such as pure mathematics, involves many devotees interested in a common subject with undesirable distractions kept to a minimum. When immersed in a scientific subject, where validity is judged by logic and observation rather than might or necessity, when one is able to escape the day-to-day absorption of the human condition, then one is able to view the world from a new perspective. One sees like the astronaut, for whom, after seeing the Earth as the single terrestrial mass, life will never be the same.

Jordan Franks
Jyväskylä, Finland
April 4, 2019

List of included articles

The thesis consists of an introductory text (Sections 1-7) and the following articles listed in chronological order of preprint appearance.

References

[A] M. Vihola, J. Helske, and J. Franks. Importance sampling type estimators based on approximate marginal MCMC. Preprint arXiv:1609.02541v5, 2016.

[B] J. Franks and M. Vihola. Importance sampling correction versus standard averages of reversible MCMCs in terms of the asymptotic variance. Preprint arXiv:1706.09873v4, 2017.

[C] J. Franks, A. Jasra, K. J. H. Law and M. Vihola. Unbiased inference for discretely observed hidden Markov model diffusions. Preprint arXiv:1807.10259v4, 2018.

[D] M. Vihola and J. Franks. On the use of ABC-MCMC with inflated tolerance and post-correction. Preprint arXiv:1902.00412, 2019.

In the introduction, the above articles are referred to by letters [A], [B], [C], and [D], whereas other article references cited in the introduction shall be referred to by numbers [1], [2], [3], etc.

The author of this dissertation has actively contributed to the research of the joint articles [A], [B], [C] and [D], and has served as the corresponding author for articles [B] and [C]. In particular, the author contributed some theoretical results for augmented Markov chains for [A], and assisted during the research and preparation of the article. Article [B] is mostly the author's own work, but benefited greatly from the second author's input, for example, regarding the topic of the paper and some technical details. The author was responsible for the efficiency considerations of multilevel Monte Carlo in [C], and for final preparation of the article. The author was responsible for the analysis of the adaptive Markov chain Monte Carlo algorithm in [D].

1. Introduction

Bayesian inference often requires the use of computational simulation methods known as Monte Carlo methods. Monte Carlo methods use probabilistic sampling mechanisms and ensemble averages to estimate expectations, such as those taken with respect to the posterior probability of a Bayesian model. Therefore, in practice on a computer, Monte Carlo methods can be computationally intensive.

A further inferential challenge arises when the likelihood function of the Bayesian model is intractable. In some important settings, it is possible to obtain an unbiased estimator for the likelihood. One such setting is the state space model, where sequential Monte Carlo supplies the unbiased estimator. In settings where unbiased estimators are not possible, approximate Bayesian computation (ABC) may be used, assuming forward generation of the model is possible and a choice of tolerance size has been made. Though only an approximation to the original Bayesian model, the ABC model comes equipped with a straightforward unbiased estimator for its ABC likelihood.

In these two example settings, a Markov chain can be run, allowing for Markov dependence in the samples, as well as the use of the unbiased estimators for the (ABC) likelihood as part of a pseudo-marginal algorithm. As a result, the samples of the Markov chain are drawn asymptotically from the (ABC) posterior, and inference is based on averaging the samples obtained. This computational approach to Bayesian inference is known as Markov chain Monte Carlo (MCMC).

This thesis is concerned with a slightly different approach, namely, where the Markov chain targets an approximate marginal of the (ABC) posterior. The subsequent importance sampling (IS) type correction is performed by a reweighting of the inexact sample output, using the unbiased estimators, which yields asymptotically exact inference for the (ABC) posterior. The use of an approximation for the Markov chain target can be computationally efficient, as can be the parallel calculation of the IS weights on a parallel computer. Some of the resulting MCMC-IS estimators are well-known, but in practice have been used only rarely, in comparison to direct MCMC. In addition, the MCMC-IS approach is shown to offer additional flexibility compared to direct MCMC.

The rest of this Section 1 is laid out as follows. We present some important notions, such as that of a statistical model, likelihood function, and Bayesian model. We briefly describe the general goal of (likelihood-based) parameter inference in statistics, as well as some of the challenges of computation which the thesis seeks to address, specifically inference aided by use of an approximation. Section 1.5 concludes with an outline of the remainder of the text.

1.1. **Likelihoods.**

A *statistical model* $(\mathsf{Y}, \mathcal{Y}, \mathscr{P})$ is composed of an *observational space* Y, together with its σ-algebra of subsets $\mathcal{Y}$, and a set $\mathscr{P}$ of probability distributions on Y [cf. 34]. We assume the family $\mathscr{P}$ is parametrised by a vector of *model parameters* $\theta \in \mathsf{T}$, with $\mathsf{T} \subset \mathbb{R}^{n_\theta}$ for some $n_\theta \geq 1$. That is,

$$\mathscr{P} = \{p^{(\theta)}\}_{\theta \in \mathsf{T}},$$

where $p^{(\theta)}(\mathrm{d}y)$ is a probability on Y, sometimes called the *data distribution*. The probability $p^{(\theta)}$ corresponds to a modeling of the dependency relationship of the observation y, considered as a random variable, on the model parameter θ.

We assume for simplicity in this introduction that $p^{(\theta)}(\mathrm{d}y)$ has a density, also denoted $p^{(\theta)}(y)$, with respect to a σ-finite reference measure on Y. Fixing the observation $y \in \mathsf{Y}$, we define the function

$$L(\theta) := p^{(\theta)}(y),$$

which is known as the *likelihood*. One type of likelihood-based inference for θ is to answer which values of θ maximise $L(\theta)$. This method of inference is known as *maximum likelihood estimation* (MLE) in a statistical model with observation [cf. 14]. In other words, MLE seeks to answer, which probability distribution on Y in $\mathscr{P}$ would most readily give rise to the observation.

1.2. **Bayesian inference.** In practice, MLE is highly dependent on the candidate set $\mathscr{P}$ of probabilities to consider. The set $\mathscr{P}$ could be parametrised by arbitrarily high dimensions of parameters, and is the result of the statistician's modeling of the dependence of the observation y on the model parameter θ. Going further, in light of this arbitrary construction of the set $\mathscr{P}$, the statistician is arguably[1] not out of bounds to specify which θ values are to be considered more probable and with more weight, based on prior knowledge or hypothesis.

This specification, for a statistical model with known observation, leads to the *Bayesian model* [cf. 33]. The *Bayesian model* consists of an assignment of a *prior probability* $\mathrm{pr}(\mathrm{d}\theta)$ to the model parameters, with density also denoted $\mathrm{pr}(\theta)$. Inference for the Bayesian model then consists of quantification of the *posterior probability*

$$\pi(\theta) := p(\theta|y) = \frac{L(\theta)\mathrm{pr}(\theta)}{p(y)}, \tag{1}$$

where the last equality, giving the posterior in terms of the likelihood and prior, is the practically useful formula of Bayes, and $p(y)$ is the *model evidence*, defined by

$$p(y) := \int L(\theta)\mathrm{pr}(\theta)\mathrm{d}\theta.$$

1.3. **Challenges for inference.** In statistical models of practical interest, the likelihood $L(\theta)$ is often *intractable*, meaning that it can not be evaluated pointwise. However, in many settings which we consider, we will see that $L(\theta)$ can be estimated unbiasedly, meaning it is possible to generate a random variable $\hat{L}_\theta$ such that $\mathbb{E}[\hat{L}_\theta] = L(\theta)$. However, construction of a reliable unbiased estimator may be neither directly available, nor efficient.

The posterior $\pi(\theta)$ of the Bayesian model is in general intractable, and can not even be estimated unbiasedly. This is often the case even if the likelihood is tractable, because of the normalisation by the model evidence in (1), which is usually computationally intensive to calculate. In case of intractable likelihood in the Bayesian model, posterior inference is even more of a challenge, and one must usually rely on ergodic averages from Markov chain Monte Carlo (MCMC). Such averages are generally asymptotically exact (i.e. *consistent*) as the number of iterations of the MCMC algorithm increases, in the sense of a law of large numbers. However, MCMC can be computationally expensive to run. It can take hours, days, weeks, or longer, in order to ensure a 'reliable' MCMC estimate, where the level of reliability can be theoretically difficult to justify.

[1]The *frequentist* approach differs from the *Bayesian* approach considered here [cf. 30].

1.4. **Approximate families.** We will see that the use of approximations can help facilitate tractable, efficient and user-friendly inference. Let $\mathscr{P}_\infty$ denote a set of (ideal) model probability distributions on Y. In many cases in practice, it may be desirable to work with a surrogate family of probability distributions $\mathscr{P}_0$. Going further, it may be desirable to work with a family

$$\mathscr{P}_\ell = \{p^{(\theta,\ell)}\}_{\theta\in\mathsf{T}},$$

of data (probability) distributions, with $\ell \in [0,\infty]$ used to indicate families of increasingly 'better' approximations. For example, inference using $\mathscr{P}_\infty$ may be too difficult to achieve or too costly, in which case using an approximate family $\mathscr{P}_\ell$ may be possible instead.

It is conceivably possible to incorporate $\mathscr{P}_\ell$ in a Bayesian inference method, which may lessen the computational cost of the algorithm, while in the end performing inference for $\mathscr{P}_\infty$. The aim of this thesis is to show general strategies in different settings where this is possible.

1.5. **Overview.** We now outline the remainder of this text[2]. The text seeks to serve as an introduction and summary for the thesis papers listed on page vi. Methodological aspects are stressed for this introduction to the articles, as are some of the supporting theoretical results. Most of the details are left to the articles. For this introductory text, we do not give algorithms and results in full generality and for all cases. Rather we focus on a few important cases. For example, we consider only a few specific Markov chains, rather than general Harris ergodic chains for the IS correction, and we focus on the use of unbiased estimators from particle filters[3] in state space models, rather than from general importance sampling distributions in latent variable models. Some more generality is provided in the original articles listed on page vi.

We begin in Section 2 with a specific problem of intractable likelihoods for statistical models, namely, that of the state space model, and review how interacting particle systems known as particle filters [44, 94] can lead to unbiased estimators of the ∞-likelihood (the likelihood corresponding to the family $\mathscr{P}_\infty$), as long as the dynamics of the state space model can be simulated. We also detail an MCMC known as the particle marginal Metropolis-Hastings (PMMH) [1] (see also [5, 9, 59, 69]), which allows for ∞-inference for the corresponding Bayesian model posterior, when unbiased ∞-likelihood estimators are available.

In Section 3, as in [A] we consider two different MCMCs, which are intended for acceleration of PMMH, and which are based on use of an approximate family $\mathscr{P}_0$ and unbiased estimators for the ∞-likelihood. These are the delayed acceptance (DA) MCMC [cf. 59, 8, 16, 17, 41, 61] and the MCMC-IS [cf. 24, 37, 38, 48, 73, A], both of which allow for unbiased estimators of the 0-likelihood and ∞-likelihood, which can be useful when deterministic approximations are not available[4]. Based on an extension of the covariance ordering [67] to the IS setting, with differing

[2]As for the intended audience, in order to keep the text of moderate size we must suppose some notions from analysis [cf. 80], probability [cf. 34], simulation [cf. 61], and statistics [cf. 33]. We try to strike a balance, to make the text of interest both to those knowledgeable and less knowledgeable in the subject matter of the thesis.

[3]also known as *sequential Monte Carlo*

[4]The references [41] for DA and [A] for MCMC-IS are most relevant in the unbiased estimator context for intractable likelihoods.

reversible stationary probabilities and with unbiased estimators, we seek to compare the algorithms in terms of statistical efficiency, as in [B].

Section 4 is concerned with a discretely and partially observed Itô diffusion, where unbiased ∞-likelihood estimates can not be directly obtained by the particle filter, because the dynamics of the diffusion can not be simulated. Instead, approximate families $\mathscr{P}_\ell$ based on Euler approximation [cf. 56] are used, along with multilevel [49, 36], randomisation [64, 77] and particle filter coupling [53] techniques, leading to an unbiased estimator for the ∞-likelihood and to the Bayesian ∞-posterior by using an MCMC-IS with randomised weights, as in [C].

Section 5 is concerned with Bayesian models with intractable likelihoods, where an unbiased estimator of the likelihood is not readily available, but where it is at least possible to generate artificial observations from $p^{(\theta)}(\mathrm{d}y')$. The approach of approximate Bayesian computation [cf. 92] is to use families $\mathscr{P}_{1/\epsilon}$ of approximations to $\mathscr{P}_\infty$, where $\mathscr{P}_{1/\epsilon}$ is indexed by $\epsilon > 0$, the 'tolerance,' which is difficult to choose. We detail an approach based on an adaptive MCMC, as well as MCMC-IS [98], with approximate confidence intervals for post-correction to finer tolerance, as in [D].

We close with a brief discussion of ideas for future work in Section 6 and provide expanded individual summaries for the thesis papers in Section 7.

2. Bayesian inference for state space models on path space

We introduce a well-known class of models based on latent variables on a state space $(\mathsf{X}, \mathcal{X})$ and conditionally independent observations on $(\mathsf{Y}, \mathcal{Y})$ which is sufficiently general and motivates a main application area of Articles [A], [B] and [C] based on unbiased estimators and approximate families $\mathscr{P}_\ell$.

2.1. Discretely-observed continuous-time path-dependent process.

To motivate this general class of models, we consider an example of continuous-time latent process. Suppose there is a process $(X'_t)_{t\geq 0}$ of *latent* or *hidden states* $X'_t \in \mathsf{X}$, where X'_t depends on $(X'_s)_{0\leq s<t}$ and the model parameter θ. Also, suppose $(Y'_t)_{t\geq 0}$ is another process (of observations), where Y'_t depends on $(X'_s)_{0\leq s\leq t}$ and θ. We make the realistic assumption[5] that only finitely many observations $\{Y'_{t_p}\}_{p=0}^n$ are gathered at observation times $\{t_p\}_{p=0}^n$.

Let us set $Y_p := Y'_{t_p}$ and $X_p := X'_{t_p}$. Let us define $X_{0:p} := (X_0, \ldots, X_p)$, and for fixed parameter value θ, consider the following dependency structure involving conditionally independent observations:

$$\cdots \longrightarrow X_{0:p-1} \longrightarrow X_{0:p} \longrightarrow X_{0:p+1} \longrightarrow \cdots$$
$$\qquad\qquad\quad \downarrow \qquad\qquad \downarrow \qquad\qquad \downarrow$$
$$\qquad\qquad\quad Y_{p-1} \qquad\quad Y_p \qquad\quad Y_{p+1}$$

Here, the arrows denote a dependency relationship, described in the following, where the initial state $X_0 \sim \eta_0^{(\theta)}$ is drawn from an initial distribution $\eta_0^{(\theta)}$. The dynamics between states (on path space) $X_{0:p-1}$ and $X_{0:p}$ is defined by a Markov probability kernel $\bar{M}_p^{(\theta,\infty)}$ (on path space), where

$$\bar{M}_p^{(\theta,\infty)}(x_{0:p-1}, \mathrm{d}x'_{0:p}) := \mathbf{1}\{x'_{0:p-1} = x_{0:p-1}\} M_p^{(\theta,\infty)}(x_{0:p-1}, \mathrm{d}x'_p),$$

[5] since the continuum can not easily be recorded by electronic or other physical means

where $M_p^{(\theta,\infty)}$ is a Markov probability kernel from X^p to X induced by the dynamics of the path-dependent continuous-time process. The observations Y_p are obtained via $Y_p \sim g_p^{(\theta)}(\,\cdot\,|X_{0:p})$, where $g_p^{(\theta)}$ is the *observational density*.

Let us set as shorthand $M_0^{(\theta,\infty)}(x_{0:-1}, \mathrm{d}x_0) := \eta_0(\mathrm{d}x_0)$ and

$$G_p^{(\theta)}(x_{0:p}) := g_p^{(\theta)}(y_p|x_{0:p}).$$

The model described above in terms of the pair $(M_p^{(\theta,\infty)}, G_p^{(\theta)})_{p=0}^n$ is known as a path-dependent *state space model* (SSM)[6], or, more succinctly, as a *Feynman-Kac model* [cf. 18].

Simulation methods for the ∞-Feynman-Kac model are impossible if the dynamics $M_p^{(\theta,\ell)}$ can not be simulated exactly. Besides some (important) exceptions, this is in general the case for continuous-time latent processes [cf. 19, Sect. 1.3]. However, we will see when we consider Itô diffusions in Section 4, that often one can obtain Euler type approximations of the original process, with precision denoted 'ℓ', leading to approximate dynamics $M_p^{(\theta,\ell)}$ between observation times [cf. 19, 56]. Using the same observational densities as for the exact model, we obtain a Feynman-Kac model $(M_p^{(\theta,\ell)}, G_p^{(\theta)})_{p=0}^n$ derived from the Euler type approximation of the dynamics.

2.2. **Model probabilities.** We now describe some of the probabilities associated to a Feynman-Kac model $(M_p^{(\theta,\ell)}, G_p^{(\theta)})_{p=0}^n$.

First, we define a bit of standard notation from analysis. If μ is a probability measure and $s \geq 1$, we denote by $L^s(\mu)$ the Banach space of real-valued functions ϕ, modulo equivalence in norm, with finite norm

$$\mu(|\phi|^s)^{\frac{1}{s}} < \infty, \qquad \text{where} \qquad \mu(\phi) := \int \phi(x)\mu(\mathrm{d}x). \tag{2}$$

Consider now the conditional ℓ-model probability on the latents, or *ℓ-smoother*,

$$p^{(\theta,\ell)}(\mathrm{d}x_{0:n}) = \frac{p_u^{(\theta,\ell)}(\mathrm{d}x_{0:n})}{p_u^{(\theta,\ell)}(1)}, \tag{3}$$

where[7]

$$p_u^{(\theta,\ell)}(\mathrm{d}x_{0:n}) = \left(\prod_{p=0}^n G_p^{(\theta)}(x_{0:p})\right)\eta_0^{(\theta)}(\mathrm{d}x_0)\prod_{p=1}^n M_p^{(\theta,\ell)}(x_{0:p-1}, \mathrm{d}x_p). \tag{4}$$

Then $p^{(\theta,\ell)}$ represents the probability to observe the latent states given the observations according to the Feynman-Kac model $(M_p^{(\theta,\ell)}, G_p^{(\theta)})_{p=0}^n$. In terms of a statistical model[8] on X^{n+1}, we have $\mathscr{P}_\ell = \{p^{(\theta,\ell)}\}_{\theta\in\mathsf{T}}$ with $p^{(\theta,\ell)}(\mathrm{d}x_{0:n})$ defined in (3), for the Feynman-Kac model.

The *joint ℓ-posterior probability* for the Bayesian model over model parameters and latent states is then

$$\pi^{(\ell)}(\mathrm{d}\theta, \mathrm{d}x_{0:n}) \propto \mathrm{pr}(\mathrm{d}\theta)p_u^{(\theta,\ell)}(\mathrm{d}x_{0:n}). \tag{5}$$

[6]As SSM is also known as a *hidden Markov model* [cf. 14], especially in the engineering disciplines.

[7]In the notation $p_u^{(\theta,\ell)}(1)$, we view 1 as the function $x_{0:n} \mapsto 1$, and the integral $p_u^{(\theta,\ell)}(\phi) = \int \phi(x_{0:n})p_u^{(\theta,\ell)}(\mathrm{d}x_{0:n})$ as in (2) for $\phi : \mathsf{X}^{n+1} \to \mathbb{R}$.

[8]really on $(\mathsf{X}\times\mathsf{Y}, \mathcal{X}\otimes\mathcal{Y})$, but we view $y_{0:n} \in \mathsf{Y}^{n+1}$ as fixed and therefore disregarded in the notation

Writing the *marginal* ℓ-likelihood as $L^{(\ell)}(\theta) := p_u^{(\theta,\ell)}(1)$ and considering the *marginal* ℓ-posterior on θ, we obtain a more familiar formula to (1) given in the beginning in Section 1.2, namely,

$$\pi_m^{(\ell)}(\mathrm{d}\theta) = \int_{\mathsf{X}^{n+1}} \pi^{(\ell)}(\mathrm{d}\theta, \mathrm{d}x_{0:n}) = \frac{\mathrm{pr}(\mathrm{d}\theta) L^{(\ell)}(\theta)}{\int \mathrm{pr}(\mathrm{d}\theta) L^{(\ell)}(\theta)}.$$

The main topic of this thesis is incorporation of ℓ-approximation within a ∞-inference method, to obtain efficient and user-friendly inference with respect to $\pi^{(\infty)}$ and $\pi_m^{(\infty)}$.

2.3. **Particle filter.** Ignoring the θ and ℓ labels, we have seen that a Feynman-Kac model (with time horizon n) is defined through a pair $(M_p, G_p)_{p=0}^n$, where,

(i) $M_p(x_{0:p-1}, \mathrm{d}x_p)$ is a Markov 'transition' kernel for $p = 1, \ldots, n$, and $M_0(x_{0:-1}, \mathrm{d}x_0) := \eta_0(\mathrm{d}x_0)$ is a probability measure, and
(ii) $G_p(x_{0:p})$ is a nonnegative 'potential' function for $0 \leq p \leq n$.

The particle filter (PF) (Algorithm 1) was popularised in [e.g. 94, 44], and allows for unbiased estimation [cf. 18, 28] of

$$p_u(\mathrm{d}x_{0:n}) = \left(\prod_{p=0}^n G_p(x_{0:p}) \right) \eta_0(\mathrm{d}x_0) \prod_{p=1}^n M_p(x_{0:p-1}, \mathrm{d}x_p), \tag{6}$$

for (traditional) SSMs that are not path-dependent, that is,

$$M_p(X_{0:p-1}, \mathrm{d}x'_p) = M_p(X_{p-1}, \mathrm{d}x'_p), \tag{7}$$
$$G_p(X_{0:p}) = G_p(X_p). \tag{8}$$

However, straightforward generalisation also allows for unbiased estimators in the path-dependent setting of Feynman-Kac models, at least when the dynamics can be simulated [cf. 18]. In addition, as is well-known, the general resampling scheme in PF (Algorithm 1) for ancestor random variables $\{A_p^{(i)}\}_{i=1}^N$ do lead to unbiased estimators, since the equality

$$\mathbb{E}\Big[\sum_{k=1}^N \mathbf{1}\{A_p^{(k)} = i\} \Big] = N \frac{V_p^{(i)}}{V_p^*},$$

is assumed satisfied for all $p \in \{0{:}n\}$ in PF (Algorithm 1) [cf. A, Prop. 20]. Such resampling schemes include the popular multinomial, stratified, residual, and systematic resampling [cf. 14, 25, 28].

The unbiased estimator, from the output $(V_n^{(i)}, \mathbf{X}^{(i)})_{i=1}^n$ of PF (Algorithm 1) run for Feynman-Kac model $(M_p, G_p)_{p=0}^n$, is obtained by setting

$$\hat{p}_u(\phi) := \sum_{i=1}^N V_n^{(i)} \phi(\mathbf{X}^{(i)}), \tag{9}$$

for $\phi \in L^1(p_u)$, which satisfies

$$\mathbb{E}\big[\hat{p}_u(\phi)\big] = p_u(\phi). \tag{10}$$

An important point is that particle approximations $\hat{p}_u(\mathrm{d}x_{0:n})$, for $p_u(\mathrm{d}x_{0:n})$ through the PF for the model $(M_p, G_p)_{p=0}^n$, are not unique [cf. 18, Sect. 2.4.2]. One standard

Algorithm 1 Particle filter for Feynman-Kac model $(M_p, G_p)_{p=0}^n$, with $N \geq 1$ particles.

In the following, the particle index i implicitly assumes all values in $\{1{:}N\}$.

(1) For initialisation,
  (i) Sample $X_0^{(i)} \sim \eta_0$. Set $\mathbf{X}^{(i)} := X^{(i)}$.
  Set $V_0^{(i)} := \frac{1}{N} G_0(\mathbf{X}^{(i)})$ and set $V_0^* := \sum_{j=1}^N V_0^{(j)}$.
  (ii) Sample random variables $\{A_0^{(k)}\}_{k=1}^N$ satisfying
  $\mathbb{E}\big[\sum_{k=1}^N \mathbf{1}\{A_0^{(k)} = i\}\big] = N V_0^{(i)} / V_0^*$.

(2) For $p = 1, \dots n$,
  (i) Sample $X_p^{(i)} \sim M_p(\mathbf{X}^{(A_{p-1}^{(i)})}, \cdot\,)$. Set $\mathbf{X}^{(i)} := (\mathbf{X}^{(A_{p-1}^{(i)})}, X^{(i)})$.
  Set $V_p^{(i)} := \big(V_{p-1}^*\big)\big(\frac{1}{N} G_p(\mathbf{X}^{(i)})\big)$ and set $V_p^* := \sum_{j=1}^N V_p^{(j)}$.
  (ii) Sample random variables $\{A_p^{(k)}\}_{k=1}^N$ satisfying
  $\mathbb{E}\big[\sum_{k=1}^N \mathbf{1}\{A_p^{(k)} = i\}\big] = N V_p^{(i)} / V_p^*$.

Output: $(V^{(i)}, \mathbf{X}^{(i)})_{i=1}^N$, where $V^{(i)} := V_n^{(i)}$.

way to obtain a different particle approximation is merely changing the Feynman-Kac model to $(\tilde{M}_p, \tilde{G}_p)_{p=0}^n$, but in such a way that

$$\Big(\prod_{p=0}^n \tilde{G}_p(x_{0:p})\Big)\tilde{\eta}_0(\mathrm{d}x_0)\prod_{p=1}^n \tilde{M}_p(x_{0:p-1}, \mathrm{d}x_p) = \Big(\prod_{p=0}^n G_p(x_{0:p})\Big)\eta_0(\mathrm{d}x_0)\prod_{p=1}^n M_p(x_{0:p-1}, \mathrm{d}x_p). \tag{11}$$

holds, and running the PF (Algorithm 1) for the new Feynman-Kac model. From (6) and (10), it follows that the unbiased estimator from the PF run for $(\tilde{M}_p, \tilde{G}_p)_{p=0}^n$ delivers the same unbiased estimation for $p_u(\mathrm{d}x_{0:n})$ corresponding to the model $(M_p, G_p)_{p=0}^n$. As an example for $(\tilde{M}_p, \tilde{G}_p)_{p=0}^n$, consider

$$\tilde{G}_p(x_{0:p}) := \frac{G_p(x_{0:p})M_p(x_{0:p-1}, \mathrm{d}x_p)}{\tilde{M}_p(x_{0:p-1}, \mathrm{d}x_p)}(x_{0:p})$$

in the sense of a Radon-Nikodým derivative [cf. 90], which always exists if M_p and $\tilde{M}_p$ admit densities and a support condition holds.

This non-uniqueness opens the door to consider more efficient PF implementations for a particular model and filtering/smoothing problem [cf. 18, 28, 45, 75]. The question of the optimal choice of $(M_p^\star, G_p^\star)_{p=0}^n$ for the smoothing problem (i.e. unbiased estimation of $p_u(\mathrm{d}x_{0:n})$) has been considered in [45]. As the optimal choice is usually not implementable, [45] suggest an adaptive iterative algorithm, based on approximating families of mixtures of normals, in order to approximately find $M_p^\star$ and $G_p^\star$ (see also e.g. [50] for a related method). Deterministic approximations, such as Laplace approximations [cf. 83], can also be used as a substitute for the optimal transition $M_p^\star$ [A] (see also [60]). We emphasise that all the above mentioned approaches to the optimal choice problem achieve unbiased estimation of $p_u(\mathrm{d}x_{0:n})$, as they use appropriately weighted potentials so that (11) holds.

Latent inference with respect to $p(\mathrm{d}x_{0:n})$ is possible through the PF, at least when the dynamics M_p can be simulated, by using a ratio estimator targeting $p_u(\phi)/p_u(1)$.

Algorithm 2 Particle marginal Metropolis-Hastings, with $m \geq 1$ iterations.

With $(\Theta_0, V_0^{(i)}, \mathbf{X}_0^{(i)})_{i=1}^N$ given, with $\sum_{i=1}^N V_0^{(i)} > 0$, for $k = 1, \ldots, m$, do:

(i) Sample $\Theta' \sim q(\cdot|\Theta_{k-1})$ from a transition kernel q on T.

(ii) Run PF (Algorithm 1) for $(M^{(\Theta',\infty)}, G^{(\Theta')})$, outputting $(V'^{(i)}, \mathbf{X}'^{(i)})_{i=1}^N$.

(iii) Accept, setting $(\Theta_k, V_k^{(i)}, \mathbf{X}_k^{(i)})_{i=1}^N \leftarrow (\Theta', V'^{(i)}, \mathbf{X}'^{(i)})_{i=1}^N$, with probability

$$\min\left\{1, \frac{\mathrm{pr}(\Theta')\left(\sum_{i=1}^N V'^{(i)}\right) q(\Theta_{k-1}|\Theta')}{\mathrm{pr}(\Theta_{k-1})\left(\sum_{i=1}^N V_{k-1}^{(i)}\right) q(\Theta'|\Theta_{k-1})}\right\}. \tag{13}$$

Otherwise, reject, setting $(\Theta_k, V_k^{(i)}, \mathbf{X}_k^{(i)})_{i=1}^N \leftarrow (\Theta_{k-1}, V_{k-1}^{(i)}, \mathbf{X}_{k-1}^{(i)})_{i=1}^N$.

That is, if $\{\hat{p}_{u,k}\}_{k=1}^m$ with $m \geq 1$ are formed[9] as in (9) from independent runs of PF (Algorithm 1) for $(M_p, G_p)_{p=0}^n$, then

$$\frac{\sum_{k=1}^m \hat{p}_{u,k}(\phi)}{\sum_{k=1}^m \hat{p}_{u,k}(1)} \xrightarrow{m\to\infty} p^{(\theta)}(\phi), \qquad \text{almost surely.} \tag{12}$$

We remark that the above estimator (12), as mentioned for example in [15, Eq. 1], is an IS analogue of the 'particle independent Metropolis-Hastings' (PIMH) [1] chain for latent smoothing. The algorithm based on (12) is completely parallelisable and does not depend on mixing of a chain, and is therefore relatively resilient in the number of particles N. Straightforward consistent estimators to construct confidence intervals are also available [cf. A, Prop. 23].

2.4. **Particle marginal Metropolis-Hastings.** The main task for which we are interested is joint ∞-inference with respect to $\pi^{(\infty)}(\mathrm{d}\theta, \mathrm{d}x_{0:n})$. So far, we only have shown how to perform ∞-inference for $p_u^{(\theta,\infty)}(\mathrm{d}x_{0:n})$ and $p^{(\theta,\infty)}(\mathrm{d}x_{0:n})$, with θ fixed. Surprisingly [cf. 5, 9, 59, 69], joint inference is possible, using an MCMC known as the *particle marginal Metropolis-Hastings* (PMMH) [1]. Assuming $q(\theta'|\theta) > 0$ for all θ, $\theta' \in \mathsf{T}$ in the PMMH chain (Algorithm 2), or a similarly mild condition ensuring Harris ergodicity of the chain [cf. 66], the estimator formed from PMMH is strongly consistent: for $f \in L^1(\pi^{(\infty)})$,

$$E_m^{PM}(f) := \frac{1}{m}\sum_{k=1}^m \sum_{i=1}^N \frac{V_k^{(i)} f(\Theta_k, \mathbf{X}_k^{(i)})}{\sum_{j=1}^N V_k^{(j)}} \xrightarrow{m\to\infty} \pi^{(\infty)}(f), \qquad \text{a.s.}^{10} \tag{14}$$

We remark about 'Metropolis-Hastings type' MCMC. The PMMH [1] is used in state space models using PFs, *pseudo-marginal* MCMC [5, 9, 59, 69] is the general term for the chain used in latent variable models with unbiased estimators, and Metropolis-Hastings MCMC [65, 48] is used in Bayesian models with tractable likelihoods. In fact, it is possible to view these 'Metropolis-Hastings type' MCMCs each as a substantiation of the other: one direction follows by viewing the pseudo-marginal MCMC and PMMH as full-dimensional Metropolis-Hastings kernels on an extended state space, while the other direction follows by trivialisation [cf. 5].

[9] Traditionally in particle filtering [cf. 28], latent inference (12) is done with $m = 1$, possibly with a final resampling to form uniformly weighted particles, but final resampling leads to higher variance of the resulting estimator and is unnecessary here.

[10] almost surely

Algorithm 3 Delayed acceptance, with $m \geq 1$ iterations, and $\epsilon \geq 0$

Given $(\Theta_0, V_0^{(i)}, \mathbf{X}_0^{(i)}, \hat{L}^{(0)}(\Theta_0))_{i=1}^N$, with $\sum_{i=1}^N V_0^{(i)} > 0$ and $\hat{L}^{(0)}(\Theta_0) > 0$.
For $k = 1, \ldots, m$, do:

(i) Sample $\Theta' \sim q(\cdot|\Theta_{k-1})$ from a transition kernel q on T.
Obtain unbiased estimate $\hat{L}^{(0)}(\Theta')$ of $L^{(0)}(\Theta')$.
Proceed to step (ii) with probability

$$\min\left\{1, \frac{\mathrm{pr}(\Theta')\big(\hat{L}^{(0)}(\Theta') + \epsilon\big)q(\Theta_{k-1}|\Theta')}{\mathrm{pr}(\Theta_{k-1})\big(\hat{L}^{(0)}(\Theta_{k-1}) + \epsilon\big)q(\Theta'|\Theta_{k-1})}\right\}. \tag{15}$$

Otherwise, reject, setting
$(\Theta_k, V_k^{(i)}, \mathbf{X}_k^{(i)}, \hat{L}^{(0)}(\Theta_k))_{i=1}^N \leftarrow (\Theta_{k-1}, V_{k-1}^{(i)}, \mathbf{X}_{k-1}^{(i)}, \hat{L}^{(0)}(\Theta_{k-1}))_{i=1}^N$.

(ii) Run PF (Algorithm 1) for $(M^{(\Theta',\infty)}, G^{(\Theta')})$, outputting $(V'^{(i)}, \mathbf{X}'^{(i)})_{i=1}^N$. Accept, setting $(\Theta_k, V_k^{(i)}, \mathbf{X}_k^{(i)}, \hat{L}^{(0)}(\Theta_k))_{i=1}^N \leftarrow (\Theta', V'^{(i)}, \mathbf{X}'^{(i)}, \hat{L}^{(0)}(\Theta'))_{i=1}^N$, with probability

$$\min\left\{1, \frac{\big(\sum_{i=1}^N V'^{(i)}\big)/\big(\hat{L}^{(0)}(\Theta') + \epsilon\big)}{\big(\sum_{i=1}^N V_{k-1}^{(i)}\big)/\big(\hat{L}^{(0)}(\Theta_{k-1}) + \epsilon\big)}\right\}. \tag{16}$$

Otherwise, reject, setting
$(\Theta_k, V_k^{(i)}, \mathbf{X}_k^{(i)}, \hat{L}^{(0)}(\Theta_k))_{i=1}^N \leftarrow (\Theta_{k-1}, V_{k-1}^{(i)}, \mathbf{X}_{k-1}^{(i)}, \hat{L}^{(0)}(\Theta_{k-1}))_{i=1}^N$.

3. Accelerations based on an approximation

The Metropolis-Hastings MCMC has served as the backbone of the MCMC revolution for half of the last century [23], while pseudo-marginal MCMC and the PMMH have been quite popular and extensively used in the current century (see [B, Sect. 1.2] for a review). Because of the importance of these MCMCs, there has been considerable interest in their possible acceleration. We focus on acceleration of the PMMH in the following.

Usually by far the most computationally intensive part of the PMMH is running the PF (Algorithm 1), for $(M_p^{(\theta,\infty)}, G_p^{(\theta)})_{p=0}^n$ with output $(V^{(i)}, \mathbf{X}^{(i)})_{i=1}^N$, to obtain the unbiased estimator

$$\hat{L}^{(\infty)}(\theta) := \hat{p}_u^{(\theta,\infty)}(1) = \sum_{i=1}^N V^{(i)}$$

of the likelihood $L^{(\infty)}(\theta)$. The idea of acceleration based on approximation is to substitute a computationally cheaper (non-negative unbiased estimator $\hat{L}^{(0)}(\theta)$ of an) approximation $L^{(0)}(\theta)$ for the ∞-likelihood, instead of using $\hat{L}^{(\infty)}(\theta)$. One would also like to maintain (strong) consistency of the resulting estimator for the ∞-posterior.

3.1. Delayed acceptance and importance sampling.

One such popular acceleration algorithm is delayed acceptance (DA) (Algorithm 3) [cf. 59, 8, 16, 17, 41, 61], with $\epsilon \geq 0$. We require that almost surely the support condition

$$\hat{L}^{(\infty)}(\theta) > 0 \implies \big(\hat{L}^{(0)}(\theta) + \epsilon\big) > 0 \tag{17}$$

holds, so that the resulting weight $\hat{L}^{(\infty)}(\theta)/\big(\hat{L}^{(0)}(\theta)+\epsilon\big)$ in Algorithm 3(ii) is guaranteed well-defined. This can be simply achieved always by choosing a regularisation

Algorithm 4 MCMC-IS. Importance sampling correction of PMMH, with $m \geq 1$ iterations, and $\epsilon \geq 0$.

(P1) Given $(\Theta_0, \hat{L}^{(0)}(\Theta_0))$, with $\hat{L}^{(0)}(\Theta_0) > 0$, for $k = 1, \ldots, m$, do:
  (i) Sample $\Theta' \sim q(\cdot|\Theta_{k-1})$ from a transition kernel q.
  (ii) Obtain unbiased estimate $\hat{L}^{(0)}(\Theta')$ of $L^{(0)}(\Theta')$.
  (iii) Accept, setting $(\Theta_k, \hat{L}^{(0)}(\Theta_k)) \leftarrow (\Theta', \hat{L}^{(0)}(\Theta'))$, with probability (15). Otherwise, reject, setting $(\Theta_k, \hat{L}^{(0)}(\Theta_k)) \leftarrow (\Theta_{k-1}, \hat{L}^{(0)}(\Theta_{k-1}))$.

(P2) For all $k \in \{1{:}m\}$,
  (i) Run PF (Algorithm 1) for $(M^{(\Theta_k,\infty)}, G^{(\Theta_k)})$, outputting $(V_k^{(i)}, \mathbf{X}_k^{(i)})_{i=1}^N$.
  (ii) Set $\xi_k(\phi) := \frac{\sum_{i=1}^N V_k^{(i)} \phi(\mathbf{X}_k^{(i)})}{\hat{L}^{(0)}(\Theta_k)+\epsilon}$, for $\phi : \mathsf{X}^{n+1} \to \mathbb{R}$. Form the estimator,

$$E_m^{IS}(f) := \frac{\sum_{k=1}^m \xi_k(f^{(\Theta_k)})}{\sum_{k=1}^m \xi_k(1)}. \tag{18}$$

constant[11] $\epsilon > 0$, leading to asymptotically exact ∞-inference. We note that step (i) in DA (Algorithm 3) effectively acts as a screening stage: only 'good' proposals proceed to step (ii), where the expensive ∞-model PF must be run. The resulting DA estimator for the ∞-posterior is the same as that of PMMH, given in (14).

As an alternative to PMMH/DA, we consider MCMC-IS (Algorithm 4) [cf. 24, 37, 38, 48, 73, A]. Here, for $f : \mathsf{T} \times \mathsf{X}^{n+1} \to \mathbb{R}$ we have set $f^{(\theta)}(x_{0:n}) = f(\theta, x_{0:n})$. Assuming the Phase 1 chain is Harris ergodic (e.g. $q(\theta, \theta') > 0$ for all $\theta, \theta' \in \mathsf{T}$) and the support condition (17) holds, like the PMMH/DA estimator, the MCMC-IS estimator is strongly consistent [A, Thm. 3]: for $f \in L^1(\pi^{(\infty)})$,

$$E_m^{IS}(f) \xrightarrow{m\to\infty} \pi^{(\infty)}(f), \qquad \text{almost surely.}$$

Phase 1 of MCMC-IS (Algorithm 1) implements a PMMH (Algorithm 2) targeting marginally

$$\pi_m^{(0)}(\theta) \propto \mathrm{pr}(\theta) L^{(0)}(\theta).$$

Phase 2 consists of independent calls of PF (Algorithm 1), and is therefore completely parallelisable, unlike DA (Algorithm 3). This allows for the possibility of substantial additional speedup on a parallel computer [cf. 58].

We remark about an acceleration technique known as 'early rejection' [93] for Metropolis-Hastings, that can sometimes be employed if the likelihood takes a special form, described below.[12] The acceleration technique also applies to DA step (i) and MCMC-IS Phase 1, if $\hat{L}^{(0)}(\theta) = L^{(0)}(\theta)$ almost surely and $\epsilon = 0$. The form required in [93] is that the 0-likelihood $L^{(0)}(\theta)$ can be written, for example, as

$$L^{(0)}(\theta) \propto \prod_{j=0}^{n} \exp\big(-l_j^{(\theta,0)}(y_j)\big)$$

with $l_j^{(\theta,0)}(y_j) \geq 0$. In this case, because the likelihood only gets smaller with more components of the product computed, the calculation of the components can be ended and the proposal rejected early in acceptance probability (15) for DA and

[11] This will be done in Algorithm 6 given later, and is linked to 'defensive importance sampling' [51].

[12] A similar idea of early cancellation as 'early rejection' has been used previously in the exact simulation literature, under the name of 'retrospective simulation' [10].

MCMC-IS, as soon as the partially computed acceptance probability in (15) becomes smaller than the uniformly generated random variable [cf. 93, Sect. 4]. The 'early rejection' trick requires a special form for the likelihood, however, and therefore is not always applicable.

3.2. **The question of relative efficiency.** The delayed acceptance and importance sampling correction are two acceleration alternatives to the standard PMMH, both of which use the same approximation and algorithmic ingredients. The question of choice of alternative methods has been remarked before [17] in the simpler setting of Metropolis-Hastings, without unbiased estimators. Article [A] introduces the IS correction in the general case of unbiased estimators in both Phase 1 and Phase 2, and seeks to compare MCMC-IS with DA in the general setting.

A numerical comparison of the methods is done in [A], where the MCMC-IS approach was found to work slightly better than DA in experiments in SSMs, even without parallelisation. As an example of a computationally intensive experiment, a stochastic volatility model was considered with observation consisting of real data from daily financial index returns spanning two decades. Laplace approximations were used to approximate the 0-likelihood, and were used as well in the IS correction, namely, for the approximation to the optimal choice[13] of Feynman-Kac model for the smoothing problem for $p_u^{(\theta,\infty)}(\mathrm{d}x_{0:n})$. With all methods making intelligent use of the Laplace approximations, MCMC-IS performed significantly better than PMMH or DA in the experiment.

In additional to the experiments, many additional potential enhancements were suggested in [A] which would improve the computational efficiency of MCMC-IS in practice, relative to DA acceleration of PMMH, even further. For example, the Phase 2 IS weights do not need to be calculated during the burn-in phase[14] and for thinned out samples of the chain[15], nor for repeated samples of the chain if the jump chain[16] is used. As well, as previously mentioned, Phase 2 admits a straightforward parallelisation for calculation of the more expensive IS weights, which significantly increases the scalability and efficiency of MCMC-IS.

3.3. **Peskun and covariance orderings of asymptotic variances.** An estimator $E_m(f)$ is said to satisfy a central limit theorem (CLT), if

$$\sqrt{m}\big[E_m(f) - \pi^{(\infty)}(f)\big] \xrightarrow{m\to\infty} \mathrm{N}\big(0, \sigma^2(f)\big), \qquad \text{in distribution.}$$

In this case, we call $\sigma^2(f)$ the *asymptotic variance* of the estimator.

Without taking into account computational factors previously mentioned (which generally support the use of MCMC-IS; see also Section 4.6), and considering just the statistical efficiency of the estimators in terms of the asymptotic variance, it was found in [B] through artificially constructed toy examples that either MCMC-IS or PMMH/DA may do arbitrary better than the other. Moreover, the examples seemed to indicate that MCMC-IS might do better in cases of practical interest, with multi-modal targets, a phenomenon remarked previously about MCMC-IS and Metropolis-Hastings [e.g. 37]. Proving that the IS acceleration is usually 'better'

[13] as discussed in Section 2.3

[14] Additionally, the debiasing tricks [cf. 39] may be effectively and efficiently used.

[15] Thinning [cf. 72] denotes the procedure, in which only every k^{th} sample of the Markov chain is kept, with say $k = 10$, in order to decrease the auto-correlation of samples.

[16] the chain formed formed from the original chain, consisting only of the accepted states of the original chain [cf. 27, A]

than DA is of course a separate matter, which can not be done based on experiments or examples alone.

We first introduce some notation and terminology. A Markov kernel K on $(\mathsf{X}, \mathcal{X})$ is said to be *reversible* with respect to a probability μ, if for all $A, B \in \mathcal{X}$,

$$\int \mu(\mathrm{d}x)K(x,\mathrm{d}y)\mathbf{1}\{x \in A,\, y \in B\} = \int \mu(\mathrm{d}y)K(y,\mathrm{d}x)\mathbf{1}\{x \in A,\, y \in B\}.$$

We also define the *Dirichlet form*

$$\mathcal{E}_K(g) := \langle g, (1-K)g\rangle_\mu$$

for $g \in L^2(\mu)$, where $\langle g_1, g_2\rangle_\mu := \int g_1(x)g_2(x)\mu(\mathrm{d}x)$, $Kg(x) := \int K(x,\mathrm{d}x')g(x')$ and $(1g)(x) = g(x)$.

The famous Peskun ordering [74, 95] says that if

$$K(x, A\backslash\{x\}) \geq L(x, A\backslash\{x\}) \qquad \mu\text{-almost every } x \in \mathsf{X},\ \forall A \in \mathcal{X}, \tag{19}$$

where K and L are two Markov kernels, both reversible with respect to a probability μ, then

$$\sigma_K^2(f) \leq \sigma_L^2(f) \qquad \forall f \in L^2(\mu), \tag{20}$$

where $\sigma_K^2(f)$ and $\sigma_L^2(f)$ denote the asymptotic variances of the K and L chains, respectively.

Consider next a popular Peskun 'type' comparison result for asymptotic variances of reversible chains, known as the covariance ordering[17] [67]: if K and L are two Markov kernels, both reversible with respect to a probability μ, and if

$$\mathcal{E}_K(g) \geq \mathcal{E}_L(g), \qquad \forall g \in L^2(\mu), \tag{21}$$

then

$$\sigma_K^2(f) \leq \sigma_L^2(f) \qquad \forall f \in L^2(\mu). \tag{22}$$

Compared to the Peskun ordering, the covariance ordering can be more useful in practice, as the criterion can distinguish better between chains on general state spaces. For example, some chains vanish along the diagonal, in which case (19) may be useless, but (21) may still be able to distinguish between these chains [cf. 67, 68].

As a simple application of the covariance ordering, let us consider the case of PMMH and DA, which are both reversible with respect to the same invariant measure (see [8] or Section 3.5). Using the identity

$$\mathcal{E}_L(g) = \frac{1}{2}\int \mu(\mathrm{d}x)L(x,\mathrm{d}y)\big(g(x)-g(y)\big)^2,$$

which holds for any μ-reversible kernel L, and that the product of the acceptance probabilities (15) and (16) in DA (Algorithm 3) is less than or equal to the acceptance probability (13) in PMMH (Algorithm 2), it can be shown [cf. 8] that the covariance ordering implies

$$\sigma_{PM}^2(f) \leq \sigma_{DA}^2(f).$$

[17] Though not mentioned by name, it was shown already in [95, Proof of Lem. 3] that the Peskun ordering is equivalent with the 'covariance' ordering.

3.4. **Peskun type ordering for importance sampling correction.** Article [B] is concerned with extending the covariance ordering to chains K and L reversible with respect to probabilities $\Pi^{(0)}$ and $\Pi^{(\infty)}$, where $\Pi^{(0)}$ and $\Pi^{(\infty)}$ may be different.

Suppose then that K and L are Harris ergodic chains on a space $(\mathsf{X}, \mathcal{X})$, where K is $\Pi^{(0)}$-reversible and L is $\Pi^{(\infty)}$-reversible. Suppose further that the Radon-Nikodým derivative[18]

$$w(x) := \frac{\mathrm{d}\Pi^{(\infty)}}{\mathrm{d}\Pi^{(0)}}(x)$$

exists. Let $\underline{c}$, $\bar{c} \geq 0$ be constants such that

$$\begin{aligned} \underline{c}\,\mathcal{E}_K(g) \leq & \mathcal{E}_L(g) \leq \bar{c}\,\mathcal{E}_K(g) \\ \underline{c} \leq & w(x) \leq \bar{c}, \end{aligned}$$

for all $x \in \mathsf{X}$ and $g \in L^2(\Pi^{(0)})$. Then [B, Thm. 2], for all $f \in L^2(\Pi^{(\infty)})$ with $\bar{f} := f - \Pi^{(\infty)}(f)$, we have

$$\sigma_K^2(f) + \mathrm{var}_{\Pi^{(0)}}(w\bar{f}) \leq \bar{c}\Big(\sigma_L^2(f) + \mathrm{var}_{\Pi^{(\infty)}}(f)\Big), \tag{23}$$

$$\sigma_K^2(f) + \mathrm{var}_{\Pi^{(0)}}(w\bar{f}) \geq \underline{c}\Big(\sigma_L^2(f) + \mathrm{var}_{\Pi^{(\infty)}}(f)\Big). \tag{24}$$

If $\Pi^{(0)} = \Pi^{(\infty)}$, then it is direct to see that (23) simplifies to the covariance ordering (22) given earlier. Versions of the orderings (23-24) also hold for when the marginal weight is bounded in a latent variable setting [B, Thm. 5], and for self-normalised estimators using jump chain representation and unbiased estimators [B, Thm. 12] to compare with pseudo-marginal type MCMC. We discuss a particular implication of these orderings in the next section, namely MCMC-IS (algorithm 4) compared to PMMH (Algorithm 2) and DA (Algorithm 3).

3.5. **Comparison results.** We are now ready to compare MCMC-IS (Algorithm 4) with PMMH (Algorithm 2) and DA (Algorithm 3) in terms of the asymptotic variance. For simplicity, we assume deterministic approximation for the 0-likelihood, that is, $\hat{L}^{(0)}(\theta) = L^{(0)}(\theta)$ almost surely.[19] We note that the MCMC-IS chain is $\Pi^{(0)}$-reversible, while the PMMH and DA chains are both $\Pi^{(\infty)}$-reversible, with probabilities defined in the following.

Article [B] shows how a comparison can be made when the (marginal) weight between the approximate and exact model posteriors w (or $\dot{w}$) is bounded (the weights w and $\dot{w}$ are defined below). This follows from the extension of the covariance ordering to the IS context with unbiased estimators, mentioned earlier.

We first need to define some notation. Let $Q_\theta^{(\infty)}(\mathrm{d}\mathbf{x}^{(1:N)}, \mathrm{d}v^{(1:N)})$ denote the law of the output $(\mathbf{X}^{(1:N)}, V^{(1:N)})$ of the PF (Algorithm 1) for the model $(M_p^{(\theta,\infty)}, G_p^{(\theta)})_{p=0}^n$. The full invariant probability of the PMMH (Algorithm 2) is then given by

$$\Pi^{(\infty)}(\mathrm{d}\theta, \mathrm{d}v^{(1:N)}, \mathrm{d}\mathbf{x}^{(1:N)}) = \frac{1}{c_\infty}\mathrm{pr}(\mathrm{d}\theta) Q_\theta^{(\infty)}(\mathrm{d}\mathbf{x}^{(1:N)}, \mathrm{d}v^{(1:N)}) \sum_{i=1}^N v^{(i)},$$

[18] This is the function w satisfying $\Pi^{(0)}(wg) = \Pi^{(\infty)}(g)$ for all $g \in L^1(\Pi^{(\infty)})$.

[19] For the general case for $\hat{L}^{(0)}(\theta)$, see [B, Thm. 14].

where c_∞ is a normalising constant. The full invariant probability of the IS corrected chain (Algorithm 4) is

$$\Pi^{(0)}(\mathrm{d}\theta, \mathrm{d}v^{(1:N)}, \mathrm{d}\mathbf{x}^{(1:N)}) = \frac{1}{c_0}\mathrm{pr}(\mathrm{d}\theta)\big(L^{(0)}(\theta)+\epsilon\big)Q_\theta^{(\infty)}(\mathrm{d}\mathbf{x}^{(1:N)}, \mathrm{d}v^{(1:N)}),$$

where c_0 is a normalising constant. We set for a function $f : \mathsf{T} \times \mathsf{X}^{n+1} \to \mathbb{R}$,

$$\hat{\zeta}(f) := \frac{\zeta(f)}{\zeta(1)}, \qquad \text{where} \qquad \zeta(f) := \sum_{i=1}^{N} V^{(i)} f(\Theta, \mathbf{X}^{(i)}).$$

Assuming

$$\hat{L}^{(\infty)}(\theta) > 0 \implies \big(L^{(0)}(\theta)+\epsilon\big) > 0 \tag{25}$$

almost surely, for some $\epsilon \geq 0$, the weights

$$w(\theta, v^{(1:N)}, \mathbf{x}^{(1:N)}) := \frac{c_0}{c_\infty}\frac{1}{L^{(0)}(\theta)+\epsilon}\sum_{i=1}^{N} v^{(i)}, \qquad \text{and} \qquad \dot{w}(\theta) := \frac{c_0}{c_\infty}\frac{L^{(\infty)}(\theta)}{L^{(0)}(\theta)+\epsilon},$$

correspond to the Radon-Nikodým derivatives between the approximate and exact model full and marginal posteriors.

Let us now describe a CLT for MCMC-IS. As before, for a function $f \in L^2(\pi^{(\infty)})$, we set $\bar{f} := f - \pi^{(\infty)}(f)$. By [A, Thm. 7(i)], for $f \in L^2(\pi^{(\infty)})$ the MCMC-IS estimator (18) satisfies a CLT, with a formula for the MCMC-IS asymptotic variance given by

$$\sigma^2_{IS}(f) = \sigma^2_{IS,1}(f) + \sigma^2_{IS,2}(f), \tag{26}$$

assuming $\sigma^2_{IS}(f) < \infty$, support condition (17) holds, and the marginal chain $(\Theta_k)_{k\geq 1}$ of MCMC-IS (Algorithm 4) is Harris ergodic[20] and aperiodic[21]. Here, $\sigma^2_{IS,1}(f)$ is the asymptotic variance of the marginal chain $(\Theta)_{k\geq 1}$, that is,

$$\frac{1}{\sqrt{m}}\sum_{k=1}^{m}\mathbb{E}\big[w(\theta, V^{(1:N)}, \mathbf{X}^{(1:N)})\hat{\zeta}(\bar{f})|\Theta_k = \theta\big] \xrightarrow{m\to\infty} \mathrm{N}\big(0, \sigma^2_{IS,1}(f)\big),$$

and

$$\sigma^2_{IS,2}(f) := \pi_m^{(0)}\big(v_{w\hat{\zeta}(\bar{f})}\big),$$

with

$$v_g(\theta) := \mathrm{var}\big(g(\theta, V^{(1:N)}, \mathbf{X}^{(1:N)})|\Theta_k = \theta\big).$$

Note the decomposition of the MCMC-IS asymptotic variance (26) into marginal MCMC and IS correction components, which may be helpful in questions of tuning and allocation of computational resources. A similar decomposition is not expected to hold for the DA asymptotic variance.

We now state the comparison results between MCMC-IS and PMMH/DA. For functions $f \in L^2(\pi^{(\infty)})$, such that the CLT and conditions given above for MCMC-IS hold, and assuming the PMMH and DA chains are Harris ergodic, we have the following comparison result [B, Thm. 14], with σ^2_L equal to σ^2_{PM} or σ^2_{DA}:

$$\sigma^2_{IS}(f) \leq (\sup \dot{w})\Big(\sigma^2_L(f) + \mathrm{var}_{\Pi^{(\infty)}}\big(\hat{\zeta}(f)\big)\Big) + 3\mathrm{var}_{\Pi^{(0)}}\big(w\hat{\zeta}(\bar{f})\big). \tag{27}$$

[20] E.g. $q(\theta, \theta') > 0$ for all $\theta, \theta' \in \mathsf{T}$.

[21] See for example [66] for this and other definitions.

Note that $\sup \dot{w} \leq \sup w$. We have, moreover, if also $\sup w < \infty$,

$$\sigma^2_{IS}(f) + \mathrm{var}_{\Pi^{(0)}}\left(w\hat{\zeta}(\bar{f})\right) \leq (\sup w)\left(\sigma^2_L(f) + \mathrm{var}_{\Pi^{(\infty)}}\left(\hat{\zeta}(f)\right)\right) \tag{28}$$

$$\sigma^2_{IS}(f) + \mathrm{var}_{\Pi^{(0)}}\left(w\hat{\zeta}(\bar{f})\right) \geq (\inf w)\left(\sigma^2_L(f) + \mathrm{var}_{\Pi^{(\infty)}}\left(\hat{\zeta}(f)\right)\right) \tag{29}$$

The results show that the asymptotic variance of MCMC-IS and PMMH/DA can be related up to additive and multiplicative constants, which can be informative by (27) in practical cases where the marginal weight $\dot{w}$ is bounded, where $\dot{w}$ relates the ratio of likelihoods. We note that (29) is usually not helpful since a positive lower bound on w is usually not possible, while (27) and (28) do not require a positive lower bound, and are therefore more generally applicable, providing theoretical guarantees for MCMC-IS in terms of PMMH/DA. Another nice facet of (27-29) is that the function $f \in L^2(\Pi^{(\infty)})$ is allowed to be a function on $\mathsf{T} \times \mathsf{X}^{n+1}$, not only on T.

Also shown in [B] is the not too surprising fact that geometric ergodicity of the MCMC-IS augmented chain is inherited by its marginal chain. This relates the fact that the convergence and mixing of the MCMC-IS chain is not affected by the noise in the Phase 2 unbiased estimators, unlike PMMH and DA, which are very dependent on the noise, and are not geometrically ergodic if the unbiased estimator is unbounded [cf. 5, B]. Of course, the asymptotic variance (26) of the MCMC-IS estimator (18) depends on the noise, but it seems it is not as harmful in the output estimator compared to in the acceptance ratios (13) and (16) of PMMH and DA, respectively. Besides convergence and mixing, geometric ergodicity is also likely helpful for example in estimation of the asymptotic variance [cf. 32], as well as in verifying convergence of adaptive MCMC schemes [cf. 6], at least based on the existing theory.

There is room for further theoretical development. For example, quantification of the error of MCMC-IS and of the asymptotic variance, could be investigated along the lines of [32, 79]. Also, in terms of non-asymptotic error bounds, results for MCMC [e.g. 100, 63, 81] could likely be extended to MCMC-IS.

4. Bayesian inference for state space models with diffusion dynamics

The PF (Algorithm 1) for the Feynman-Kac model $(M_p^{(\theta,\infty)}, G_p^{(\theta)})$ requires that the samples can be drawn from the Markov transition kernels $M_p^{(\theta,\infty)}$. However, as discussed at the end of Section 2.1, in many settings important for real applications, the assumption that the dynamics can be simulated does not hold.

We consider the case where the model $(M_p^{(\theta,\infty)}, G^{(\theta)})$ stems from a discretely and partially observed Itô diffusion process. Suppose $(X'_t)_{t\geq 0}$ solves an Itô stochastic differential equation of the form

$$\mathrm{d}X'_t = a^{(\theta)}(X'_t)\mathrm{d}t + b^{(\theta)}(X'_t)\mathrm{d}W_t, \qquad t \geq 0,$$

where $\{W_t\}_{t\geq 0}$ is a standard Brownian motion. As in Section 2.1, we assume that there is some observational process $(Y'_t)_{t\geq 0}$, and that observations $\{Y'_{t_p}\}_{p=0}^n$ are obtained at discrete times $\{t_p\}_{p=0}^n$. With $X_p := X'_{t_p}$ and $Y_p := Y'_{t_p}$, and with $G_p^{(\theta)}(x_p) := g_p^{(\theta)}(y_p|x_p)$, we obtain a model $(M_p^{(\theta,\infty)}, G_p^{(\theta)})_{p=0}^n$ which additionally satisfies the SSM conditions (7-8).

In some, essentially one-dimensional diffusion settings, where the Lamperti transformation [cf. 70] can be applied, ∞-inference is possible for $p^{(\theta,\infty)}$ [10, 11, 31] and $\pi^{(\infty)}$ [87, 97]. Article [C] attempts to extend to more settings ∞-inference for $p^{(\theta,\infty)}$

and $\pi^{(\infty)}$ in a computationally feasible way. The approach of [C] is based on Euler approximations of the dynamics [cf. 56], multilevel Monte Carlo (MLMC) [49, 36], a particle filter coupling [53], *debiasing* tricks for MLMC [64, 77], and an IS type correction [A]. We introduce each of these in turn in the following.

4.1. **Euler approximations.** The Euler approximation amounts to defining a discretisation size $h_\ell \propto 2^{-\ell}$ for $\ell \in \mathbb{N} \cup \{0\}$, and replacing the dynamics of the latent process $(X'_t)_{t\geq 0}$ with a discrete-time Markov chain,

$$X'_{t+h_\ell} = X'_t + a^{(\theta)}(X'_t)h_\ell + b^{(\theta)}(X'_t)(W_{t+h_\ell} - W_t).$$

Here, $(W_t)_{t\geq 0}$ is a standard Brownian motion, so that $W_{t+h_\ell} - W_t \sim \mathrm{N}(0, h_\ell)$ is independent of X'_u, $u \leq t$.

The approximate dynamics corresponds to an approximate transition $M_p^{(\theta,\ell)}$, which, together with the conditionally independent observations, results in a model $(M_p^{(\theta,\ell)}, G_p^{(\theta)})$ satisfying the SSM conditions (7-8), with ℓ-smoother $p^{(\theta,\ell)}(\mathrm{d}x_{0:n})$ given in (3) in Section 2.2, and with joint ℓ-posterior $\pi^{(\ell)}(\mathrm{d}\theta, \mathrm{d}x_{0:n})$ given in (5).

Joint ℓ-inference for $\pi^{(\ell)}$ is possible using PMMH (Algorithm 2) [1], which has been quite popular in the setting of diffusions [cf. 42]. Another ℓ-inference method [53], which uses PMMH together with a 'multilevel' decomposition, is discussed in the following section. We reiterate that, in distinction to these methods, the goal in [C] is to develop a ∞-inference method (which is also computationally efficient).

4.2. **Multilevel Monte Carlo.** The idea of MLMC is based on telescoping sums, where each summand is coupled in such a way that leads to lower variance of the resulting estimator [49, 36]. The multilevel decomposition used in [53], for ℓ_F-inference in partially observed diffusions, is based on the telescoping sum in terms of expectations of normalised probabilities,

$$\pi^{(\ell_F)}(\phi) = \sum_{\ell=1}^{\ell_F} \left(\pi^{(\ell)}(\phi) - \pi^{(\ell-1)}(\phi) \right) + \pi^{(0)}(\phi),$$

with $\ell_F \geq 1$ ideally taken quite large. PMMH chains are run at level ℓ and at level $\ell - 1$ in each summand, and are coupled to each other using the 'approximate coupling' described below.

In [C], such a telescoping sum is used not to target an expectation (and where the normalising constants must be simultaneously estimated in each summand), but rather an integral taken with respect to an unnormalised ℓ_F-smoother,

$$p_u^{(\theta,\ell_F)}(\phi) = \sum_{\ell=1}^{\ell_F} \left(p_u^{(\theta,\ell)}(\phi) - p_u^{(\theta,\ell-1)}(\phi) \right) + p_u^{(\theta,0)}(\phi), \tag{30}$$

with $\ell_F \geq 1$ ideally taken quite large. The quality of the approximation as measured by the variance depends on the coupling used for each increment

$$p_u^{(\theta,\ell)}(\phi) - p_u^{(\theta,\ell-1)}(\phi).$$

The algorithm used in [C] to unbiasedly estimate this difference is given in Algorithm 5, which we refer to as the 'delta PF' (ΔPF). The coupling used is the 'approximate coupling' of [53]. This coupling is based on a change of measure of the Feynman-Kac model on a joint path space, which, together with an importance sampling correction of the particle filter output, leads to the ΔPF.

Algorithm 5 Delta particle filter (ΔPF) for $(\check{M}_p^{(\theta,\ell)}, \check{G}_p^{(\theta)})_{p=0}^n$, with $\ell \geq 1$ and with $N \geq 1$ particles.

(i) Run PF (Algorithm 1) for $(\check{M}_p^{(\theta,\ell)}, \check{G}_p^{(\theta)})_{p=0}^n$, outputting $(V_n^{(i)}, \check{X}_{0:n}^{(i)})_{i=1}^N$.

(ii) Output $\Delta^{(\theta,\ell)}$, where, for $\phi : \mathsf{X}^{n+1} \to \mathbb{R}$,

$$\Delta^{(\theta,\ell)}(\phi) := \sum_{i=1}^N V_n^{(i)} \left(\overline{w}_\ell(\check{X}_{0:n}^{(i)}) \phi(X_{0:n}^{(\ell,i)}) - \underline{w}_\ell(\check{X}_{0:n}^{(i)}) \phi(X_{0:n}^{(\ell-1,i)}) \right)$$

where

$$\overline{w}_\ell(\check{X}_{0:n}) := \frac{\prod_{p=0}^n G_p^{(\theta)}(X_{0:p}^{(\ell)})}{\prod_{p=0}^n \check{G}_p^{(\theta)}(\check{X}_{0:p})} \qquad \text{and} \qquad \underline{w}_\ell(\check{X}_{0:n}) := \frac{\prod_{p=0}^n G_p^{(\theta)}(X_{0:p}^{(\ell-1)})}{\prod_{p=0}^n \check{G}_p^{(\theta)}(\check{X}_{0:p})}.$$

4.3. **Coupling of Feynman-Kac models.** Suppose $(M_p^{(\theta,\ell)}, G_p^{(\theta)})_{p=0}^n$ and $(M_p^{(\theta,\ell-1)}, G_p^{(\theta)})_{p=0}^n$ are two Feynman-Kac models. We describe a coupling of them as follows. For some fixed $\ell \geq 1$, $\check{M}_p^{(\theta,\ell)}(\check{x}_{0:p-1}, \mathrm{d}\check{x}_p)$ is assumed to be a coupling of the ℓ and $\ell-1$ level transitions, that is,

$$\check{M}_p^{(\theta,\ell)}(\check{x}_{0:p-1}, A \times \mathsf{X}) = M_p^{(\theta,\ell)}(x_{0:p-1}^{(\ell)}, A),$$
$$\check{M}_p^{(\theta,\ell-1)}(\check{x}_{0:p-1}, \mathsf{X} \times A) = M_p^{(\theta,\ell-1)}(x_{0:p-1}^{(\ell-1)}, A),$$

for $A \in \mathcal{B}(\mathsf{X})$ and with the notation $\check{x}_{0:p} = (x_{0:p}^{(\ell)}, x_{0:p}^{(\ell-1)})$ denoting an element in the space $\mathsf{X}^{2(p+1)}$, and we set

$$\check{G}_{0:p}^{(\theta)}(\check{x}_{0:p}) = \frac{1}{2}\left(G_p^{(\theta)}(x_{0:p}^{(\ell)}) + G_p^{(\theta)}(x_{0:p}^{(\ell-1)}) \right). \tag{31}$$

The dynamics $\check{M}_p^{(\theta,\ell)}$ is typically obtained in the diffusion context by using a common Brownian path for mesh discretisation levels ℓ and $\ell-1$. Other choices for $\check{G}_p^{(\theta)}$ are possible then the choice (31) used in [C]. The important point is that $\check{G}_p^{(\theta)}(\check{x}_{0:p}) > 0$ whenever $G_p^{(\theta)}(x_{0:p}^{(\ell)}) > 0$ or $G_p^{(\theta)}(x_{0:p}^{(\ell-1)}) > 0$. This ensures that the estimator $\Delta^{(\theta,\ell)}(\phi)$ from the ΔPF (Algorithm 5) is unbiased [C, Prop. 3]: for bounded $\phi : \mathsf{X}^{n+1} \to \mathbb{R}$,

$$\mathbb{E}[\Delta^{(\theta,\ell)}(\phi)] = p_u^{(\theta,\ell)}(\phi) - p_u^{(\theta,\ell-1)}(\phi).$$

We can then estimate $p_u^{(\theta,\ell_F)}$ unbiasedly using MLMC. Namely,

$$\mathbb{E}\big[I_{m_0}^{(\theta,0)}(\phi) + I_{m_{1:F}}^{(\theta,1:\ell_F)}(\phi) \big] = p_u^{(\theta,\ell_F)}(\phi), \tag{32}$$

where

$$I_{m_0}^{(\theta,0)}(\phi) := \frac{1}{m_0} \sum_{i=1}^{m_0} \hat{p}_{u,i}^{(\theta,0)}(\phi),$$

with $\{\hat{p}_{u,i}^{(\theta,0)}(\phi)\}_{i=1}^{m_0}$ independently run versions of the estimator $\hat{p}_u^{(\theta,0)}(\phi) = \sum_{i=1}^N V^{(i)} \phi(\mathbf{X}^{(i)})$ from the output $(V^{(1:N)}, \mathbf{X}^{(1:N)})$ of the PF (Algorithm 1) run for the model $(M_p^{(\theta,0)}, G_p^{(\theta)})_{p=0}^n$, and where

$$I_{m_{1:F}}^{(\theta,1:\ell_F)}(\phi) := \sum_{\ell=1}^{\ell_F} \frac{1}{m_\ell} \sum_{i=1}^{m_\ell} \Delta_i^{(\theta,\ell)}(\phi),$$

with $\{\Delta_i^{(\theta,\ell)}(\phi)\}_{i=1}^{m_\ell}$ estimators formed from independent runs of the ΔPF (Algorithm 5) run for the model $(\check{M}_p^{(\theta,\ell)}, \check{G}_p^{(\theta)})_{p=0}^n$.

The approach based on (32) allows for efficient MLMC estimation of the unnormalised ℓ_F-smoother $p_u^{(\theta,l_F)}$, over the latent states. If we were content with joint ℓ_F-inference, then we could apply already the IS type correction of MCMC as in Algorithm 4, with regularised 'likelihood' estimate

$$L(\Theta_k) := I_{m_0}^{(\Theta_k,0)}(1) + \epsilon$$

in the acceptance ratio (15), with $\epsilon \geq 0$, and with IS weights

$$\xi_k(\phi) := \frac{I_{m_0}^{(\Theta_k,0)}(\phi) + I_{m_{1:F}}^{(\Theta_k,1:\ell_F)}(\phi)}{I_{m_0}^{(\Theta_k,0)}(1) + \epsilon},$$

which are allowed to take negative values [cf. A]. This would provide an efficient MLMC alternative method to the PMMH or the algorithm in [53] for inference with respect to $\pi^{(l_F)}$. Instead, we wish to go one (infinite!) step further, and target $\pi^{(\infty)}$.

4.4. **Debiasing techniques.** Debiased MLMC [64, 77, 96] is based on randomising the running level used in deterministic MLMC (with a reweighting), as follows.

We assume that $(p_\ell)_{\ell\geq 1}$ is a probability mass function (p.m.f.) on $\mathbb{N}$ satisfying $p_\ell > 0$ for all $\ell \geq 1$. We also assume that

$$p_u^{(\theta,\ell)}(\phi) \xrightarrow{\ell\to\infty} p_u^{(\theta,\infty)}(\phi),$$

for all bounded $\phi : \mathsf{X}^{n+1} \to \mathbb{R}$, which is not too difficult to verify in our setting under certain boundedness assumptions, because of the known convergence properties of the Euler approximation [cf. 56]. With $L \sim (p_\ell)$, the *single-term debiased MLMC estimator* of [77] in our case is given by $p_L^{-1}\Delta^{(\theta,L)}(\phi)$, which satisfies

$$\mathbb{E}[p_L^{-1}\Delta^{(\theta,L)}(\phi)] = p_u^{(\theta,\infty)}(\phi) - p_u^{(\theta,0)}(\phi).$$

Adding an independent 'zeroth level' estimate $\hat{p}_u^{(\theta,0)}(\phi) := \sum_{i=1}^N V^{(i)}\phi(\mathbf{X}^{(i)})$, formed from the output $(V^{(i)}, \mathbf{X}^{(i)})_{i=1}^N$ of PF (Algorithm 1) run for the model $(M_p^{(\theta,0)}, G_p^{(\theta)})_{p=0}^n$, we set

$$\tilde{\Delta}^{(\theta)}(\phi) := \frac{1}{p_L}\Delta^{(\theta,L)}(\phi) + \hat{p}_u^{(\theta,0)}(\phi), \tag{33}$$

to obtain that

$$\mathbb{E}[\tilde{\Delta}^{(\theta)}(\phi)] = p_u^{(\theta,\infty)}(\phi).$$

By using a self-normalised estimator to take care of the normalising constant, this already allows for consistent inference over the latents. That is, as in (12) of Section 2.3, if $\{\tilde{\Delta}_k^{(\theta)}\}_{k=1}^m$ for $m \geq 1$ are independently run to form estimator functionals of the form (33), then

$$\frac{\sum_{k=1}^m \tilde{\Delta}_k^{(\theta)}(\phi)}{\sum_{k=1}^m \tilde{\Delta}_k^{(\theta)}(1)} \longrightarrow p^{(\theta,\infty)}(\phi), \qquad \text{almost surely,}$$

as $m \to \infty$ [C, Prop. 7].

4.5. **Joint inference using importance sampling type correction.** Recall that our original goal was joint ∞-inference (for $\pi^{(\infty)}$). To do this, we will use Algorithm 6, which is similar to Algorithm 4, but which uses a multilevel IS type correction based on the randomised ΔPF output. Consistency was also detailed in [A] for IS type correction involving negative weights as in Algorithm 6, which can occur frequently in the multilevel context which we consider here.

Algorithm 6 MCMC-IS for joint ∞-inference for diffusions based on debiased IS type correction, with $m \geq 1$, $\epsilon \geq 0$, p.m.f. (p_ℓ) on $\mathbb{N}$, and $N_\ell \geq 1$ for all $\ell \geq 0$.

(P1) With $(\Theta_0, V_0^{(i)}, \mathbf{X}_0^{(i)})_{i=1}^{N_0}$ given, for $k = 1, \ldots, m$, do:

(i) Sample $\Theta' \sim q(\,\cdot\,|\Theta_{k-1})$ from a transition kernel q.

(ii) Run PF (Algorithm 1) for $(M_p^{(\Theta',0)}, G_p^{(\Theta')})$, with output $(V'^{(i)}, \mathbf{X}'^{(i)})_{i=1}^{N_0}$.

(iii) Accept, setting $(\Theta_k, V_k^{(i)}, \mathbf{X}_k^{(i)})_{i=1}^{N_0} \leftarrow (\Theta', V'^{(i)}, \mathbf{X}'^{(i)})_{i=1}^{N_0}$, with probability

$$\min\left\{1, \frac{\mathrm{pr}(\Theta')\big(\epsilon + \sum_{i=1}^{N_0} V'^{(i)}\big) q(\Theta_{k-1}|\Theta')}{\mathrm{pr}(\Theta_{k-1})\big(\epsilon + \sum_{i=1}^{N_0} V_{k-1}^{(i)}\big) q(\Theta'|\Theta_{k-1})}\right\}.$$

Otherwise, reject, setting $(\Theta_k, V_k^{(i)}, \mathbf{X}_k^{(i)})_{i=1}^{N_0} \leftarrow (\Theta_{k-1}, V_{k-1}^{(i)}, \mathbf{X}_{k-1}^{(i)})_{i=1}^{N_0}$.

(P2) For all $k \in \{1{:}m\}$,

(i) Sample $L_k \sim (p_\ell)$.

(ii) Run ΔPF (Algorithm 5) for $(\check{M}^{(\Theta_k, L_k)}, \check{G}^{(\Theta_k)})$ with N_{L_k} particles, outputting $\Delta^{(\Theta_k, L_k)}$.

With $\xi_k(\phi)$ defined for $\phi : \mathsf{X}^{n+1} \to \mathbb{R}$, form the estimator

$$E_m^{IS}(f) := \frac{\sum_{k=1}^m \xi_k(f^{(\Theta_k)})}{\sum_{k=1}^m \xi_k(1)}, \quad \xi_k(\phi) := \frac{p_{L_k}^{-1}\Delta^{(\Theta_k, L_k)}(\phi) + \sum_{i=1}^{N_0} V_k^{(i)} \phi(\mathbf{X}_k^{(i)})}{\epsilon + \sum_{i=1}^{N_0} V_k^{(i)}}.$$

The likelihood support condition (17) mentioned for Algorithm 4 can be achieved by using $\epsilon > 0$.[22] We also need finiteness of the variance of the randomised ΔPF, $p_L^{-1}\Delta^{(\theta,L)}$, in order to guarantee that the debiased MLMC works correctly [cf. 77], and that the MCMC-IS (Algorithm 6) can have finite asymptotic variance [cf. C, Prof. 13]. That is, we need to show that

$$\mathrm{var}\left(\frac{1}{p_L}\Delta^{(\theta,L)}(\phi)\right) = \sum_{\ell \geq 1} \frac{\mathbb{E}\big[(\Delta^{(\theta,\ell)}(\phi))^2\big]}{p_\ell} - \Big(p_u^{(\theta,\infty)}(\phi) - p_u^{(\theta,0)}(\phi)\Big)^2 \tag{34}$$

is finite, uniformly in $\theta \in \mathsf{T}$. This requires showing that the variance of $\Delta^{(\theta,\ell)}(\phi)$ decays at a sufficient rate relative to p_ℓ as ℓ increases.

Under some standard (stringent) assumptions used elsewhere in the literature, the results of the technical analysis are formulated in [C, Cor. 9]. In the case of standard Euler approximation, the result says that

$$\mathbb{E}\big[(\Delta^{(\theta,\ell)}(\phi))^2\big] \leq C\Big(\frac{2^{-\ell}}{N_\ell} + 2^{-2\ell}\Big), \tag{35}$$

where $C > 0$ is a constant which does not depend on $N_\ell \geq 1$, $\ell \geq 1$, or $\theta \in \mathsf{T}$, where N_ℓ particles are used in the ΔPF run at level ℓ. Hence, with $N_\ell = N$ constant, by taking $p_\ell \propto 2^{-r\ell}$, with $r < 1$, (34) will be finite. More generally, if $N_\ell \propto 2^{\rho\ell}$ with $\rho \in [0,1]$, then we see that we can take $p_\ell \propto 2^{-r\ell}$ with $r < 1 + \rho$, so that (34) will be finite.

The assumptions needed to prove the bound (35) in [C] are on the diffusion [e.g. 56], in terms of uniform ellipticity and globally Lipschitz diffusion terms, as well as on the Feynman-Kac model [e.g. 18], in terms of globally Lipschitz potentials and transitions and lower and upper bounded potentials. The results of the analysis are based on a global error martingale decomposition [cf. 18] in terms of the local

[22] It is closely linked to 'defensive importance sampling' [51], but its optimal choice in terms of efficiency is not known.

sampling error of the particle filter run for the coupled Feynman-Kac model, and on an analysis of the ΔPF in the diffusion context.

4.6. **Computational efficiency and allocations.** We have seen that under some assumptions, the finiteness of the variance of the randomised ΔPF can be verified for any $N_\ell \geq 1$ and for sufficiently heavy-tailed (p_ℓ) [C, Cor. 9]. However, the use of a heavy-tailed p.m.f. (p_ℓ) can lead to excessive use of computational resources, and we must therefore try to use thinner-tailed p.m.f.s (p_ℓ) and optimal number of particles N_ℓ at level ℓ in order to minimise the inverse relative efficiency (IRE) [40] which measures the computational cost.

Let $(\Theta_k)_{k\geq 1}$ be the marginal Markov chain of Algorithm 6, and $L_k \sim (p_\ell)$ for $k \geq 1$. With terminology similar to [40], who consider the i.i.d.[23] case for $(\tau_k)_{k\geq 1}$, we assume that the *total computational cost* to run Algorithm 6 for m iterations is

$$\mathscr{C}(m) := \sum_{k=1}^{m} \tau_k,$$

where $(\tau_k)_{k\geq 1}$ are conditionally independent positive random variables given $(\Theta_k, L_k)_{k\geq 1}$, where τ_k depends only on Θ_k and L_k. Given some *budget* $\kappa \in \mathbb{R}_{\geq 0}$, the *realised length* of the chain is

$$\mathscr{M}(\kappa) := \max\left\{m \in \mathbb{N}_{\geq 0} | \mathscr{C}(m) \leq \kappa\right\}.$$

Then, if for some number $\tau > 0$,

$$\frac{1}{m}\sum_{k=1}^{m} \tau_k \xrightarrow{m\to\infty} \tau, \qquad \text{almost surely,}$$

and if the MCMC-IS estimator satisfies a CLT with asymptotic variance $\sigma^2(f)$, then [40, C]

$$\sqrt{\kappa}\left[E^{IS}_{\mathscr{M}(\kappa)}(f) - \pi^{(\infty)}(f)\right] \xrightarrow{\kappa\to\infty} \mathrm{N}\big(0, \tau\sigma^2(f)\big), \qquad \text{in distribution,}$$

and $\tau\sigma^2(f)$ is the IRE. We thus extend the discussion of [40] to non-i.i.d. $(\tau_k)_{k\geq 1}$.

Using this computational efficiency framework, similar to [77] who consider the i.i.d. case in traditional MLMC, it is possible to consider the matter of computational complexity and optimal allocation of resources in Algorithm 6. Suppose a CLT holds for the MCMC-IS estimator of Algorithm 6 with finite asymptotic variance [cf. C, Prop. 13]. Let $\epsilon > 0$ and $\delta \in (0, 1)$ be given. In order to have

$$\mathbb{P}[|E^{IS}_m(f) - \pi^{(\infty)}(f)| \leq \epsilon] \geq 1 - \delta,$$

by the Chebyshev inequality and using the standard m^{-1} mean squared error convergence rate for MCMC, we need that m is of order ϵ^{-2}, denoted $m = O(\epsilon^{-2})$.[24] The question is then how we can minimise the computational complexity given by $\mathscr{C}(m)$ when $m = O(\epsilon^{-2})$, by adjusting p_ℓ and N_ℓ, while keeping the variance (34) finite.[25] Assuming

$$\mathbb{E}[\tau_k | \Theta_k = \theta,\, L_k = \ell] \leq C 2^{\ell(1+\rho)},$$

[23] independent and identically distributed

[24] That is, with $m = m(\epsilon)$, we have $m(\epsilon)/\epsilon^{-2} \to C$ as $\epsilon \to 0$, some $C > 0$.

[25] Besides for the debiasing [77] to work, the asymptotic variance [see C, Prop. 13] of the MCMC-IS estimator of Algorithm 6 has a part from the marginal MCMC, as well as from the IS type correction. The latter is finite if the variance (34) is finite.

where C does not depend on θ or ℓ, then it is shown in [C, Prop. 24] that for all $q > 2$, $\eta > 1$, the computational cost

$$O\big(\epsilon^{-2}|\log_2 \epsilon|^q\big) \tag{36}$$

can be obtained for sufficiently small ϵ, if p_ℓ and N_ℓ are chosen to be

$$p_\ell \propto 2^{-\ell(1+\rho)}\ell[\log_2(\ell+1)]^\eta \qquad \text{and} \qquad N_\ell \propto 2^{\rho\ell} \tag{37}$$

for $\rho \in [0,1]$. This choice for p_ℓ and N_ℓ ensures that the variance (34) is finite, and suggests[26] the choice

$$p_\ell \propto 2^{-\ell(1+\rho)} \qquad \text{and} \qquad N_\ell \propto 2^{\rho\ell} \tag{38}$$

for $\rho \in [0,1]$. The computational cost (36) is the same as that of [77, Prop. 4] for the single-term estimator in traditional, randomised MLMC. It is also very close to

$$O\big(\epsilon^{-2}(\log_2 \epsilon)^2\big),$$

(recall $q > 2$), which is the well-known computational complexity order [36] in the traditional, deterministic MLMC.

The result (37) shows that in case of Euler approximation, there is in fact a parametrisation of recommended choices for particle number N_ℓ and p.m.f. (p_ℓ), all of which share the same order of computational complexity to obtain a given precision, under certain assumptions such as previously explained for τ_k. Then (37) (or the simplified suggestion (38)) should lead to a proper usage of computational resources, in order to keep both the asymptotic variance and the total cost jointly small, and therefore the IRE small. The *order* of computational complexity is the same along the parametrisation in terms of $\rho \in [0,1]$, but it is still unknown whether a certain choice of ρ will usually lead to the best choice for N_ℓ and (p_ℓ). In an experiment in [C] concerning a geometric Brownian motion, the choice $\rho = 0$ performed better than the choice $\rho = 1$ in the allocation (37). We leave, for now, the optimal choice of ρ for future research and experiment.

5. Inference via approximate Bayesian computation

We assume a Bayesian model as in Section 1.2, with fixed observation denoted $y^* \in \mathsf{Y}$, prior $\mathrm{pr}(\theta)$, and likelihood $L(\theta) = p^{(\theta)}(y^*)$, which is assumed to be intractable. Although the data distribution $p^{(\theta)}(\,\cdot\,)$ can not be evaluated, we assume that it is possible to sample data $y \sim p^{(\theta)}(\,\cdot\,)$ from it. Let $d(y, y')$ be a pseudo-metric[27] on Y^2. With *tolerance* $\epsilon > 0$, we then define

$$p_u^{(\theta,1/\epsilon)}(\mathrm{d}y) := p^{(\theta)}(\mathrm{d}y)\mathbf{1}\big(d(y,y^*) \le \epsilon\big),^{28}$$

Approximate Bayesian computation (ABC) (see [92] for a review) is based on using the family $\mathscr{P}_{1/\epsilon} := \{p^{(\theta,1/\epsilon)}\}_{\theta\in\mathsf{T}}$ of approximate probabilities, where

$$p^{(\theta,1/\epsilon)}(\mathrm{d}y) := \frac{p_u^{(\theta,1/\epsilon)}(\mathrm{d}y)}{L^{(1/\epsilon)}(\theta)},$$

[26] by disregarding the factor $\ell[\log_2(\ell+1)]^\eta$ in (37)

[27] That is, for all $y_1, y_2, y_3 \in \mathsf{Y}$, it holds $d(y_1,y_2) \ge 0$, $d(y_1,y_2) = d(y_2,y_1)$, and $d(y_1,y_3) \le d(y_1,y_2)+d(y_2,y_3)$. For example, $d(y_1,y_2) = \|s(y_1) - s(y_2)\|$ where $s : \mathsf{Y} \to \mathbb{R}^{n_y}$ is some (summary) statistic [cf. 76].

[28] The quantity '$1/\epsilon$' can be thought of as denoting the level of 'precision.'

with ABC likelihood,

$$L^{(1/\epsilon)}(\theta) := \int p^{(\theta)}(\mathrm{d}y)\mathbf{1}\big(d(y, y^*) \leq \epsilon\big).$$

Then $\mathscr{P}_{1/\epsilon}$ become families of increasingly 'better' approximations as ϵ goes to 0. However, it is important to keep in mind that it is only approximate even in the limit, since in general,

$$L^{(\infty)}(\theta) := \lim_{\epsilon \to 0} L^{(1/\epsilon)}(\theta) \neq L(\theta).^{29}$$

The ABC posterior is then given by

$$\pi^{(1/\epsilon)}(\theta) \propto \mathrm{pr}(\theta) L^{(1/\epsilon)}(\theta).$$

A method of inference for the ABC posterior which we consider is the ABC-MCMC (Algorithm 7), as suggested by [62], which may also be viewed as a pseudo-marginal MCMC [5], with

$$\mathbb{E}_\theta\big[\mathbf{1}\big\{d(Y, y^*) \leq \epsilon\big\}\big] = L^{(1/\epsilon)}(\theta).$$

5.1. **Choosing the tolerance in ABC-MCMC.** The choice of tolerance ϵ is a difficult question in ABC-MCMC [cf. 91]. Namely, a large choice of ϵ leads to large bias, but to computational inefficiency if ϵ is small. To see this, note that if ϵ is small, then a proposed state is hardly ever accepted, since $\mathbf{1}\big\{d(Y'_k, y^*) \leq \epsilon\big\}$ is usually 0. If ϵ is large, then $L^{(1/\epsilon)}(\theta) \approx 1$ is nearly constant in θ and so ABC-MCMC is essentially targeting the prior model, which is uninformative for Bayesian posterior inference.

Article [D] attempts to deal with the issue of tolerance choice in ABC-MCMC, by using an inflated and adaptively tuned tolerance parameter in order to maximise efficiency of the MCMC, and then to use a post-correction, importance sampling step, to remove bias [98] as well as to quantify uncertainty with proposed approximate confidence intervals.

The tolerance adaptive ABC-MCMC (Algorithm 8), which is run during burn-in for some number of iterations n_b, is an adaptive MCMC [cf. 6] targeting a user-specified overall acceptance probability $\alpha^* \in (0, 1)$. In experiments in [D], a value of $\alpha^* = 10\%$ was used, which ensures sufficient mixing and number of different samples from the MCMC. We provide convergence theorems in [D] for the adaptive algorithm under two sets of assumptions. The simpler set of assumptions essentially requires that the proposal $q(\theta'|\theta) > 0$ is uniformly bounded away from zero, and ϵ_k is bounded away from zero for all $k \geq 1$ almost surely. The former assumption on q is removed in the more general set of assumptions. Removing the assumption on ϵ_k might be possible, based on projections [cf. 4].

5.2. **Approximate confidence intervals.** An approximate estimator for the asymptotic variance of the post-corrected ABC-MCMC has been suggested in [D, Alg. 6], which can be used for the construction of (approximate) confidence intervals.

Suppose that $\hat{\tau}_{\epsilon_0}(f)$ is an estimate for the integrated auto-correlation time for ABC-MCMC(ϵ_0),

$$\tau_{\epsilon_0}(f) := \sum_{k \geq 1} \mathrm{Corr}\big(f(\vartheta_0), f(\vartheta_k)\big), \qquad \vartheta_0 \sim \pi^{(1/\epsilon_0)}(\,\cdot\,), \tag{40}$$

[29] This is in general the case. However, there can be equality if, for example, $d(y, y')$ is a metric, or, in particular, a metric formed from composition of a sufficient statistic with a Euclidean norm $\|\cdot\|$ [cf. 76].

Algorithm 7 ABC-MCMC(ϵ). Given $\Theta_0 \in \mathsf{T}$ with $\mathrm{pr}(\Theta_0) > 0$, run the following for $k = 0, \dots n-1$:

(i) Sample $\Theta'_k \sim q(\,\cdot\,|\Theta_k)$.
(ii) Sample $Y'_k \sim p^{(\Theta'_k)}(\,\cdot\,)$.
(iii) Accept, setting $(\Theta_{k+1}, Y_{k+1}) \leftarrow (\Theta'_k, Y'_k)$, with probability $\alpha_\epsilon(\Theta_k, \Theta'_k, Y'_k)$, where

$$\alpha_\epsilon(\theta, \theta', y') := \min\left\{1, \frac{\mathrm{pr}(\theta')q(\theta|\theta')}{\mathrm{pr}(\theta)q(\theta'|\theta)}\right\}\mathbf{1}\big\{d(y', y^*) \le \epsilon\big\}. \tag{39}$$

Else, reject, by setting $(\Theta_{k+1}, Y_{k+1}) \leftarrow (\Theta_k, Y_k)$.

Algorithm 8 TA(n_b). Given $\Theta_0 \in \mathsf{T}$ with $\mathrm{pr}(\Theta_0) > 0$, $\epsilon_0 := d(Y_0, y^*) > 0$ with $Y_0 \in \mathsf{Y}$, $\alpha^* = .1$, and step sizes $\gamma_k = k^{-2/3}$.

For $k = 0, \dots, n_b - 1$,

(i) Sample $\Theta'_k \sim q(\,\cdot\,|\Theta_k)$.
(ii) Sample $Y'_k \sim p^{(\Theta'_k)}(\,\cdot\,)$.
(iii) Accept, setting $(\Theta_{k+1}, Y_{k+1}) \leftarrow (\Theta'_k, Y'_k)$, with probability $\alpha_{\epsilon_k}(\Theta_k, \Theta'_k, Y'_k)$, with α_ϵ defined in (39). Otherwise, reject, setting $(\Theta_{k+1}, Y_{k+1}) \leftarrow (\Theta_k, Y_k)$.
(iv) $\log \epsilon_{k+1} \leftarrow \log \epsilon_k + \gamma_k\big(\alpha^* - \alpha_{\epsilon_k}(\Theta_k, \Theta'_k, Y'_k)\big)$.

Output $(\Theta_{n_b}, \epsilon_{n_b})$.

perhaps using a windowed sample auto-correlation estimator [cf. 47]. Also define the following estimator for the function variance,

$$S_{\epsilon_0,\epsilon}(f) := \sum_{k=1}^{n} \frac{\mathbf{1}\big(d(Y_k, y^*) \le \epsilon\big)\big(f(\Theta_k) - E_{\epsilon_0,\epsilon}(f)\big)^2}{\big(\sum_{j=1}^{n} \mathbf{1}\big(d(Y_j, y^*) \le \epsilon\big)\big)^2}. \tag{41}$$

The approximate confidence interval then takes the form

$$\Big[E_{\epsilon_0,\epsilon}(f) \pm \beta\sqrt{\hat{\tau}_{\epsilon_0}(f) S_{\epsilon_0,\epsilon}(f)}\Big],$$

where $\beta > 0$ corresponds to the standard normal quantile.

We remark that there is some theoretical backing for the approximate confidence interval, based on an exact formula for the integrated auto-correlation time of the post-corrected chain [D, Thm. 7]. The relevance of the approximate confidence intervals is also verified in some experiments in [D].

5.3. **Adaptive ABC-MCMC with post-correction.** The approach of [D] then takes the form of Algorithm 9. In regards to the adaptive ABC-MCMC, also the proposal covariance matrix q is best updated as in [46, 3]. The estimator $E_{\epsilon_0,\epsilon}(f)$ can be calculated effortlessly for all $\epsilon \in (0, \epsilon_0]$ by sorting beforehand the samples Θ_k according to their corresponding distances T_k.

In experiments in [D], for example in a Lotka-Volterra model involving two reagents and three reactions [cf. 13], it was found that Algorithm 9 delivers a robust approach to inference in ABC models. In particular, the post-processing estimators were found to be competitive with direct ABC-MCMC with pre-tuned tolerance and starting value, the approximate confidence interval provided good coverage, and the adaptive ABC-MCMC allowed for essentially arbitrary initial choice of tolerance and starting value from the prior.

Algorithm 9 Given n_b, $n \geq 1$ perform the following:

(i) Run TA(n_b) (Algorithm 8), and call the output (Θ_0, ϵ_0).
(ii) Run ABC-MCMC(ϵ_0) (Algorithm 7) for n iterations, with starting values (Θ_0, ϵ_0), outputting $(\Theta_k, Y_k)_{k=1}^n$.
(iii) For all $\epsilon \leq \epsilon_0$, an estimator for $\pi^{(1/\epsilon)}(f)$ is given by

$$E_{\epsilon_0,\epsilon}(f) := \frac{\sum_{k=1}^n \mathbf{1}\{d(Y_k, y^*) \leq \epsilon\} f(\Theta_k)}{\sum_{k=1}^n \mathbf{1}\{d(Y_k, y^*) \leq \epsilon\}}.$$

(iv) With $\hat{\tau}_\epsilon(f)$ an estimate of (40), $S_{\epsilon_0,\epsilon}(f)$ calculated as in (41), and $\beta > 0$ corresponding to the desired standard normal quantile, report the approximate confidence interval

$$\left[E_{\epsilon_0,\epsilon}(f) \pm \beta\sqrt{\hat{\tau}_{\epsilon_0}(f) S_{\epsilon_0,\epsilon}(f)}\right].$$

Compared to direct ABC-MCMC(ϵ), the approach based on slightly inflated tolerance and post-correction, was shown to be competitive in experiments in [D]. An upper bound for the asymptotic variance of the ABC-MCMC(ϵ_0) with post-correction to ϵ, in terms of that of a direct ABC-MCMC(ϵ), is given in [D, Thm. 8]. It is a direct application of the Peskun type ordering for importance sampling [B] stated previously in (23), where the upper bound guarantee becomes an equality as $|\epsilon_0 - \epsilon| \to 0$.

5.4. **Convergence of the tolerance adaptive ABC-MCMC.** We briefly discuss the general approach to the convergence proofs of the tolerance adaptive ABC-MCMC. To obtain a setup fitting within the framework of stochastic approximation [cf. 3, 4], we write the tolerance adaptation update in Algorithm 8(iv) as

$$\begin{aligned}\log \epsilon_{k+1} &= \log \epsilon_k + \gamma_{k+1} H_{\epsilon_k}(\Theta_k, \Theta'_k, Y'_k) \\ &= \log \epsilon_k + \gamma_{k+1} h(\epsilon_k) + \gamma_{k+1}\eta_{k+1},\end{aligned}$$

where $H_\epsilon(\theta, \theta', y') := \alpha^* - \alpha_\epsilon(\theta, \theta', y')$, with α_ϵ defined in (39), with 'mean field'

$$h(\epsilon) := \int \pi^{(\epsilon)}(\mathrm{d}\theta) q(\theta, \mathrm{d}\theta') p^{(\theta')}(\mathrm{d}y') H_\epsilon(\theta, \theta', y'),$$

and centred 'noise' sequence $\eta_{k+1} := H_{\epsilon_k}(\Theta_k, \Theta'_k, Y'_k) - h(\epsilon_k)$. In this common framework for stochastic approximation algorithms, we can apply [4, Theorem 2.3], which implies that the key lemma for the proof of convergence of the tolerance adaptive ABC-MCMC (Algorithm 8) essentially reduces to showing that the noise sequence η_k is asymptotically controlled,

$$\lim_{j\to\infty} \sup_{n \geq j} \left| \sum_{k=j}^n \gamma_k \eta_k \right| = 0, \qquad \text{almost surely,}$$

[D, Lemma 20]. This relies on various ancillary results, such as monotonicity of the map $\epsilon \mapsto h(\epsilon)$, continuity and contraction properties of the Markov kernels, and a generalisation of the 'proposal augmentation' from Metropolis-Hastings chains [85, 82] to 'proposal-rejection' chains. Here, we call a kernel K a 'proposal-rejection' kernel if it is reversible and can be written as

$$K(\theta, \mathrm{d}\theta') = q(\mathrm{d}\theta'|\theta)\alpha(\theta, \theta') + \left(1 - \int q(\mathrm{d}\vartheta|\theta)\alpha(\theta, \vartheta)\right)\delta_\theta(\mathrm{d}\theta'), \tag{42}$$

where $\alpha(\theta, \theta') \in [0, 1]$ is a measurable function on T^2. By marginalising away the auxiliary variable in ABC-MCMC(ϵ) (7), we obtain such a 'proposal-rejection' kernel, with

$$\alpha(\theta, \theta') = \min\left\{1, \frac{\mathrm{pr}(\theta')q(\theta|\theta')}{\mathrm{pr}(\theta)q(\theta'|\theta)}\right\} L^{(1/\epsilon)}(\theta'),$$

which is clearly not a Metropolis-Hastings kernel any longer.

Non-standard theoretical challenges of the tolerance adaptive ABC-MCMC (Algorithm 8) are that the invariant measure $\pi^{(1/\epsilon_k)}$ is changing at each iteration, and that the chain is technically a pseudo-marginal. Regarding this latter point, however, we do have simplification to independent refreshments of the auxiliary variable y', because of the use of a simple cut-off function $\mathbf{1}\{d(\,\cdot\,, y^*) \leq \epsilon\}$ in the acceptance ratio. As mentioned in Section 5.1, essentially, the convergence theorems for the adaptation are formulated in a simpler setting of uniform ergodicity, as well as for simultaneously geometrically ergodic 'proposal-rejection' chains, obtained by only considering the marginal chain $(\Theta_k)_{k\geq 1}$ of the original chain $(\Theta_k, Y_k)_{k\geq 1}$, on possibly unbounded state space domains.

6. Discussion and directions for future work

In this thesis, various old and new Monte Carlo estimators are presented. A defining feature of the estimators suggested is that they involve an IS type correction of samples drawn according to an intermediate approximate distribution. Basic convergence properties of the suggested estimators are established, and efficiency of these algorithms is studied and related to standard direct methods used hitherto commonly in practice.

There is still much interesting work that could be done in regards to the use of these estimators in different settings and with different approximations. Experimental results have been promising, and suggest further comparisons could be made, for example, of PIMH and its IS analogue (12). In the parameter inference setting, there have been many MCMC implementations making use of an approximation by applying delayed acceptance [see B, Section 7.2], but very few using MCMC-IS (see [73] for one other non-academic example). One of the main goals of [A] is to bring attention to MCMC-IS, that it represents a viable approach, which enjoys flexibility in implementation and theoretical backing.

In the filtering and smoothing context, the approach for optimal selection of Feynman-Kac model for the smoothing problem [45] based on deterministic approximations, as used in [A] and further developed in [60] for an extended class of models, could be further developed. These approximations could also be based on various other non-linear filters and smoothers [cf. 84].

There are various directly applicable innovations which could be incorporated into MCMC-IS, and we mention a few. Quasi-Monte Carlo may be helpful in MCMC-IS, whether in the MCMC [cf. 86] or in the PF [35]. Work on exact simulation [12] techniques for diffusions [10, 11, 31] (see also the recent preprint [97]) and jump-diffusions [43] using continuous-time IS techniques is showing progress, and suggests parameter inference methods for partially observed versions could be developed, at least in the one-dimensional setting, using the MCMC-IS framework, with IS correction based on a PF using exact simulation dynamics, or based on other types of randomised weights, which may freely assume negative values in the IS correction.

It would be of interest to adapt the tuning guidelines [29] (see also [89]) for the PF when used in the PMMH, to the case when used within MCMC-IS. The formula (26)

for the MCMC-IS asymptotic variance, which decomposes into marginal chain and IS correction parts, could also be useful in this regard. More generally, beyond PMMH, it would be beneficial to use better scaling MCMCs within MCMC-IS, for example, particle Gibbs [1], which is known to scale very well with backward sampling [cf. 57]. Additional annealing steps may be useful, as part of the Metropolis-Hastings with asymmetric acceptance ratio (MHAAR) approach [cf. 2]

In [B], more practical examples could be given showing bounds of likelihood ratios and usefulness of the results in practice. Further comparisons could be made, for example, with annealed IS [71] correction versus multi-stage DA [8]. Two different extensions of traditional DA correction were introduced in [B], and it would be interesting to study the stability properties of these new DA corrections, for example, along the lines of [5, 88]. Other more sophisticated reversible chains as in [2] with IS correction could be considered and compared. The effect of debiasing tricks [39] could be compared between MCMC-IS and pseudo-marginal type MCMC, where the coupling time integral to the debiasing approach may be considerably less for MCMC-IS if Phase 1 is based on deterministic approximation and Phase 2 involves noisy unbiased estimators.

There are many settings where there is a multilevel type structure and the debiasing techniques can be applied. In the joint inference setting, the IS-debiasing method as presented in [C] allows for an efficient debiasing strategy for joint inference using Euler approximations. The results could be generalised to Itô diffusions with time-dependent path-dependent coefficients, and to general resampling schemes in the ΔPF besides multinomial resampling. It would be nice to apply the IS-debiasing strategy in various settings, for example, to jump-diffusions [cf. 22, 54]. The coupling [53] and multilevel approach to the (unnormalised) smoothing problem [C], with possible randomised MLMC correction [64, 77], could be applied, for example, to the problem of calculation of normalisation constants [cf. 20] important for model selection. The optimal choice of coupled dynamics and potential could be studied, where we remark that the coupled potentials may be made level dependent, which is an additional degree of freedom. It would also be of interest to study stability and limit theorems [15, 18, 26] of these coupled PFs [cf. 55] based on change of reference measure and IS reweighting for use in unnormalised multilevel estimators as in [C].

We are currently looking into optimal tuning of the regularisation constant in the approximate likelihood estimator within MCMC-IS, which is connected to 'defensive importance sampling' [51]. The question of efficiency and proper allocation of resources of the MCMC-IS carries over to the multilevel and PF setting, where additionally multilevel aspects play a rôle. The question of optimal scaling particles versus level in the sub-canonical regime associated to Euler approximations was not entirely conclusive in [C]. It would be interesting to study this phenomenon in more depth. This may entail adapting the non-canonical CLT of [99] in the diffusion setting to the partially observed diffusion setting where number of particles and particle approximation variances are additional factors.

Applied in the ABC context in [98], the post-correction (or trimming) over a range of tolerances is a methodological approach applicable in other Monte Carlo settings where IS can be applied at small additional cost, for example, in the MLMC context, with the sum of multilevel increments computed sequentially over an increasing range of the fine tolerances, with corresponding plots. In such settings, it may also be possible to derive analogous approximate confidence intervals for the resulting

estimators as in [D]. The tolerance adaptive ABC-MCMC in [D] was based on targeting a user-specified overall acceptance probability, and we chose a target close to the rule of thumb from the more general random walk PM literature [89]. It may be interesting to adapt the assumptions of [29, 89] to ones more resembling the ABC context, in order to find a perhaps different 'rule of thumb' for ABC-MCMC. The tolerance adaptation was also found to benefit the covariance adaptation during the burn-in, likely due to the improved mixing in the initial stages of the algorithm. It would be of interest to study this phenomenon and the interplay of different optimisation criteria in more depth, following, for example, the theoretical developments of adaptive MCMC as in [3, 4, 6].

The 'proposal-rejection' chains (42), which were considered for DA correction [B] and 'proposal augmentation' [D], are generalisations of Metropolis-Hastings chains, which include DA [8], PM [5], MHAAR [2] and marginalised ABC-MCMC [D]. Although 'proposal-rejection' chains technically include pseudo-marginal chains, we lay here particular emphasis on the possible course of study of (simpler) chains on the marginal (parameter) space, without auxiliary variable extensions like in pseudo-marginal MCMC. Many results worked out for Metropolis-Hastings can likely be extended to the marginal-space 'proposal-rejection' setting. For example, the waste-recyclers of [21, 82, 85], originally for Metropolis-Hastings, could be extended to 'proposal-rejection' chains. Some convergence analysis has been done for pseudo-marginal Metropolis-Hastings chains [cf. 5, 7] and some of this type of analysis could possibly be adapted to marginal-space 'proposal-rejection' chains. Following the line of argument of [52, 78], who show geometric ergodicity of symmetric random walk Metropolis-Hastings essentially if the target has exponential or lighter tails and a certain contour condition holds, it would be interesting to work out conditions for a similar type of result for the more general sub-class of marginal-space 'proposal-rejection' chains.

7. Summary of articles

7.1. **Article [A].** Convergence properties are established for Markov chain Monte Carlo (MCMC) algorithms using an additional importance sampling (IS) type correction of approximate sample output of the Markov chain. Included is the interesting case where the approximate chain itself stems from a pseudo-marginal chain. The asymptotic variance from the proven central limit theorems is shown to decouple over the approximate marginal chain and the IS correction, which can be useful for questions of optimal allocation of computational resources. Particular strengths of the approach are highlighted, such as the efficient use of a jump chain, thinning, and straightforward parallelisation. Abstract properties of the augmented Markov chains corresponding to the MCMC-IS method are established. Experiments in state space models compare the MCMC-IS method with existing popular direct methods, and show the viability of the MCMC-IS approach in the state space models context.

7.2. **Article [B].** The asymptotic variance of the MCMC-IS is compared to that of the direct MCMC methods. This is based on an extension of the existing covariance comparison result for direct chains to the context of comparison of one MCMC-IS to one direct chain. The extension also allows for use of unbiased estimators in the MCMC-IS Phase 1 and 2, as well as the use of a jump chain. Provided examples show that there can be no strict ordering between MCMC-IS and direct MCMC, as either may perform arbitrarily better than the other. Theoretical results are provided,

which show upper and lower bounds for the MCMC-IS asymptotic approach in relation to an analogous direct MCMC method. The upper bound is satisfied in practice when approximations are reasonably accurate, and provides guarantees for the MCMC-IS asymptotic variance in terms of direct pseudo-marginal and delayed acceptance analogues. In the latent variable setting, this is the case in the sense of finite supremum norm of the ratio of likelihoods. Ergodicity and mixing of the MCMC-IS is shown to be less affected by noise of the Phase 2 unbiased estimators compared to pseudomarginal direct MCMC. The results help justify the viability of the MCMC-IS approach as a competing method to a direct approach.

7.3. **Article [C].** The question of joint inference for a challenging class of state space models is considered, where the underlying process is a diffusion process arising as a solution to a stochastic differential equation, which can not be simulated exactly. Noisy non-linear observations are obtained at some discrete points in time. Bayesian inference is performed using the IS debiasing approach, where, namely, an IS type correction, based on debiased multilevel Monte Carlo, a particle filter coupling, and Euler approximations, is used for an approximate MCMC targeting a coarse-model approximate distribution. Convergence of the method to the exact posterior is verified under standard conditions on the state space model and Euler type approximations found in the literature. From asymptotic efficiency and cost considerations, suggested allocations for computational resources are given, which help ensure efficient use of the algorithm.

7.4. **Article [D].** The use of a slightly inflated tolerance is suggested in the context of approximate Bayesian computation (ABC) MCMC, along with subsequent post-correction based on trimming or IS correction of the sample output, over a (continuous) range of decreasing tolerances. Approximate confidence intervals for the resulting estimators are provided, which enjoy theoretical backing as well as good coverage in the experiments considered. An adaptive ABC-MCMC is also proposed, which finds a suitable (inflated) tolerance based on acceptance rate as the proxy. Convergence theorems for the adaptation under simple and more general conditions are provided. The tolerance adaptation worked well when used together with proposal covariance adaptation, in experiments which confirmed the suitability of the method based on adaptive ABC-MCMC and post-correction.

References

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

[2] C. Andrieu, A. Doucet, S. Yıldırım, and N. Chopin. On the utility of Metropolis-Hastings with asymmetric acceptance ratio. Preprint arXiv:1803.09527, 2018.

[3] C. Andrieu and É. Moulines. On the ergodicity properties of some adaptive MCMC algorithms. *Ann. Appl. Probab.*, 16(3):1462–1505, 2006.

[4] C. Andrieu, É. Moulines, and P. Priouret. Stability of stochastic approximation under verifiable conditions. *SIAM J. Control Optim.*, 44(1):283–312, 2005.

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

[6] C. Andrieu and J. Thoms. A tutorial on adaptive MCMC. *Statist. Comput.*, 18(4):343–373, Dec. 2008.

[7] C. Andrieu and M. Vihola. Convergence properties of pseudo-marginal Markov chain Monte Carlo algorithms. *Ann. Appl. Probab.*, 25(2):1030–1077, 04 2015.

[8] M. Banterle, C. Grazian, A. Lee, and C. P. Robert. Accelerating Metropolis-Hastings algorithms by delayed acceptance. Preprint arXiv:1503.00996v2, 2015.

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

[10] A. Beskos, O. Papaspiliopoulos, and G. O. Roberts. Retrospective exact simulation of diffusion sample paths with applications. *Bernoulli*, pages 1077–1098, 2006.

[11] A. Beskos, O. Papaspiliopoulos, G. O. Roberts, and P. Fearnhead. Exact and computationally efficient likelihood-based estimation for discretely observed diffusion processes. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 68(3):333–382, 2006. (with discussion).

[12] A. Beskos and G. Roberts. Exact simulation of diffusions. 15(4):2422–2444, 11 2005.

[13] R. J. Boys, D. J. Wilkinson, and T. B. Kirkwood. Bayesian inference for a discretely observed stochastic kinetic model. 18(2):125–135, 2008.

[14] O. Cappé, E. Moulines, and T. Rydén. *Inference in Hidden Markov Models*. Springer, 2005.

[15] N. Chopin. Central limit theorem for sequential Monte Carlo methods and its application to Bayesian inference. 32(6):2385–2411, 2004.

[16] J. A. Christen and C. Fox. Markov chain Monte Carlo using an approximation. *J. Comput. Graph. Statist.*, 14(4), 2005.

[17] T. Cui, Y. Marzouk, and K. Willcox. Scalable posterior approximations for large-scale Bayesian inverse problems via likelihood-informed parameter and state reduction. *J. Comput. Phys.*, 315:363–387, 2016.

[18] P. Del Moral. *Feynman-Kac Formulae*. Springer, 2004.

[19] P. Del Moral. *Mean field simulation for Monte Carlo integration*. Chapman & Hall/CRC, 2013.

[20] P. Del Moral, A. Jasra, K. J. H. Law, and Y. Zhou. Multilevel sequential Monte Carlo samplers for normalizing constants. *ACM Trans. on Modeling and Computer Simulation (TOMACS)*, 27(3):20, 2017.

[21] J.-F. Delmas and B. Jourdain. Does waste recycling really improve the multi-proposal Metropolis–Hastings algorithm? an analysis based on control variates. 46(4):938–959, 2009.
[22] S. Dereich. Multilevel Monte Carlo algorithms for Lévy-driven SDEs with Gaussian correction. 21(1):283–311, 2011.
[23] P. Diaconis. The Markov chain Monte Carlo revolution. *Bull. Amer. Math. Soc.*, 46(2):179–205, 2009.
[24] H. Doss. Discussion: Markov chains for exploring posterior distributions. *Ann. Statist.*, 22(4):1728–1734, 1994.
[25] R. Douc, O. Cappé, and E. Moulines. Comparison of resampling schemes for particle filtering. In *Proc. Image and Signal Processing and Analysis, 2005*, pages 64–69, 2005.
[26] R. Douc and E. Moulines. Limit theorems for weighted samples with applications to sequential Monte Carlo methods. In *ESAIM: Proceedings*, volume 19, pages 101–107. EDP Sciences, 2007.
[27] R. Douc, C. P. Robert, et al. A vanilla Rao-Blackwellization of Metropolis-Hastings algorithms. *Ann. Statist.*, 39(1):261–277, 2011.
[28] A. Doucet and A. Johansen. A tutorial on particle filtering and smoothing: Fifteen years later. In *Handbook of nonlinear filtering*, volume 12, pages 656–704. 2009.
[29] A. Doucet, M. Pitt, G. Deligiannidis, and R. Kohn. Efficient implementation of Markov chain Monte Carlo when using an unbiased likelihood estimator. *Biometrika*, 102(2):295–313, 2015.
[30] B. Efron and T. Hastie. *Computer age statistical inference*. Cambridge, 2016.
[31] P. Fearnhead, K. Łatuszyński, G. Roberts, and G. Sermaidis. Continious-time importance sampling: Monte Carlo methods which avoid time-discretisation error. Preprint arXiv:1712.06201, 2017.
[32] J. M. Flegal and G. L. Jones. Batch means and spectral variance estimators in Markov chain Monte Carlo. *Ann. Statist.*, 38(2):1034–1070, 2010.
[33] A. Gelman, J. Carlin, H. Stern, and D. Rubin. *Bayesian Data Analysis*. Chapman & Hall/CRC, 1995.
[34] H.-O. Georgii. *Stochastics: introduction to probability and statistics*. Walter de Gruyter, 2012.
[35] M. Gerber and N. Chopin. Sequential quasi-Monte Carlo. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 77(3):509–579, 2015.
[36] M. B. Giles. Multilevel Monte Carlo path simulation. *Oper. Res.*, 56(3):607–617, 2008.
[37] W. Gilks and G. Roberts. Strategies for improving MCMC. In *Markov chain Monte Carlo in practice*, volume 6, pages 89–114. 1996.
[38] P. W. Glynn and D. L. Iglehart. Importance sampling for stochastic simulations. *Management Science*, 35(11):1367–1392, 1989.
[39] P. W. Glynn and C.-H. Rhee. Exact estimation for Markov chain equilibrium expectations. 51(A):377–389, 2014.
[40] P. W. Glynn and W. Whitt. The asymptotic efficiency of simulation estimators. *Oper. Res.*, 40(3):505–520, 1992.
[41] A. Golightly, D. Henderson, and C. Sherlock. Delayed acceptance particle MCMC for exact inference in stochastic kinetic models. *Statist. Comput.*, 25, 2015.

[42] A. Golightly and D. Wilkinson. Bayesian parameter inference for stochastic biochemical network models using particle Markov chain Monte Carlo. *Interface focus*, 1(6):807–820, 2011.
[43] F. Gonçalves, K. Łatuszyński, and G. Roberts. Exact Monte Carlo likelihood-based inference for jump-diffusion processes. Preprint arXiv:1707.00332, 2017.
[44] N. J. Gordon, D. J. Salmond, and A. F. M. Smith. Novel approach to nonlinear/non-Gaussian Bayesian state estimation. *IEE Proceedings-F*, 140(2):107–113, 1993.
[45] P. Guarniero, A. Johansen, and A. Lee. The iterated auxiliary particle filter. *J. Amer. Statist. Assoc.*, 112(520):1636–1647, 2017.
[46] H. Haario, E. Saksman, and J. Tamminen. An adaptive Metropolis algorithm. *Bernoulli*, 7(2):223–242, 2001.
[47] F. Harris. On the use of windows for harmonic analysis with the discrete Fourier transform. *Proc. IEEE*, 66(1):51–83, 1978.
[48] W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109, Apr. 1970.
[49] S. Heinrich. Multilevel Monte Carlo methods. In *Large-scale scientific computing*, pages 58–67. Springer, 2001.
[50] J. Heng, A. Bishop, G. Deligiannidis, and A. Doucet. Controlled sequential Monte Carlo. Preprint arXiv:1708.08396v2, 2017.
[51] T. Hesterberg. Weighted average importance sampling and defensive mixture distributions. *Technometrics*, 37(2):185–194, 1995.
[52] S. F. Jarner and E. Hansen. Geometric ergodicity of Metropolis algorithms. *Stochastic Process. Appl.*, 85(2):341–361, 2000.
[53] A. Jasra, K. Kamatani, K. J. Law, and Y. Zhou. Bayesian static parameter estimation for partially observed diffusions via multilevel Monte Carlo. *SIAM J. Sci. Comp.*, 40:A887–A902, 2018.
[54] A. Jasra, K. J. Law, and P. P. Osei. Multilevel particle filters for Lévy-driven stochastic differential equations. Preprint arXiv:1804.04444, 2018.
[55] A. Jasra and F. Yu. Central limit theorems for coupled particle filters. Preprint arXiv:1810.04900, 2018.
[56] P. Kloeden and E. Platen. *Numerical solution of stochastic differential equations*. Springer, 3rd edition, 1999.
[57] A. Lee, S. Singh, and M. Vihola. Coupled conditional backward sampling particle filter. (arXiv:1806.05852), 2018.
[58] A. Lee, C. Yau, M. B. Giles, A. Doucet, and C. C. Holmes. On the utility of graphics cards to perform massively parallel simulation of advanced Monte Carlo methods. *J. Comput. Graph. Statist.*, 19(4):769–789, 2010.
[59] L. Lin, K. Liu, and J. Sloan. A noisy Monte Carlo algorithm. *Phys. Rev. D*, 61, 2000.
[60] F. Lindsten, J. Helske, and M. Vihola. Graphical model inference: Sequential Monte Carlo meets deterministic approximations. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 8190–8200. Curran Associates, Inc., 2018.
[61] J. S. Liu. *Monte Carlo strategies in scientific computing*. Springer-Verlag, New York, 2003.
[62] P. Marjoram, J. Molitor, V. Plagnol, and S. Tavaré. Markov chain Monte Carlo without likelihoods. *Proc. Natl. Acad. Sci. USA*, 100(26):15324–15328,

2003.

[63] P. Mathé and E. Novak. Simple Monte Carlo and the Metropolis algorithm. *J. Complexity*, 23(4-6):673–696, 2007.

[64] D. McLeish. A general method for debiasing a Monte Carlo estimator. *Monte Carlo Methods Appl.*, 17(4):301–315, 2011.

[65] N. Metropolis, A. Rosenbluth, M. Rosenbluth, and A. Teller. Equation of state calculations by fast computing machines. *Chem. Phys.*, 21(6), 1953.

[66] S. Meyn and R. L. Tweedie. *Markov chains and stochastic stability*. Cambridge University Press, second edition, 2009.

[67] A. Mira and C. Geyer. Ordering Monte Carlo Markov Chains. Technical report, School of Statistics, University of Minnesota, 1999.

[68] A. Mira and F. Leisen. Covariance ordering for discrete and continuous time Markov chains. *Statistica Sinica*, pages 651–666, 2009.

[69] J. Møller, A. Pettitt, R. Reeves, and K. Berthelsen. An efficient Markov chain Monte Carlo method for distributions with intractable normalising constants. *Biometrika*, 93(2):451–458, 2006.

[70] J. K. Møller and H. Madsen. From state dependent diffusion to constant diffusion in stochastic differential equations by the Lamperti transform. *IMM-Technical Report-2010-16*, 2010.

[71] R. M. Neal. Annealed importance sampling. *Statist. Comput.*, 11(2):125–139, 2001.

[72] A. Owen. Statistically efficient thinning of a Markov chain sampler. *J. Comput. Graph. Statist.*, (just-accepted), 2017.

[73] P. Parpas, B. Ustun, M. Webster, and Q. K. Tran. Importance sampling in stochastic programming: A Markov chain Monte Carlo approach. *INFORMS J. Comput.*, 27(2):358–377, 2015.

[74] P. H. Peskun. Optimum Monte-Carlo sampling using Markov chains. *Biometrika*, 60(3):607–612, 1973.

[75] M. Pitt and N. Shephard. Filtering via simulation: auxiliary particle filters. *J. Amer. Statist. Assoc.*, 94(446):590–599, 1999.

[76] D. Prangle. Summary statistics. In S. Sisson, Y. Fan, and M. Beaumont, editors, *Handbook of Markov chain Monte Carlo*, pages 125–152. Chapman & Hall/CRC, 2018.

[77] C.-H. Rhee and P. W. Glynn. Unbiased estimation with square root convergence for SDE models. *Oper. Res.*, 63(5):1026–1043, 2015.

[78] G. Roberts and R. Tweedie. Geometric convergence and central limit theorems for multidimensional Hastings and Metropolis algorithms. *Biometrika*, 83(1):95–110, 1996.

[79] V. Roy, A. Tan, and J. M. Flegal. Estimating standard errors for importance sampling estimators with multiple Markov chains. Preprint arXiv:1509.06310, 2015.

[80] W. Rudin. *Principles of mathematical analysis*. McGraw-Hill, 3rd edition, 1976.

[81] D. Rudolf. Explicit error bounds for Markov chain Monte Carlo. Preprint arXiv:1108.3201, 2011.

[82] D. Rudolf and B. Sprungk. On a Metropolis-Hastings importance sampling estimator. Preprint arXiv:1805.07174, 2018.

[83] H. Rue, S. Martino, and N. Chopin. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *J.*

R. Stat. Soc. Ser. B Stat. Methodol., 71(2):319–392, 2009. (with discussion).
[84] S. Särkkä. *Bayesian filtering and smoothing*. Cambridge University Press, 2013.
[85] I. Schuster and I. Klebanov. Markov chain importance sampling - a highly efficient estimator for MCMC. Preprint arXiv:1805.07179, 2018.
[86] T. Schwedes and B. Calderhead. Quasi Markov chain Monte Carlo methods. Preprint arXiv:1807.00070, 2018.
[87] G. Sermaidis, O. Papaspiliopoulos, G. Roberts, A. Beskos, and P. Fearnhead. Markov chain Monte Carlo for exact inference for diffusions. *Scand. J. Statist*, 40(2):294–321, 2013.
[88] C. Sherlock and A. Lee. Variance bounding of delayed-acceptance kernels. Preprint arXiv:1706.02142, 2017.
[89] C. Sherlock, A. H. Thiery, G. O. Roberts, and J. S. Rosenthal. On the efficiency of pseudo-marginal random walk Metropolis algorithms. *Ann. Statist.*, 43(1):238–275, 2015.
[90] A. N. Shiryaev. *Probability*. Springer-Verlag, New York, second edition, 1996.
[91] S. Sisson and Y. Fan. ABC samplers. In S. Sisson, Y. Fan, and M. Beaumont, editors, *Handbook of Markov chain Monte Carlo*. Chapman & Hall/CRC, 2018.
[92] S. A. Sisson, Y. Fan, and M. Beaumont. *Handbook of approximate Bayesian computation*. Chapman & Hall/CRC, 2018.
[93] A. Solonen, P. Ollinaho, M. Laine, H. Haario, J. Tamminen, and H. Järvinen. Efficient MCMC for climate model parameter estimation: parallel adaptive chains and early rejection. *Bayesian Analysis*, 7(3):715–736, 2012.
[94] L. Stewart and P. McCarty. Use of Bayesian belief networks to fuse continuous and discrete information for target recognition, tracking, and situation assessment. In *Signal Processing, Sensor Fusion, and Target Recognition*, volume 1699, pages 177–186, 1992.
[95] L. Tierney. A note on Metropolis-Hastings kernels for general state spaces. *Ann. Appl. Probab.*, 8(1):1–9, 1998.
[96] M. Vihola. Unbiased estimators and multilevel Monte Carlo. *Oper. Res.*, 66(2):448–462, 2018.
[97] Q. Wang, V. Rao, and Y. W. Teh. An exact auxiliary variable Gibbs sampler for a class of diffusions. Preprint arXiv:1903.10659, 2019.
[98] D. Wegmann, C. Leuenberger, and L. Excoffier. Efficient approximate Bayesian computation coupled with Markov chain Monte Carlo without likelihoods. *Genetics*, 182(4):1207–1218, 2009.
[99] Z. Zheng, J. Blanchet, and P. Glynn. Rates of convergence and CLTs for subcanonical debiased MLMC. In A. Owen and P. Glynn, editors, *Monte Carlo and quasi-Monte Carlo methods*. Springer Proc. Math. Stat., 2018.
[100] K. Łatuszyński, B. Miasojedow, and W. Niemiro. Nonasymptotic bounds on the estimation error of MCMC algorithms. *Bernoulli*, 19(5A):2033–2066, 11 2013.

Article [A]

Importance sampling type estimators based on approximate marginal MCMC

Matti Vihola, Jouni Helske and Jordan Franks

Preprint arXiv:1609.02541v5, 2016.

IMPORTANCE SAMPLING TYPE ESTIMATORS BASED ON APPROXIMATE MARGINAL MCMC

MATTI VIHOLA, JOUNI HELSKE, AND JORDAN FRANKS

Abstract. 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 [44, 64, 102].

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.

2010 *Mathematics Subject Classification.* Primary 65C60; secondary 60J22, 65C05, 65C35, 65C40.

Key words and phrases. Delayed acceptance, exact approximation, importance sampling, Markov chain Monte Carlo, parallel computing, particle filter, state space model, unbiased estimator.

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. 20, 38, 46], and used, for instance, to estimate Bayes factors using a single MCMC output [21]. Related confidence intervals have been suggested based on regeneration [11] and in case of multiple Markov chains [94]. Using unbiased estimators of importance weights in this context has been suggested at least in [65, 68], who consider marginal inference with a generalisation of the pseudo-marginal method, allowing for likelihood estimators that may take negative values, and in [82] with data sub-sampling.

Nested or compound sampling has also appeared in many forms in the Monte Carlo literature. The SMC^2 algorithm [13] is based on an application of nested sequential Monte Carlo steps, which has similarities with our framework, and the IS^2 method [96] 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) [15, 33, 67], and recently developed pseudo-marginal type MCMC [4, 65] such as particle MCMC [2], grouped independence Metropolis-Hastings [9], approximate Bayesian computation (ABC) MCMC [69], the algorithm for estimation of discretely observed diffusions suggested in [10], and annealed IS [57, 74]. Theoretical advances of pseudo-marginal methods [3, 6, 7, 14, 27, 66, 92] 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 [4, 62]. Adding dependence to the estimators [cf. 7], such as using the recently proposed correlated version of the pseudo-marginal MCMC [18], 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 [93] 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; see Section 2.3.

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 3), based on a concept (Definition 2), which we call a 'proper weighting' scheme (following the terminology of Liu [67]), 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 detail proper weighting schemes in the state space models (SSMs) using sequential Monte Carlo (SMC) in Section 8. We then focus on SSMs with linear-Gaussian state dynamics in Section 9, and show how a Laplace approximation can be used both for approximate inference, and for construction of efficient proper weighting schemes. Section 10 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 11 with Poisson observations, with a stochastic volatility model and with a discretely observed geometric Brownian motion. Section 12 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) [87] and the particle MCMC methods (PMCMC) [2], 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.

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 $(\Theta_k)_{k=1,\ldots,n}$ targeting an approximate hyperparameter posterior $\pi_a(\theta) \propto \mathrm{pr}(\theta)L_a(\theta)$.

Phase 2: For each Θ_k, sample $(V_k^{(i)}, X_k^{(i)})_{i=1,\dots,m}$ 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 essential conditions required for the validity of the estimator are:

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. Section 8 reviews how (unnormalised) importance sampling and particle filter lead to such schemes. There is also a (mild) integrability condition, which $(W_k^{(i)}, X_k^{(i)})$ must satisfy in order to guarantee a strong convergence $E_n(f) \to \mathbb{E}_\pi[f(\Theta, X)]$. When $V_k^{(i)} \geq 0$ almost surely, it suffices that $|f|$ 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 $(\tilde{V}_k^{(i)}, \tilde{X}_k^{(i)})_{i=1,\dots,m}$ 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. 4], 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

(i) 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 $\mathbb{E}[\tilde{U}_k \mid \tilde{\Theta}_k = \theta] = L_a(\theta)$.

(ii) 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 $(V_k^{(i)}, X_k^{(i)})_{i=1,\dots,m}$ 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 (Θ_k, U_k), 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 15 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 12).

Note that the variables $(V_k^{(i)}, X_k^{(i)})$ 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 [18], which relies on a particular form of $V_k^{(i)}$ and U_k. If positively correlated structure is unavailable, $(V_k^{(i)}, X_k^{(i)})$ may be constructed independent 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 $(W_k^{(i)}, X_k^{(i)})$ 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. 42]. Such an algorithm defines a Markov chain $(\Theta_k, W_k^{(i)}, X_k^{(i)})_{k\geq 1}$, 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, otherwise go to (DA 2).

DA 2: Conditional on $\tilde{\Theta}_k$, draw $(\tilde{V}_k^{(i)}, \tilde{X}_k^{(i)})$ which satisfies (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 $(\tilde{\Theta}_k, \tilde{W}_k^{(i)}, \tilde{X}_k^{(i)})$, otherwise reject.

If the pseudo-marginal method is used in DA 1, the value $L_a(\Theta_k)$ is replaced with the related likelihood estimator. Under essentially the same assumptions as required by our scheme, and additionally requiring that $\tilde{W}_k^{(i)} \geq 0$, the DA scheme described above 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.

(i) Phase 2 corrections are entirely independent 'post-processing' of Phase 1 MCMC output $(\Theta_k)_{k=1,\ldots,n}$, which is easy to implement efficiently using parallel or distributed computing. This is unlike DA 1 and DA 2, which must be iterated sequentially.

(ii) 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.

(iii) 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. 78].

(iv) 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. 6]. It is also easier to implement efficiently. For instance, popular adaptive MCMC methods which rely on acceptance rate optimisation [5, and references therein] are directly applicable.

(v) Reversibility of the MCMC kernel P in DA 1 is necessary, but not required for the Phase 1 MCMC.

(vi) 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 [37, 47] or debiasing [71, 84, 98] are applicable. (See also the discussion in [52] why pseudo-marginal method may not be applicable at all in such a context.)

(vii) 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.

The further work [35] considers the relationship between IS and DA in terms of the asymptotic variance.

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 '$\mathrm{d}x$' on a σ-algebra denoted with a corresponding calligraphic letter, such as $\mathcal{X}$. Product spaces 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

$\nu(\mathrm{d}x) := \nu(x)\mathrm{d}x$.[1] 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 $\mathbb{N} := \{1, 2, \ldots\}$. 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 $x_{a:b} = (x_a, \ldots, x_b)$, $x^{(a:b)} = (x^{(a)}, \ldots, x^{(b)})$. 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 $x^{(i_{1:T})} = (x^{(i_1)}, \ldots, x^{(i_T)})$ and $x_{1:T}^{(i_{1:T})} = (x_1^{(i_1)}, \ldots, x_T^{(i_T)})$. We also use double-indexing, such as $x_k^{(1:m,1:n)} = (x_k^{(1,1)}, \ldots, x_k^{(1,m)}, x_k^{(2,1)}, \ldots, x_k^{(m,n)})$.

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. 75, 95], although in some cases careless implementation could lead to a non-Harris chain [cf. 85]. 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 1. 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 1 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. 20, 38]

$$\frac{\sum_{k=1}^n w_u(\Theta_k)g(\Theta_k)}{\sum_{j=1}^n w_u(\Theta_j)} = \frac{n^{-1}\sum_{k=1}^n w_u(\Theta_k)g(\Theta_k)}{n^{-1}\sum_{j=1}^n w_u(\Theta_j)} \xrightarrow[\text{a.s.}]{n\to\infty} \frac{\pi_a(w_u g)}{\pi_a(w_u)} = \pi_m(g),$$

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 to estimate the numerator and the denominator, in place of $w_u(\Theta_k)$ and $g(\Theta_k)$, under

[1] Note that our definition is set-theoretic support of the density, and differs in general from the support of the measure ν (on a topological space X).

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 2. Suppose Assumption 1 holds, and let $(S_k)_{k\geq 1}$, where $S_k = \big(A_k, B_k\big) \in \mathbb{R}^2$, 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 $m^{(1)}(\theta) := \mathbb{E}\big[|A_k| + |B_k| \;\big|\; \Theta_k = \theta\big]$.

We record the following simple statement which guarantees consistency under Assumption 2.

Lemma 1. *If Assumption 2 holds for some $g \in L^1(\pi_m)$, then*

$$E_n(g) := \frac{\sum_{k=1}^n A_k}{\sum_{j=1}^n B_j} \xrightarrow[\text{a.s.}]{n\to\infty} \pi_m(g).$$

The proof of Lemma 1 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 $\check{\pi}(\mathrm{d}\theta \times \mathrm{d}s) := \pi_a(\mathrm{d}\theta) Q(\theta, \mathrm{d}s)$, where $Q(\theta, A) := \mathbb{P}(S_k \in A \mid \Theta_k = \theta)$; see Lemma 24 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 2 with $g = f^*$ and therefore guarantees consistency for a class of functions $f \in \mathcal{L} \subset L^1(\pi)$.

Definition 2 ($\mathcal{L}$-Proper weighting scheme)**.** Suppose Assumption 1 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 $m_f^{(1)}(\theta) := \mathbb{E}\big[|\xi_k(f)| \;\big|\; \Theta_k = \theta\big]$.

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 2. Regarding Definition 2:

(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). Further, if (i) holds for both f and $|f|$, then (ii) holds, because $|\xi_k(f)| \leq \xi_k(|f|)$.
(ii) When certain multilevel [37, 47] or debiasing methods [cf. 39, 71, 84] 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.

(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*. Our results extend also to such generalisation, which may be a useful interpretation for instance in the context of Rao-Blackwellisation, where ξ_k could be mixtures of Gaussians.

The following consistency result is a direct consequence of Lemma 1.

Theorem 3. *If $(\xi_k)_{k\geq 1}$ form a $\mathcal{L}$-proper weighting scheme, then the IS type estimator is consistent, that is,*

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

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

Proposition 4. *Suppose Assumption 1 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*

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

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, we record the following simple observations how a proper weighting property is inherited in sub-sampling, which may be useful for instance due to memory constraints.

Proposition 5. *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 recording a complementary statement about convex combinations, allowing to merge multiple proper sampling schemes.

Proposition 6. *Suppose $(\xi_{k,j})_{k\geq 1}$ forms a $\mathcal{L}$-proper weighting scheme for each $j \in \{1{:}N\}$, then, for any constants $\beta_1, \ldots, \beta_N \geq 0$ 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 chains, which coincides with the limiting variance of related estimators in case a central limit theorem (CLT) holds.

Definition 3. 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

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

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 7. *Suppose that the conditions of Theorem 3 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) $\sum_{n=1}^{\infty} n^{-3/2} \big\{ \pi_m\big(\big[\sum_{k=0}^{n-1} P^k(\mu_{\bar{f}}) \big]^2 \big) \big\}^{1/2} < \infty$,

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

$$\sqrt{n}[E_n(f) - \pi(f)] \xrightarrow{n\to\infty} N\big(0, \sigma_f^2\big), \quad \text{where} \quad \sigma_f^2 := \frac{\mathrm{Var}(\mu_{\bar{f}}, P) + \pi_a(v)}{c_w^2}$$

in distribution, where $v(\theta) := \mathrm{Var}\big(\xi_k(\bar{f}) \,\big|\, \Theta_k = \theta\big)$.

Remark 8. In case of reversible chains, the condition in Theorem 7 (i) is essentially optimal, and the CLT relies on a result due to Kipnis and Varadhan [58]. 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 [53, 86]. In case $(\Theta_k)_{k\geq 1}$ is sub-geometric, such as polynomial, extra conditions are required; see for instance [54]. The condition (ii) which applies for non-reversible chains is also nearly optimal, and relies on a result due to Maxwell and Woodroofe [70]. See also the review on Markov chain CLTs by Jones [55].

Note that the latter term $\pi_a(v)$ in the asymptotic variance expression 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 9. *Suppose $f \in \mathcal{L} \cap L^2(\pi)$ and $\pi_a(v) < \infty$ where v is defined in Theorem 7, and also $\pi_a(m_{\mathbf{1}}^{(2)}) < \infty$. Then, the estimator*

$$v_n := \frac{\sum_{k=1}^{n} \big(\xi_k(f) - \xi_k(\mathbf{1}) E_n(f)\big)^2}{\big(\sum_{j=1}^{n} \xi_j(\mathbf{1})\big)^2}$$

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

Proof of Theorem 9 is given in Appendix B.

The estimator nv_n in Theorem 9 provides a consistent estimate for the CLT variance $\sigma_{\bar{f}}^2/n$ when P corresponds to i.i.d. sampling, in which case $\mathrm{Var}(\mu_{\bar{f}}, P) = \pi_a(\mu_{\bar{f}}^2)$. Typically, $\mathrm{Var}(\mu_{\bar{f}}, P) \geq \pi_a(\mu_{\bar{f}}^2)$ (which is always true when P is positive), and then nv_n provides a lower bound of the variance. It can provide useful information about the importance sampling noise contribution, and may be used as an optimisation criteria when adjusting the accuracy of the related estimators. Generic Markov chain asymptotic variance estimators (see, e.g., the review [32] and references therein) may also be used with IS correction, by estimating the asymptotic variance of $n^{-1}\sum_{k=1}^n \xi_k(f)$ and dividing it by $[n^{-1}\sum_{k=1}^n \xi_k(\mathbf{1})]^2$.

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. 19, 23, 27].

Definition 4 (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:

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

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

Remark 10. 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 $(\tilde{\Theta}_k)$.

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 3. Suppose that Assumption 1 holds, and let $(\tilde{\Theta}_k, N_k)_{k\geq 1}$ denote the corresponding jump chain (Definition 4). 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 2 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 $\bar{m}^{(1)}(\theta) := \sup_{n\in\mathbb{N}} \mathbb{E}\big[|\xi_k(f)| \,\big|\, \Theta_k = \theta, N_k = n\big]$.

Theorem 11. *Suppose Assumption 3 holds, then,*

$$E_n(f) := \frac{\sum_{k=1}^n N_k\xi_k(f)}{\sum_{j=1}^n N_j\xi_j(\mathbf{1})} \xrightarrow[\text{a.s.}]{n\to\infty} \pi(f). \tag{4}$$

The proof follows from Lemma 1 because $(\tilde{\Theta}_k)$ is Harris ergodic with invariant probability $\tilde{\pi}_a(\theta) \propto \pi_a(\theta)\alpha(\theta)$; see Proposition 27 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 27), and therefore $\mathbb{E}[N_k \mid \tilde{\Theta}_k = \theta] = 1/\alpha(\theta)$.

Remark 12. Regarding Assumption 3:

(i) Condition (ii) in Assumption 3 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 $(\xi_{k,1}, \ldots, \xi_{k,N_k})$ (cf. Proposition 6), 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. Section 8) 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 (4) 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 in [23].

Let us finally consider a central limit theorem corresponding to the estimator in Theorem 11, whose proof is given in Appendix C.

Theorem 13. *Suppose Assumption 3 holds,* $(\tilde{\Theta}_k)_{k\geq 1}$ *is aperiodic,* $f \in \mathcal{L} \cap L^2(\pi)$,

$$\pi_a\big(\alpha \tilde{m}^{(2)}\big) < \infty, \qquad \text{where} \qquad \tilde{m}^{(2)}(\theta) := \mathbb{E}\big[N_k^2 |\xi_k(\bar{f})|^2 \;\big|\; \tilde{\Theta}_k = \theta\big], \tag{5}$$

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* $g - Pg = \mu_{\bar{f}}$.

Then, the estimator $E_n(f)$ *in* (4) *satisfies*

$$\sqrt{n}\big[E_n(f) - \pi(f)\big] \xrightarrow{n\to\infty} N(0, \sigma^2) \qquad \text{in distribution,}$$

where the limiting variance can be given as:

$$\sigma^2 = \frac{\pi_a(\alpha)}{c_w^2}\Big[\mathrm{Var}\big(\mu_{\bar{f}}, P\big) + \pi_a(\alpha \tilde{v})\Big], \tag{6}$$

where $\tilde{v}(\theta) := \mathbb{E}\big[N_k^2 \mathrm{Var}\big(\xi_k(\bar{f}) \;\big|\; \tilde{\Theta}_k = \theta, N_k\big) \;\big|\; \tilde{\Theta}_k = \theta\big]$.

Let us briefly discuss the conditions and implications of Theorem 13 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

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

and so $\pi_a(b) < \infty$ guarantees (5). For example, if $(\Theta_k)_{k\geq 1}$ is L^2-geometrically ergodic, then the acceptance probability is (essentially) bounded away from zero [86], and $g := \sum_{k\geq 0} P^k \mu_{\bar{f}} \in L^2(\pi_a)$ 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 6) which do not depend on N_k,

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

Then, $\pi_a(\alpha\tilde{v}) = \pi_a(\hat{v})$, which leads to an asymptotic variance that coincides with simple IS correction (cf. Theorem 7).

Remark 14. In the non-reversible case, our CLT only applies when a solution $g \in L^2(\pi_a)$ to the Poisson equation $g - Pg = \mu_{\bar{f}}$ exists. We believe that the result holds more generally, but this requires showing that the jump chain $(\tilde{\Theta}_k)_{k\geq 1}$ inherits a central limit theorem from the base chain $(\Theta_k)_{k\geq 1}$ under more general conditions.

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 formalise next a pseudo-marginal Markov chain $(\Theta_k, \Phi_k)_{k\geq 1}$ on $\mathsf{T} \times \mathsf{S}_\Phi$. Let $\Theta_0 \in \mathsf{T}$ and $\Phi_0 \in \mathsf{S}_\Phi$ such that $U(\Phi_0) > 0$, and for $k \geq 1$, iterate

(i) Generate $\tilde{\Theta}_k \sim q(\Theta_{k-1}, \cdot\,)$ and $\tilde{\Phi}_k \sim Q_a(\tilde{\Theta}_k, \cdot\,)$.

(ii) 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 Markov chain $(\Theta_k, \Phi_k)_{k\geq 1}$ 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. 6]:

Assumption 4. 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, $(\Theta_k, \Phi_k)_{k\geq 1}$ is easily shown to be Harris ergodic under minimal conditions.

Let us consider next an abstract minimal condition which ensures consistency of an IS type estimator. We discuss practically relevant sufficient conditions later in Proposition 17.

Assumption 5. Suppose Assumption 1 holds, $(\Theta_k, \Phi_k)_{k\geq 1}$ is Harris ergodic, $c_m > 0$ is a constant, and let $(P_k)_{k\geq 1}$ be conditionally independent given $(\Theta_k, \Phi_k)_{k\geq 1}$, such that the distribution of each $P_k = (M_k, V_k^{(1:M_k)}, X_k^{(1:M_k)})$ depends only on (Θ_k, Φ_k), 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) $\iint Q_a(\theta, \mathrm{d}\phi)\mathbb{I}\,(U(\phi) > 0)\,\mathbb{E}\big[|\zeta_k(f)| \;\big|\; \Theta_k = \theta, \Phi_k = \phi\big]\mathrm{d}\theta < \infty$.

Proposition 15. *Suppose Assumption 4 and 5 hold, and* $\mathbf{1} \in \mathcal{L}$. *Then, Theorem 3 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 15 follows by noting a proper weighting scheme involving the augmented approximate marginal distribution π_a° and target distribution π° (Lemma 16), and Theorem 3.

Lemma 16. *Suppose the conditions of Proposition 15 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 2, 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

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

by Assumption 5 (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 5 (ii). □

Let us finally consider different conditions, which guarantee Assumption 5 (i); the integrability Assumption 5 (ii) may be shown similarly.

Proposition 17. *Assumption 5* (i) *holds if one of the following hold:*

(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

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

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

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

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

Remark 18. Proposition 17 (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 17 (ii) may be used directly to verify the validity of an MCMC version of the lazy ABC algorithm [81]. 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 17 (iii) demonstrates that strict positivity is not necessary, but in this case a delicate dependency structure is required.

8. General state space models and sequential Monte Carlo

State space models (SSM) are latent variable models which are commonly used in time series analysis [cf. 12]. 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 the observations $Y_{1:T}$ are conditionally independent with $Y_i \sim g_t^{(\theta)}(\cdot \mid Z_t)$.

This section reviews general techniques to generate random variables $V_\theta^{(1:m)}$ and $X_\theta^{(1:m)}$ for which $\zeta_\theta(h) := \sum_{i=1}^m V_\theta^{(i)} h(X^{(i)})$ satisfy:

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

for any θ and for some class of functions $h : \mathsf{S}_z^T \to \mathbb{R}$. These random variables may be used in order to construct a proper weighting; see Corollary 21 below.

Simple IS correction may be applied directly (see Proposition 4). 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 6, allowing for instance variance reduction by using pairs of antithetic variables [29].

For the rest of the section, we focus on the particle filter (PF) algorithm [43]; see also the monographs [12, 16, 24]. We consider a generic version of the algorithm, with the following components [cf. 16]:

(i) 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 $z_{1:t-1} \in \mathsf{S}_z^{t-1}$.
(ii) Potential functions: $G_t : \mathsf{S}_z^t \to \mathbb{R}_+$.
(iii) 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 two conditions are minimal for consistency:

Assumption 6. 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 6 (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 9, inspired by the 'twisted SSM' approach of [45]. It allows a 'look-ahead' strategy based on approximations of the full smoothing distributions $q^{(\theta)}(z_{1:T} \mid y_{1:T})$. Assumption 6 (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. 22].

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

Algorithm 1 (Particle filter)**.** Initial state:

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

(ii) Sample $A_{t-1}^{(1:m)} \sim \mathrm{Res}(\cdot \mid \bar{\omega}_{t-1}^{(1:m)})$.
(iii) Sample $Z_t^{(i)} \sim M_t(\cdot \mid \bar{Z}_{t-1}^{(A_{t-1}^{(i)})})$ and set $\bar{Z}_t^{(i)} = (\bar{Z}_{t-1}^{(A_{t-1}^{(i)})}, Z_t^{(i)})$.
(iv) 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 19. If $\omega_t^* = 0$, then Algorithm 1 is terminated immediately, and all the estimators considered (cf. Proposition 20) equal zero.

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 scattered in the literature [e.g. 16, 79], and some may be stated under slightly more stringent conditions, but a self-contained and concise proof of Proposition 20 may be found in [99].

Proposition 20. *Let $\theta \in \mathsf{T}$ be fixed, assume* Res*, M_t and G_t satisfy Assumption 6, 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 1, 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 (iiic).

The estimator in Proposition 20 (i) was called the filter-smoother in [59]. This property was shown in [16, Theorem 7.4.2] in case of multinomial resampling, and extended later [cf. 2]. The statement holds also when the PF is applied with a general sequence of distributions rather than the SSM [16]. Proposition 20 (ii) corresponds to backwards simulation smoothing [41]. Drawing a single backward trajectory is, perhaps surprisingly, probabilistically equivalent to subsampling one trajectory from the filter-smoother

estimate in Proposition 20 (i) [26]. However, drawing several trajectories independently as in Proposition 20 (ii) may lead to lower variance estimators. Proposition 20 (iii) and its special case (iv) correspond to the forward-backward smoother [25]; see also [12]. 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 its square complexity in m makes it inefficient with large m.

We next formally state how Proposition 20 allows to use Algorithm 1 to derive a proper weighting scheme.

Corollary 21. *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 1 with $\theta = \Theta_k$, as defined in Proposition 20* (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 2), for the following classes of functions, respectively:*

(i) $\mathcal{L} = L^1(\pi)$, (iii) $\mathcal{L} = \{f \in L^1(\pi) \,:\, f(\theta, x_{1:T}) = \hat{f}(\theta, x_{t-1:t})\}$,

(ii) $\mathcal{L} = L^1(\pi)$, (iv) $\mathcal{L} = \{f \in L^1(\pi) \,:\, f(\theta, x_{1:T}) = \hat{f}(\theta, x_t)\}$.

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

Remark 22. The latter two cases in Corollary 21 are stated for a single marginal (pair), but it is clear that we may calculate estimates simultaneously for several marginal (pairs), so that Proposition 20 (iii) is applicable for every function which is of the form $\sum_{t=1}^{T-1} f_t(\theta, x_{t:t+1})$ and Proposition 20 (iv) for a function of the form $\sum_{t=1}^{T} f_t(\theta, x_t)$. See also the general discussion of smoothing functionals in [12, §4.1.2].

We state finally an implication of Proposition 20 outside the main focus of this paper, in general SSM smoothing context (with fixed θ). This result is widely known among particle filtering experts, but appears not to be widely adopted.

Proposition 23. *Suppose $\theta \in \mathsf{T}$ is fixed, and let $(V_k^{(1:m)}, X_k^{(1:m)})_{k\geq 1}$ correspond to independent realisations of random variables defined in Proposition 20.*

(i) If the conditions of Proposition 20 are satisfied, then the estimator

$$E_n(h) := \frac{\sum_{k=1}^n \zeta_k(h)}{\sum_{j=1}^n \zeta_j(\mathbf{1})} \xrightarrow[\text{a.s.}]{n\to\infty} \mu_h := \int p^{(\theta)}(x_{1:T} \mid y_{1:T}) h(x_{1:T}) \mathrm{d}x_{1:T}.$$

(ii) If also $\sigma_*^2 := \mathbb{E}\big[|\zeta_1(\bar{h})|^2\big] < \infty$, where $\bar{h} = h - \mu_h$, then

$$\sqrt{n}\big[E_n(h) - \mu_h\big] \xrightarrow[\mathrm{d}]{n\to\infty} N(0, \sigma^2), \qquad \text{where} \qquad \sigma^2 := \frac{\sigma_*^2}{p^{(\theta)}(y_{1:T})^2}.$$

(iii) If in addition $\mathbb{E}\big[|\zeta_1(\mathbf{1})|^2\big] < \infty$, then $nv_n \to \sigma^2$, almost surely, where

$$v_n := \frac{\sum_{k=1}^n \big(\zeta_k(h) - \zeta_k(\mathbf{1})E_n(h)\big)^2}{\big(\sum_{j=1}^n \zeta_k(\mathbf{1})\big)^2}.$$

Proof is similar to Theorem 9 in Appendix B.

The estimator $E_n(h)$ in Proposition 23 is an importance sampling analogue of the particle independent Metropolis-Hastings (PIMH) algorithm suggested in [2]. Unlike

the PIMH, calculation of $E_n(h)$ is parallelisable, and allows for straightforward consistent confidence intervals $\left[E_n(f) \pm \beta\sqrt{v_n}\right]$, where β corresponds to the desired standard Gaussian quantile. Calculation of consistent confidence intervals for a single realisation of a particle smoothing algorithm requires sophisticated techniques [63]. Another promising method recently suggested in [50] relies on unbiased estimators obtained by coupling of conditional sequential Monte Carlo and debiasing tricks as in [39, 71, 84].

9. State space models with linear-Gaussian state dynamics

We consider a special case of the general SSM in Section 8, 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.

9.1. **Marginal approximation.** The scheme we consider here is based on [28, 90], 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:

(i) $R_t^{(\theta)} = -[D_t^{(2)}(H_t^{(\theta)}\tilde{z}_t)]^{-1}$ and $\tilde{y}_t^{(\theta)} = H_t^{(\theta)}\tilde{z}_t + R_t^{(\theta)} D_t^{(1)}(H_t^{(\theta)}\tilde{z}_t)$

(ii) 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 typically take less than 10 iterations [29].

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 [31] 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.

9.2. **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. Such a scheme using the simulation smoother [30] antithetic variables, we call **SPDK**, following [90].

We consider also several variants of M_t and G_t in the particle filter discussed in Section 8. The bootstrap filter [43], 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 [45, 101], we consider also the choice

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

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 [45]. As they are often approximately so, we believe that this choice, which we call ψ**-APF** following [45], can provide substantial benefits over BSF.

10. 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 $X = (Z_1, \ldots, Z_T)$ are assumed to follow the law of $(\tilde{Z}_{t_1}, \ldots, \tilde{Z}_{t_T})$, 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. 60]. The denser the time-discretisation mesh used, the less bias introduced. However, 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 (Section 8) 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 17 (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. 89]. The correction phase

could be also based on exact sampling for diffusions [10], which allow for elimination of the discretisation bias entirely.

The recent work [34] details how unbiased inference is also possible with IS type correction, using randomised multilevel Monte Carlo.

11. Experiments

We did experiments for our generic framework with SSMs, using Laplace approximations (Section 9) and an approximation based on coarsely discretised diffusions (Section 10). 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 $\tilde{p}^{(\Theta_k)}(z_{1:T} \mid y_{1:T})$.
- **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 12 (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 [97], targeting the acceptance rate 0.234. In DA, the adaptation was based on the first stage acceptance probability.

All the experiments were conducted in `R` [83] using the `bssm` package which is available online [48]. The experiments were run on a Linux server with eight octa-core Intel Xeon E7-8837 2.67GHz processors with total 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 [40].

Further details and results of the experiments may be found in the preprint version of our article [99].

11.1. 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 5, 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 9.2. For BSF and ψ-APF, the filter-smoother estimates as in Proposition 20 (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 [27]. 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

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 676s.

		BSF			SPDK				ψ-APF			
	AI	DA	IS1	IS2	PM	DA	IS1	IS2	PM	DA	IS1	IS2
Time	54	281	600	166	78	61	71	53	115	78	83	62
σ_η	0.039	0.721	0.535	0.336	0.060	0.047	0.056	0.042	0.082	0.068	0.065	0.049
σ_ξ	0.042	0.676	0.537	0.278	0.064	0.052	0.059	0.044	0.091	0.068	0.069	0.051
u_1	0.561	0.911	0.609	0.406	0.063	0.055	0.057	0.042	0.097	0.071	0.076	0.053
u_{100}	1.211	1.049	0.623	0.441	0.072	0.059	0.067	0.052	0.106	0.075	0.074	0.060
Time	11	235	549	120	35	17	28	10	72	34	38	19
σ_η	0.012	0.596	0.476	0.218	0.025	0.013	0.022	0.008	0.052	0.028	0.030	0.015
σ_ξ	0.052	0.564	0.530	0.197	0.029	0.015	0.025	0.009	0.061	0.031	0.034	0.017
u_1	0.085	0.779	0.527	0.273	0.027	0.016	0.023	0.009	0.056	0.030	0.033	0.015
u_{100}	0.333	0.804	0.563	0.305	0.034	0.016	0.027	0.010	0.068	0.034	0.036	0.019

posterior mean of θ. We kept the same m for all methods, for comparability, even though in some cases optimal choice might differ [91].

11.1.1. *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, leading to $\delta \approx 0.1$, and $m = 200$ for BSF with $\delta \approx 1.2$. 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 0.104, and 0.234 for all others. All exact methods led to essentially the same overall mean estimate $(0.093, 0.016, -0.075, 2.618\text{–}2.619)$ 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 $(-0.064, 2.629)$ and $(-0.065, 2.631)$ with local and global approximation, respectively. IS2-BSF outperformed DA-BSF by about a factor of two 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.

11.1.2. *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

TABLE 2. IREs for SV model. Times are in hours. AI^{G} is with global approximation and IS2^8 is with 8 parallel cores. For DA-BSF, IREs are one and time 46.3h.

			BSF		SPDK				ψ-APF			
	AI	AI^{G}	IS2	IS2^8	PM	DA	IS1	IS2	PM	DA	IS1	IS2
Time	1.3	0.2	25.2	4.6	4.4	1.9	2.8	1.5	2.4	1.4	1.5	1.3
ϕ	0.083	0.062	0.304	0.050	1.015	0.696	0.684	0.483	0.021	0.024	0.009	0.017
σ_η	0.726	0.298	0.483	0.096	3.090	3.307	0.603	0.710	0.044	0.055	0.016	0.028
ν	0.008	0.747	0.287	0.042	1.208	2.544	0.228	0.404	0.026	0.027	0.010	0.020
Z_1	0.133	0.035	0.321	0.071	3.054	1.883	0.346	0.373	0.029	0.026	0.007	0.018
Z_{5473}	1.887	0.417	0.540	0.112	6.574	1.871	0.444	0.810	0.057	0.064	0.012	0.039

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-known exponential deterioration in T, unlike the particle filter which often scales much better in T [100]. In addition, it is known that for this particular model, the importance weights may have large variability [61, 80]. 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 $m = 300$ 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.

11.2. **Discretely observed Geometric Brownian motion.** Our last experiment was about a discretely observed diffusion as discussed in Section 10. 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 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.

11.3. **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 [27] 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

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.

Init.	Mean						IRE				
		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

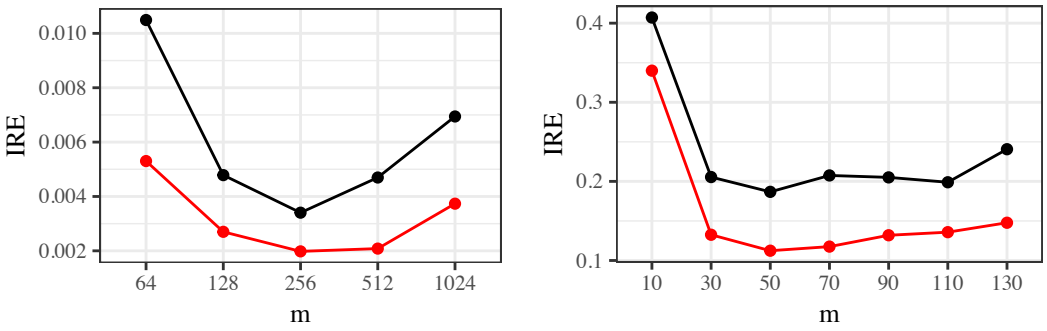


FIGURE 1. Average IRE of $(\sigma_\eta, \sigma_\xi, Z_1, Z_{100})$ in the Poisson model with BSF (left) and of $(\phi, \sigma_\eta, \nu, Z_1, Z_{5473})$ in the SV model with ψ-APF (right). DA is shown in black and IS2 in red.

less than DA uniformly in terms of IRE. In the Poisson-BSF case, the choice $m = 200$ based on [27] 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 [27]. 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.

12. 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 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)

[35]. Our followup work [35] 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. The IS weight may also be regularised easily 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 [49]. Other generic safe IS schemes may also be useful [cf. 77], 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. 5]. 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 [12], or quasi-Monte Carlo updates [36], apply directly. PFs can also be useful beyond the state space models context [17]. Twisted particle filters [1, 101] 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 [34]. We are currently investigating various other instances of our framework. Laplace approximations are available for a wider class of Gaussian latent variable models beyond SSMs [cf. 87]. Variational approximations [8, 56] and expectation propagation [73] have been found useful in a wide variety of models. In the SSM context, various non-linear filters could also be applied [cf. 88]. Our framework provides a generic validation mechanism for approximate inference, where assessment of bias is difficult in general [cf. 76]. 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 in [39] and further explored in [50, 51], may also lead to useful proper weighting schemes.

Acknowledgements

The authors have been supported by an Academy of Finland research fellowship (grants 274740, 284513 and 312605). We thank Christophe Andrieu, Arnaud Doucet, Anthony Lee and Chris Sherlock for many insightful remarks.

Appendix A. Properties of augmented Markov chains

Throughout this section, suppose that K is a Markov kernel on X and $Q(x, B)$ 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 24. *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*

- $\varphi_P \in \mathrm{irr}(K) \implies \varphi_{\check{P}}(\mathrm{d}x \times \mathrm{d}s) := \varphi_P(\mathrm{d}x)Q(x, \mathrm{d}s) \in \mathrm{irr}(\check{K})$,

- $\varphi_{\check{P}} \in \mathrm{irr}(\check{K}) \implies \varphi_P(\mathrm{d}x) := \varphi_{\check{P}}(\mathrm{d}x \times \mathsf{S}) \in \mathrm{irr}(K)$.

(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.

(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* $n \geq 1$*,*

$$(\check{K}^n h)(x, s) = (K^n m_h)(x).$$

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

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

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

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 25. *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 $\check{K}$.

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

by stationarity. For $\ell \geq 1$, Lemma 24 (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 $n \geq 1$

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

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

Lemma 26. *Suppose K is Harris ergodic and aperiodic, and $h \in L^2_0(\check{\nu})$. The CLT*

$$\frac{1}{\sqrt{n}} \sum_{k=1}^{n} h(X_k, S_k) \xrightarrow{n\to\infty} N\big(0, \mathrm{Var}(m_h, K) + \nu(v_h)\big) \tag{10}$$

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

(i) K is reversible and $\mathrm{Var}(m_h, K) < \infty$.
(ii) $\sum_{n=1}^{\infty} n^{-3/2} \big\{\nu\big(\big[\sum_{i=0}^{n-1} K^i m_h\big]^2\big)\big\}^{1/2} < \infty$.
(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 reversible case (i) follows from Lemma 25 and the Kipnis and Varadhan CLT [58], which implies (10) for the initial distribution $\check{\nu}$. The jump chain is Harris by Lemma 24 (iv), so [72, Proposition 17.1.6] guarantees (10) for every initial distribution.

The case (ii) follows similarly, but relies on a result due to Maxwell and Woodroofe [70], which guarantees (10) from $\check{\nu}$-almost every starting point, if $\sum_{n=1}^{\infty} n^{-3/2} \big\{\check{\nu}\big(\big[\sum_{i=0}^{n-1} \check{K}^i h\big]^2\big)\big\}^{1/2} < \infty$. Notice that for $n \geq 2$ by Lemma 24 (v),

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

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 [72, Theorem 17.4.4]. □

Appendix B. Proofs about CLT and asymptotic variance

Proof of Theorem 7. Whenever $\sum_{i=1}^{n} \xi_i(\mathbf{1}) > 0$, we may write

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

The denominator converges to $c_w > 0$ almost surely, som by Slutsky's lemma, it is enough to show that the numerator converges in distribution to $N\big(0, \mathrm{Var}(\nu_{\bar{f}}, P) + \pi_a(v)\big)$. This follows from Lemma 26 (i) and (ii), under conditions (i) and (ii), respectively. □

Proof of Theorem 9. For n large enough such that $\sum_{j=1}^{n} \xi_j(\mathbf{1}) > 0$, we may write

$$n v_n = \frac{\frac{1}{n} \sum_{k=1}^{n} \big(\xi_k(f) - \xi_k(\mathbf{1}) E_n(f)\big)^2}{\big(\frac{1}{n} \sum_{j=1}^{n} \xi_j(\mathbf{1})\big)^2}.$$

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

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

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 $D_n^2\big(n^{-1} \sum_{k=1}^{n} \xi_k^2(\mathbf{1})\big) \to 0$ and $2D_n\big(n^{-1} \sum_{k=1}^{n} \xi_k(\mathbf{1}) \xi_k(\bar{f})\big) \to 0$. □

Appendix 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 [23, Lemma 1] and [19], which are stated for more specific cases.

Proposition 27. *Suppose (X_k) is a Markov chain with kernel K and $(\tilde{X}_k)$ the corresponding jump chain with holding times (N_k) (Definition 4). 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 in [23]. The invariance (iii) follows from

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

and the reversibility is shown in [23]. 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

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

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

We now state results about the asymptotic variance of the jump chain, complementing the reversible case characterisation of [19, 27].

Proposition 28. *Let $f \in L^2_0(\tilde{\nu})$. With the notation of Proposition 27,*

(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

$$\mathrm{Var}(f, \tilde{K}) = \nu(a)^{-1}\big[\mathrm{Var}(af, K) - \nu\big(a(1-a)f^2\big)\big]. \tag{11}$$

(ii) If there exists a function $g \in L^2(\nu)$ which satisfies $g - Kg = af$, then $\mathrm{Var}(f, \tilde{K}) < \infty$, $\mathrm{Var}(af, K) < \infty$, (11) holds, $g - \tilde{K}g = f$ and $g \in L^2(\tilde{\nu})$.

Proof. The reversible case (i) is a restatement of [19, Theorem 1].

Consider then (ii). By Proposition 27 (i), we obtain for any $h : \mathsf{X} \to \mathbb{R}$ with Kh well-defined,

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

Consequently, we observe that $g - \tilde{K}g = a^{-1}\big(g - Kg\big) = f$ implying (ii). Because $g \in L^2(\tilde{\nu})$, Lemma 26 (iii) and a straightforward calculation yield

$$\begin{aligned}\mathrm{Var}(f, \tilde{K}) &= \tilde{\nu}\big(g^2 - (\tilde{K}g)^2\big) \\ &= 2\langle g, g - \tilde{K}g\rangle_{\tilde{\nu}} - \langle g - \tilde{K}g, g - \tilde{K}g\rangle_{\tilde{\nu}} \\ &= \nu(a)^{-1}\big[2\langle g, g - Kg\rangle_{\nu} - \nu(af^2)\big],\end{aligned}$$

where $\langle f, g\rangle_\nu := \int f(x)g(x)\nu(\mathrm{d}x)$. Similarly, by Lemma 26 (iii)

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

which allows us to conclude. □

Proof of Theorem 13. Whenever $\sum_{j=1}^n \xi_j(\mathbf{1}) > 0$, we may write

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

We shall show below that the CLT holds for the numerator, with asymptotic variance $\sigma^2 := \big[\mathrm{Var}(\mu_{\bar{f}}, P) + \pi_a(\alpha\tilde{v})\big]/\pi_a(\alpha)$. 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 (5).

In case (i) holds, also $\tilde{P}$ and $\check{P}$ are reversible by Proposition 27 (iii) and Lemma 24 (iii). Lemma 26 (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:

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

where, by the variance decomposition formula,

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

Proposition 28 (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 28 (ii) implies that $g - \tilde{P}g = \mu_{\bar{f}}/\alpha$, and $g \in L^2(\tilde{\pi})$. Lemma 26 (iii) implies the CLT, and together with Proposition 28 (ii) leads to $\mathrm{Var}(h, \check{P}) = \sigma^2$. □

References

[1] J. Ala-Luhtala, N. Whiteley, K. Heine, and R. Piché. An introduction to twisted particle filters and parameter estimation in non-linear state-space models. *IEEE Trans. Signal Process.*, 64(18):4875–4890, 2016.

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

[3] C. Andrieu, A. Lee, and M. Vihola. Uniform ergodicity of the iterated conditional SMC and geometric ergodicity of particle Gibbs samplers. *Bernoulli*, 24(2):842–872, 2018.

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

[5] C. Andrieu and J. Thoms. A tutorial on adaptive MCMC. *Statist. Comput.*, 18(4):343–373, Dec. 2008.

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

[7] C. Andrieu and M. Vihola. Establishing some order amongst exact approximations of MCMCs. *Ann. Appl. Probab.*, 26(5):2661–2696, 2016.

[8] M. J. Beal. *Variational algorithms for approximate Bayesian inference*. PhD thesis, University College London, 2003.

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

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

[11] S. Bhattacharya. Consistent estimation of the accuracy of importance sampling using regenerative simulation. *Statist. Probab. Lett.*, 78(15):2522–2527, 2008.

[12] O. Cappé, E. Moulines, and T. Rydén. *Inference in Hidden Markov Models*. Springer, 2005.

[13] N. Chopin, P. Jacob, and O. Papaspiliopoulos. SMC^2: A sequential Monte Carlo algorithm with particle Markov chain Monte Carlo updates. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 75(3):397–426, 2013.

[14] N. Chopin and S. S. Singh. On particle Gibbs sampling. *Bernoulli*, 21(3):1855–1883, 2015.

[15] J. A. Christen and C. Fox. Markov chain Monte Carlo using an approximation. *J. Comput. Graph. Statist.*, 14(4), 2005.

[16] P. Del Moral. *Feynman-Kac Formulae*. Springer, 2004.

[17] P. Del Moral, A. Doucet, and A. Jasra. Sequential Monte Carlo samplers. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 68(3):411–436, 2006.

[18] G. Deligiannidis, A. Doucet, M. K. Pitt, and R. Kohn. The correlated pseudo-marginal method. Preprint arXiv:1511.04992v3, 2015.

[19] G. Deligiannidis and A. Lee. Which ergodic averages have finite asymptotic variance? *Ann. Appl. Probab.*, 28(4):2309–2334, 2018.

[20] H. Doss. Discussion: Markov chains for exploring posterior distributions. *Ann. Statist.*, 22(4):1728–1734, 1994.

[21] H. Doss. Estimation of large families of Bayes factors from Markov chain output. *Statist. Sinica*, pages 537–560, 2010.

[22] R. Douc, O. Cappé, and E. Moulines. Comparison of resampling schemes for particle filtering. In *Proc. Image and Signal Processing and Analysis, 2005*, pages 64–69, 2005.

[23] R. Douc and C. P. Robert. A vanilla Rao-Blackwellization of Metropolis-Hastings algorithms. *Ann. Statist.*, 39(1):261–277, 2011.

[24] A. Doucet, N. de Freitas, and N. Gordon. *Sequential Monte Carlo Methods in Practice*. Springer-Verlag, New York, 2001.

[25] A. Doucet, S. Godsill, and C. Andrieu. On sequential Monte Carlo sampling methods for Bayesian filtering. *Statist. Comput.*, 10(3):197–208, 2000.

[26] A. Doucet and A. Lee. Sequential Monte Carlo methods. In M. Drton, S. Lauritzen, M. Matthuis, and M. Wainwright, editors, *Handbook of Graphical Models*. CRC press, to appear.

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

[28] J. Durbin and S. J. Koopman. Monte Carlo maximum likelihood estimation for non-Gaussian state space models. *Biometrika*, 84(3):669–684, 1997.
[29] J. Durbin and S. J. Koopman. 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, 2000.
[30] J. Durbin and S. J. Koopman. A simple and efficient simulation smoother for state space time series analysis. *Biometrika*, 89:603–615, 2002.
[31] J. Durbin and S. J. Koopman. *Time series analysis by state space methods*. Oxford University Press, New York, 2nd edition, 2012.
[32] J. M. Flegal and G. L. Jones. Batch means and spectral variance estimators in Markov chain Monte Carlo. *Ann. Statist.*, 38(2):1034–1070, 2010.
[33] C. Fox and G. Nicholls. Sampling conductivity images via MCMC. In K. V. Mardia, C. A. Gill, and R. G. Aykroyd, editors, *Proceedings in The Art and Science of Bayesian Image Analysis*, pages 91–100. Leeds University Press, 1997.
[34] J. Franks, A. Jasra, K. Law, and M. Vihola. Unbiased inference for discretely observed hidden markov model diffusions. Preprint arXiv:1807.10259, 2018.
[35] J. Franks and M. Vihola. Importance sampling and delayed acceptance via a Peskun type ordering. Preprint arXiv:1706.09873, 2017.
[36] M. Gerber and N. Chopin. Sequential quasi-Monte Carlo. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 77(3):509–579, 2015.
[37] M. B. Giles. Multilevel Monte Carlo path simulation. *Oper. Res.*, 56(3):607–617, 2008.
[38] P. W. Glynn and D. L. Iglehart. Importance sampling for stochastic simulations. *Management Science*, 35(11):1367–1392, 1989.
[39] P. W. Glynn and C.-H. Rhee. Exact estimation for Markov chain equilibrium expectations. *J. Appl. Probab.*, 51(A):377–389, 2014.
[40] P. W. Glynn and W. Whitt. The asymptotic efficiency of simulation estimators. *Oper. Res.*, 40(3):505–520, 1992.
[41] S. J. Godsill, A. Doucet, and M. West. Monte Carlo smoothing for nonlinear time series. *J. Amer. Statist. Assoc.*, 99(465):156–168, 2004.
[42] A. Golightly, D. A. Henderson, and C. Sherlock. Delayed acceptance particle MCMC for exact inference in stochastic kinetic models. *Statist. Comput.*, 25(5):1039–1055, 2015.
[43] N. J. Gordon, D. J. Salmond, and A. F. M. Smith. Novel approach to nonlinear/non-Gaussian Bayesian state estimation. *IEE Proceedings-F*, 140(2):107–113, 1993.
[44] P. J. Green, K. Łatuszyński, M. Pereyra, and C. P. Robert. Bayesian computation: a summary of the current state, and samples backwards and forwards. *Statist. Comput.*, 25(4):835–862, 2015.
[45] P. Guarniero, A. M. Johansen, and A. Lee. The iterated auxiliary particle filter. *J. Amer. Statist. Assoc.*, 112(520):1636–1647, 2017.
[46] W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109, Apr. 1970.
[47] S. Heinrich. Multilevel Monte Carlo methods. In *Large-scale scientific computing*, pages 58–67. Springer, 2001.
[48] J. Helske and M. Vihola. *bssm: Bayesian inference of non-linear and non-Gaussian state space models in R*, 2017. `https://CRAN.R-project.org/package=bssm`.
[49] T. Hesterberg. Weighted average importance sampling and defensive mixture distributions. *Technometrics*, 37(2):185–194, 1995.
[50] P. E. Jacob, F. Lindsten, and T. B. Schön. Smoothing with couplings of conditional particle filters. *J. Amer. Statist. Assoc.*, to appear. Preprint arXiv:1701.02002v1.
[51] P. E. Jacob, J. O'Leary, and Y. F. Atchadé. Unbiased Markov chain Monte Carlo with couplings. Preprint arXiv:1708.03625v1, 2017.
[52] P. E. Jacob and A. H. Thiery. On nonnegative unbiased estimators. *Ann. Statist.*, 43(2):769–784, 2015.
[53] S. F. Jarner and E. Hansen. Geometric ergodicity of Metropolis algorithms. *Stochastic Process. Appl.*, 85(2):341–361, 2000.
[54] S. F. Jarner and G. O. Roberts. Convergence of heavy-tailed Monte Carlo Markov chain algorithms. *Scand. J. Stat.*, 34(4):781–815, Dec. 2007.

[55] G. L. Jones. On the Markov chain central limit theorem. *Probab. Surv.*, 1:299–320, 2004.
[56] M. I. Jordan. Graphical models. *Statist. Sci.*, pages 140–155, 2004.
[57] G. Karagiannis and C. Andrieu. Annealed importance sampling reversible jump MCMC algorithms. *J. Comput. Graph. Statist.*, 22(3):623–648, 2013.
[58] C. Kipnis and S. S. Varadhan. Central limit theorem for additive functionals of reversible Markov processes and applications to simple exclusions. *Comm. Math. Phys.*, 104(1):1–19, 1986.
[59] G. Kitagawa. Monte Carlo filter and smoother for non-Gaussian nonlinear state space models. *J. Comput. Graph. Statist.*, 5(1):1–25, 1996.
[60] P. E. Kloeden and E. Platen. *Numerical solution of stochastic differential equations*. Springer, 1992.
[61] S. J. Koopman, N. Shephard, and D. Creal. Testing the assumptions behind importance sampling. *J. Econometrics*, 149(1):2 – 11, 2009.
[62] A. Lee and K. Łatuszynski. Variance bounding and geometric ergodicity of Markov chain Monte Carlo kernels for approximate Bayesian computation. *Biometrika*, 101(3):655–671, 2014.
[63] A. Lee and N. Whiteley. Variance estimation and allocation in the particle filter. Preprint arXiv:1509.00394v2, 2015.
[64] A. Lee, C. Yau, M. B. Giles, A. Doucet, and C. C. Holmes. On the utility of graphics cards to perform massively parallel simulation of advanced Monte Carlo methods. *J. Comput. Graph. Statist.*, 19(4):769–789, 2010.
[65] L. Lin, K. Liu, and J. Sloan. A noisy Monte Carlo algorithm. *Phys. Rev. D*, 61, 2000.
[66] F. Lindsten, R. Douc, and E. Moulines. Uniform ergodicity of the Particle Gibbs sampler. *Scand. J. Stat.*, 42(3):775–797, 2015.
[67] J. S. Liu. *Monte Carlo Strategies in Scientific Computing*. Springer-Verlag, New York, 2003.
[68] A.-M. Lyne, M. Girolami, Y. Atchade, H. Strathmann, and D. Simpson. On Russian roulette estimates for Bayesian inference with doubly-intractable likelihoods. *Statist. Sci.*, 30(4):443–467, 2015.
[69] P. Marjoram, J. Molitor, V. Plagnol, and S. Tavaré. Markov chain Monte Carlo without likelihoods. *Proc. Natl. Acad. Sci. USA*, 100(26):15324–15328, 2003.
[70] M. Maxwell and M. Woodroofe. Central limit theorems for additive functionals of Markov chains. *Ann. Probab.*, 28(2):713–724, 2000.
[71] D. McLeish. A general method for debiasing a Monte Carlo estimator. *Monte Carlo Methods Appl.*, 17(4):301–315, 2011.
[72] S. Meyn and R. L. Tweedie. *Markov Chains and Stochastic Stability*. Cambridge University Press, 2nd edition, 2009.
[73] T. P. Minka. Expectation propagation for approximate Bayesian inference. In *Proceedings of the Seventeenth conference on Uncertainty in artificial intelligence*, pages 362–369, 2001.
[74] R. M. Neal. Annealed importance sampling. *Statist. Comput.*, 11(2):125–139, 2001.
[75] E. Nummelin. MC's for MCMC'ists. *Int. Statist. Rev.*, 70(2):215–240, 2002.
[76] H. E. Ogden. On asymptotic validity of naive inference with an approximate likelihood. *Biometrika*, 104(1):153–164, 2017.
[77] A. Owen and Y. Zhou. Safe and effective importance sampling. *J. Amer. Statist. Assoc.*, 95(449):135–143, 2000.
[78] A. B. Owen. Statistically efficient thinning of a Markov chain sampler. *J. Comput. Graph. Statist.*, 26(3):738–744, 2017.
[79] M. K. Pitt, R. dos Santos Silva, P. Giordani, and R. Kohn. On some properties of Markov chain Monte Carlo simulation methods based on the particle filter. *J. Econometrics*, 171(2):134–151, 2012.
[80] M. K. Pitt, M.-N. Tran, M. Scharth, and R. Kohn. On the existence of moments for high dimensional importance sampling. Preprint arXiv:1307.7975, 2013.
[81] D. Prangle. Lazy ABC. *Statist. Comput.*, 26(1-2):171–185, 2016.
[82] M. Quiroz, M. Villani, and R. Kohn. Exact subsampling MCMC. Preprint arXiv:1603.08232v2, 2016.
[83] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2016.

[84] C.-H. Rhee and P. W. Glynn. Unbiased estimation with square root convergence for SDE models. *Oper. Res.*, 63(5):1026–1043, 2015.
[85] G. O. Roberts and J. S. Rosenthal. Harris recurrence of Metropolis-within-Gibbs and trans-dimensional Markov chains. *Ann. Appl. Probab.*, 16(4):2123–2139, 2006.
[86] G. O. Roberts and R. L. Tweedie. Geometric convergence and central limit theorems for multi-dimensional Hastings and Metropolis algorithms. *Biometrika*, 83(1):95–110, 1996.
[87] H. Rue, S. Martino, and N. Chopin. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 71(2):319–392, 2009.
[88] S. Särkkä. *Bayesian filtering and smoothing*. Cambridge University Press, 2013.
[89] D. Sen, A. H. Thiery, and A. Jasra. On coupling particle filter trajectories. *Statist. Comput.*, 28(2):461–475, 2018.
[90] N. Shephard and M. K. Pitt. Likelihood analysis of non-Gaussian measurement time series. *Biometrika*, 84(3):653–667, 1997.
[91] C. Sherlock, A. H. Thiery, and A. Lee. Pseudo-marginal Metropolis–Hastings sampling using averages of unbiased estimators. *Biometrika*, 104(3):727–734, 2017.
[92] C. Sherlock, A. H. Thiery, G. O. Roberts, and J. S. Rosenthal. On the efficiency of pseudo-marginal random walk Metropolis algorithms. *Ann. Statist.*, 43(1):238–275, 2015.
[93] S. S. Singh, F. Lindsten, and E. Moulines. Blocking strategies and stability of particle Gibbs samplers. *Biometrika*, 104(1):953–969, 2017.
[94] A. Tan, H. Doss, and J. P. Hobert. Honest importance sampling with multiple Markov chains. *J. Comput. Graph. Statist.*, 24(3):792–826, 2015.
[95] L. Tierney. Markov chains for exploring posterior distributions. *Ann. Statist.*, 22(4):1701–1728, 1994.
[96] M.-N. Tran, M. Scharth, M. K. Pitt, and R. Kohn. Importance sampling squared for Bayesian inference in latent variable models. Preprint arXiv:1309.3339v3, 2014.
[97] M. Vihola. Robust adaptive Metropolis algorithm with coerced acceptance rate. *Statist. Comput.*, 22(5):997–1008, 2012.
[98] M. Vihola. Unbiased estimators and multilevel Monte Carlo. *Oper. Res.*, 66(2):448–462, 2017.
[99] M. Vihola, J. Helske, and J. Franks. Importance sampling type estimators based on approximate marginal Markov chain Monte Carlo. Preprint arXiv:1609.02541v3, 2017.
[100] N. Whiteley. Stability properties of some particle filters. *Ann. Appl. Probab.*, 23(6):2500–2537, 2013.
[101] N. Whiteley and A. Lee. Twisted particle filters. *Ann. Statist.*, 42(1):115–141, 2014.
[102] D. J. Wilkinson. Parallel Bayesian computation. In E. J. Kontoghiorghes, editor, *Handbook of Parallel Computing and Statistics*, pages 481–512. Chapman & Hall/CRC, 2005.

University of Jyväskylä, Department of Mathematics and Statistics, P.O.Box 35, FI-40014 University of Jyväskylä, Finland
Email address, Matti Vihola: matti.vihola@iki.fi

Linköping University, Department of Science and Technology, Media and Information Technology, Campus Norrköping SE-601 74 Norrköping, Sweden
Email address, Jouni Helske: jouni.helske@iki.fi

University of Jyväskylä, Department of Mathematics and Statistics, P.O.Box 35, FI-40014 University of Jyväskylä, Finland
Email address, Jordan Franks: jordan.j.franks@jyu.fi

Article [B]

Importance sampling correction versus standard averages of reversible MCMCs in terms of the asymptotic variance

Jordan Franks and Matti Vihola

Preprint arXiv:1706.09873v4, 2017.

IMPORTANCE SAMPLING CORRECTION VERSUS STANDARD AVERAGES OF REVERSIBLE MCMCS IN TERMS OF THE ASYMPTOTIC VARIANCE

JORDAN FRANKS AND MATTI VIHOLA

Abstract. We establish an ordering criterion for the asymptotic variances of two consistent Markov chain Monte Carlo (MCMC) estimators: an importance sampling (IS) estimator, based on an approximate reversible chain and subsequent IS weighting, and a standard MCMC estimator, based on an exact reversible chain. Essentially, we relax the criterion of the Peskun type covariance ordering in order to consider two different invariant probabilities, and obtain, in place of a strict ordering of asymptotic variances, a bound of the asymptotic variance of IS by that of the direct MCMC. Simple examples show that IS can have arbitrarily better or worse asymptotic variance than Metropolis-Hastings and delayed acceptance (DA) MCMC. Our ordering implies that IS is guaranteed to be competitive up to a factor depending on the supremum of the (marginal) IS weight. We elaborate upon the criterion in case of unbiased estimators as part of an auxiliary variable framework. We show how the criterion implies asymptotic variance guarantees for IS in terms of pseudomarginal (PM) and DA corrections, essentially if the ratio of exact and approximate likelihoods is bounded. We also show that convergence of the IS chain can be less affected by unbounded high-variance unbiased estimators than PM and DA chains.

1. Introduction

Markov chain Monte Carlo (MCMC) algorithms are important for sampling. They are widely applicable and asymptotically exact under mild hypotheses. As they take considerable time to run, it is of interest to know which MCMCs are more efficient. The standard measure of statistical efficiency for MCMCs is the asymptotic variance, as it corresponds with the limiting variance of a $\sqrt{n}$-central limit theorem (CLT) (cf. Proposition 1). A famous ordering criterion for the asymptotic variances of two reversible Markov chains is the Peskun ordering [50, Thm. 2.1.1], extended to general state spaces by Tierney [61, Thm. 4], and elaborated upon in [43, Thm. 4.2] in terms of the lag-1 auto-covariance, whence it is sometimes called the 'covariance ordering' [see also 61, Proof of Lem. 3]. The result has been applied and extended to various settings, e.g. continuous-time chains [36, 44], Gibbs [4] and hybrid [1, 41] samplers, and to pseudomarginal (PM) chains [3, 8, 9, 23, 59], where it is used in particular for the proof of the 'convex order' criterion for PM chains [9]. The aforementioned orderings have in common that the two chains being compared share the same invariant probability, at least marginally.

Key words and phrases. Asymptotic variance, delayed acceptance, importance sampling, Markov chain Monte Carlo, pseudomarginal algorithm, unbiased estimator.

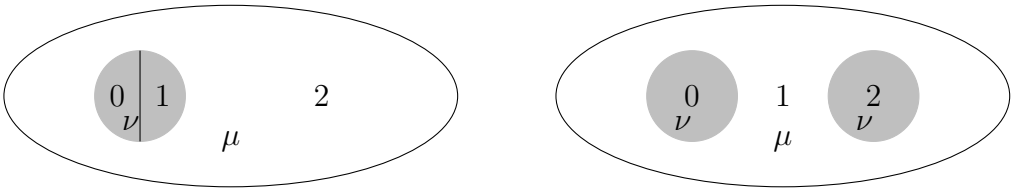
1.1. **Ordering criterion.** We suggest a Peskun type ordering for importance sampling (IS) MCMC estimators to compare with standard MCMC estimators. IS is based on a chain targeting an approximate probability μ of the target probability ν of interest. A final IS correction phase is then used, which involves an IS weight w satisfying $\mu(wf) = \nu(f)$ for suitable functions f [cf. 21, 28, 29, 31, 49, 65]. We seek to compare IS with the typical MCMC, i.e. standard averages of a reversible chain targeting ν. Instead of a *strict* ordering as in the covariance ordering, we obtain a *quasi*-ordering involving constants depending on w and the function variance (Theorem 2). A product space version for augmented IS kernels (Theorem 5) will turn out to be particularly useful when unbiased estimators are introduced as part of an auxiliary variable framework.

1.2. **Popular direct MCMCs.** The workhorse of the reversible MCMC world is the Metropolis-Hastings algorithm [cf. 31], or equivalently, its novel reformulation in terms of unbiased estimators, the PM algorithm [38, 6]. A PM variant, known as delayed acceptance (DA) [cf. 38, 40, 17], has drawn considerable attention recently as a means to accelerate PM [cf. 10, 18, 20, 24, 30, 52, 56, 57, 58, 60, 65]; see §3.2.1 and §6 for in fact two different possible types of DA 'correction,' and §7.3 for examples of DA in different settings. Although the statistical efficiency of DA is worse than PM by the covariance ordering, the overall computational efficiency of DA can be better, as judged by empirical wall-clock time to reach a certain confidence interval assuming the chains start at stationarity. The acceleration is based on decreasing the number of expensive calculations in the standard PM implementation by using an intermediate approximation as a 'screening' step (cf. Algorithm 4, p. 14).

1.3. **The IS vs. direct MCMC question.** As IS and DA are consistent MCMCs which can use the same intermediate approximation, and along with PM are asymptotically exact, there is a choice about which algorithm to use. A study of self-normalised IS versus the independence Metropolis-Hastings has been made in [39], who also explains why the objective function plays a rôle in IS, which can be *super-efficient*, i.e. better than sampling i.i.d. from the target distribution (but IS can also be worse). Asymptotic variances are explicitly computed and compared in some discrete examples in [12] who find that IS and Metropolis-Hastings can be competitive, but that Metropolis-Hastings can do much better (see also [11, §4.2]). On the other hand, [62] study independent IS with unbiased estimators, and find that this performs better than PM in their experiments (see also [16]). The algorithm of [51] in the approximate Bayesian computation setting involving a two-phase IS approach is also found to perform better experimentally than a direct approach. The IS versus DA question is noted in [18, §3.3.3], who mention the likely improvement of IS over DA in massive parallelisation. A methodological comparison of the alternatives in the general MCMC and joint inference context is made in [65], who investigate empirically the relative efficiencies, finding that IS and DA can be competitive, with IS doing slightly better than DA in their experiments, with little or no parallelisation. The gap widens with increased parallelisation, a known strength of the IS correction [cf. 18, 35, 51, 65].

1.4. **Either may do arbitrarily better in asymptotic variance.** Ignoring computational and implementation aspects for the most part (but see the discussion in §7), this paper seeks to address the question generally in terms of the asymptotic variances. The prior references and the examples provided in Appendix D show that the answer can not be completely simple. We give toy examples where either IS or PM/DA can do (arbitrarily) better than the other in terms of asymptotic variance (cf. Figure 1, and Appendix D for details).

FIGURE 1. Marked regions for examples



(A) 'direct MCMC better' case (B) 'IS better' case

1.5. **Intuition why IS can help with multimodal targets.** As the examples in Appendix D show, IS can do well when μ is uniform or close to uniform, which allows for good mixing between modes of the target [cf. 28]. The benefit of a slow transition to the target is well-known throughout the IS repertoire, e.g. in sequential IS [cf. 40] and annealed IS [45]. The possible mixing improvement of the IS first phase as a 'warm start' is not shared by PM/DA, which targets ν directly. In the simple setting of the examples, where no unbiased estimators are used, Corollary 4 gives guarantees that IS performs competitively with PM/DA if the IS weight is bounded, which is always true if μ is the uniform density and ν is a bounded density.

1.6. **Unbiased estimators.** After extending Theorem 2 to an auxiliary variable framework suitable for pseudomarginal chains and unbiased estimators (Theorem 12), we show a quasi-ordering for IS and PM/DA (Theorem 14), implying asymptotic variance guarantees for IS in terms of PM/DA. When the IS weight is estimated unbiasedly, the essential assumption for our ordering to hold is boundedness of the IS weight estimator conditional means, not necessarily the IS weight estimator itself. Also, the objective function may depend on the latents, which is usually not the case for Peskun type orderings for PM chains (cf. Remark 6(i)). These relaxations are ultimately due to the augmented kernel structure of the IS chain (17) (cf. Lemma 22(iii)).

1.7. **Convergence considerations.** We complement our ordering results by showing that the IS chain is geometrically or uniformly ergodic if and only if the IS base chain has the corresponding property (Lemma 22). The error of MCMC is due to two factors: the distance of the chain from stationarity, related e.g. by the burn-in time, and the Monte Carlo error, related by the asymptotic variance [cf. 33, 46]. In case of unbiased estimators for the weight, we describe how an IS chain can be geometrically ergodic, in contradistinction to PM/DA (cf. §7.1).

As geometrically ergodic chains converge geometrically fast from all initial positions, IS may be a good choice when little is known about the IS weight but good approximate (marginal) Markov chains are available (cf. §7.2). Also, geometric and uniform ergodicity are likely to benefit adaptive MCMC [cf. 7], at least based on the existing theory [cf. 5, 54], as well as the construction of estimators for the asymptotic variance [cf. 25]. The minimal requirement on the IS weight is a simple support condition (Assumption 1 or Assumption 4(iii)), which can often be ensured easily in practice (cf. Remark 8(ii) of §5).

1.8. **Possible extensions.** Although we only apply our criterion (Theorem 12) to a comparison of IS with PM/DA (Theorem 14), and allude to the possible utility of approximate Gibbs samplers in §3.2.2, our criterion may also be relevant for a comparison of IS with other direct reversible MCMCs, such as 'MHAAR' [3] and 'correlated PM' [19]. By decreasing the variance occurring in the PM acceptance ratio, these algorithms seek to improve upon mixing properties of PM type chains. Our result (Theorem 14) applies to a comparison of IS versus 'DA correction' (19) of approximate reversible chains, such as approximate versions of the previously mentioned chains. Further studies may be interesting in the specific settings of e.g. annealed IS [45], likelihood-free inference [cf. 51], multi-stage DA [10], multilevel Monte Carlo [cf. 20], and sequential Monte Carlo [cf. 16, 40]. See also §7.3 for possible application settings.

1.9. **Related work.** Earlier studies involving IS and direct MCMC have been made in e.g. [12, 39, 51, 62, 65] (cf. §1.3). We consider here general reversible Markov chains, in particular PM/DA, and seek a Peskun type ordering of the asymptotic variances. Our elaboration of the IS correction with the use of unbiased estimators in §5 aligns with the IS type correction of [65], but we only consider here nonnegative IS weights (and reversible base chains). The work [65] includes consistency and CLT results for the IS type correction, as well as implementation and computational efficiency comparisons for IS, PM, and DA in experiments in state space models.

1.10. **Outline.** After preliminaries in §2, we state in §3 the Peskun type ordering result for normalised IS (Theorems 2), its implication for IS versus PM/DA in a simple setting (Corollary 4), and augmented IS kernels (Theorem 5). We define jump chains and self-normalised importance sampling (SNIS) in §4, before proceeding to §5, where we consider a general auxiliary variable framework which accommodates IS and PM type schemes that use unbiased estimators. The PM type algorithms and kernels which we consider are given in §6, and we compare them with IS (Theorem 14). We discuss some stability, implementation, and computational efficiency considerations in §7. Proofs of the Peskun type orderings are given in Appendix A. Dirichlet form bounds and proof of the main comparison application (Theorem 14) are found in Appendix B. Appendix C mentions some properties of augmented chains. Appendix D contains the examples mentioned earlier in §1.4.

2. Notation and definitions

2.1. **Notation.** The spaces $\mathbf{X}$ we consider are assumed equipped with a σ-algebra, denoted $\mathcal{B}(\mathbf{X})$, and with a σ-finite dominating measure, denoted '$\mathrm{d}x$.' Product spaces will be assumed equipped with their product σ-algebras and corresponding product measures. If μ is a probability density on $\mathbf{X}$, we denote the corresponding probability measure with the same symbol, so that $\mu(\mathrm{d}x) = \mu(x)\mathrm{d}x$.

For $p \in [1, \infty)$, we denote by $L^p(\mu)$ the Banach space of equivalence classes of measurable $f : \mathbf{X} \to \mathbb{R}$ satisfying $\|f\|_p < \infty$ under the norm $\|f\|_{L^p(\mu)} := \{\int |f(x)|^p \mu(\mathrm{d}x)\}^{1/p}$. We similarly define $L^\infty(\mu)$ under the norm $\|f\|_\infty := \mu\text{-}\operatorname{ess\,sup}_{x\in\mathbf{X}} |f(x)|$. We denote by $L^p_0(\mu)$ the subset of $L^p(\mu)$ with $\mu(f) = 0$, where $\mu(f) := \int f(x)\mu(\mathrm{d}x)$. For $f \in L^1(\mu)$ and $K_x(\mathrm{d}x')$ a Markov kernel on $\mathbf{X}$, we define $\mu K(A) := \int \mu(\mathrm{d}x) K_x(A)$ for $A \in \mathcal{B}(\mathbf{X})$, $Kf(x) := \int K_x(\mathrm{d}x') f(x')$, and inductively $K^n f(x) := K^{n-1}(Kf)(x)$ for $n \geq 2$. For $f, g \in L^2(\mu)$, we define $\langle f, g \rangle_\mu := \int f(x)g(x)\mu(\mathrm{d}x)$, $\|f\|_\mu := (\langle f, f\rangle_\mu)^{1/2}$, and $\operatorname{var}_\mu(f) := \mu(f^2) - \mu(f)^2$.

For $m \in \mathbb{N}$ and $x^{(i)} \in \mathbf{X}$ for $i = 1, \ldots m$, we write $x^{(1:m)} := (x^{(1)}, \ldots, x^{(m)})$. If $x = x^{(1:m)}$, then $x^{(-i)}$ is the vector of length $m-1$ obtained from x by deleting the ith entry. Throughout, ν will denote the target probability of interest, and for $\varphi \in L^1(\nu)$ we set $\bar{\varphi} := \varphi - \nu(\varphi)$, element of $L^1_0(\nu)$.

2.2. **Definitions.** Let μ and ν be σ-finite measures on $\mathbf{X}$. If $\mu(A) = 0$ implies $\nu(A) = 0$ for all $A \in \mathcal{B}(\mathbf{X})$, we say that ν is *absolutely continuous* with respect to μ, and write $\nu \ll \mu$. Suppose $\nu \ll \mu$. A *Radon-Nikodým derivative* of ν with respect to μ is a measurable function $\frac{d\nu}{d\mu}(x)$ on $\mathbf{X}$ such that $\mu(\frac{d\nu}{d\mu} f) = \nu(f)$ for all $f \in L^1(\nu)$. If also μ and ν are probability densities, then it is easy to see that $\frac{d\nu}{d\mu}(x)$ exists in $L^1(\mu)$, and is equivalent with $\frac{\nu(x)}{\mu(x)}$.

Let μ be a probability on $\mathbf{X}$. A Markov chain K on $\mathbf{X}$ is *μ-invariant* if $\mu K = \mu$. If also $\langle f, Kf\rangle_\mu \geq 0$ for all $f \in L^2(\mu)$, then K is *positive*. If $\mu(\mathrm{d}x)K_x(\mathrm{d}x') = \mu(\mathrm{d}x')K_{x'}(\mathrm{d}x)$, then K is said to satisfy *detailed balance* with respect to μ, or briefly, K is *μ-reversible*. This implies that K is μ-invariant, and that the *Dirichlet form* $\mathcal{E}_K(f)$ for $f \in L^2(\mu)$ satisfies

$$\mathcal{E}_K(f) := \langle f, (1-K)f\rangle_\mu = \frac{1}{2}\int \mu(\mathrm{d}x)K_x(\mathrm{d}x')\big(f(x) - f(x')\big)^2. \tag{1}$$

Definition 1 (Harris ergodic)**.** A Markov chain K is *μ-Harris ergodic* if K is μ-invariant, ψ-irreducible, and Harris recurrent.

See [42] for details, and for the definition of ψ-irreducibility. Most MCMC schemes are Harris ergodic, although a careless implementation can lead to a non-Harris chain [cf. 53].

Definition 2 (Asymptotic variance)**.** Let (X_k) be a μ-Harris ergodic Markov chain with transition K. For $f \in L^2(\mu)$ the *asymptotic variance* of f with respect to K is defined, whenever the limit exists in $[0, \infty]$, as

$$\operatorname{var}(K, f) := \lim_{n\to\infty} \frac{1}{n}\mathbf{E}\Big[\Big(\sum_{k=1}^{n}[f(X_k^{(s)}) - \mu(f)]\Big)^2\Big], \tag{2}$$

where $(X_k^{(s)})$ denotes a stationary version of the chain (X_k), i.e. $X_0^{(s)} \sim \mu$.

For reversible K, which is the focus of this paper, $\mathrm{var}(K, f)$ always exists in $[0, \infty]$ [cf. 61]. Moreover, a CLT holds under general conditions.

Proposition 1. *[34, Cor. 1.5] and [42, Prop. 17.1.6] Let $(X_k)_{k\geq 1}$ be an aperiodic μ-reversible Harris ergodic Markov chain with transition K. If $f \in L^2(\mu)$ and $\mathrm{var}(K, f) < \infty$, then, for all initial distributions,*

$$\frac{1}{\sqrt{n}}\Big(\sum_{k=1}^{n}[f(X_k) - \mu(f)]\Big) \xrightarrow{n\to\infty} \mathcal{N}\big(0, \mathrm{var}(K, f)\big), \qquad \textit{in distribution}, \tag{3}$$

where $\mathcal{N}(a, b^2)$ is a normal random variable with mean a and variance b^2.

See [42] for definition of aperiodic. Proposition 1 above explains the importance of the asymptotic variance, since it is the CLT limiting variance.

3. Peskun type ordering for normalised importance sampling

3.1. **General case.** Let μ and ν be probability measures on a space $\mathbf{X}$, and let $w : \mathbf{X} \to [0, \infty)$ be a nonnegative measurable function.

Assumption 1 (Importance sampling)**.** A triplet (μ, ν, w) is such that $\nu \ll \mu$ and $w(x) = \frac{d\nu}{d\mu}(x)$ is the Radon-Nikodým derivative.

Assumption 2. A heptuple $(\mu, \nu, w, K, L, \underline{c}, \overline{c})$ is such that (μ, ν, w) satisfies Assumption 1, K and L are Harris ergodic Markov chains reversible with respect to μ and ν, respectively, and the constants $\underline{c}$, $\overline{c} \geq 0$ satisfy

(a) $\underline{c}\,\mathcal{E}_K(g) \leq \mathcal{E}_L(g) \leq \overline{c}\,\mathcal{E}_K(g)$, for all $g \in L^2(\mu)$, and
(b) $\underline{c} \leq w \leq \overline{c}$, μ-a.e.

Theorem 2. *If Assumption 2 holds, then for all $\varphi \in L^2(\nu)$,*

$$\mathrm{var}(K, w\varphi) + \mathrm{var}_\mu(w\bar{\varphi}) \leq \overline{c}\,\big[\mathrm{var}(L, \varphi) + \mathrm{var}_\nu(\varphi)\big], \tag{4}$$

$$\mathrm{var}(K, w\varphi) + \mathrm{var}_\mu(w\bar{\varphi}) \geq \underline{c}\,\big[\mathrm{var}(L, \varphi) + \mathrm{var}_\nu(\varphi)\big]. \tag{5}$$

Remark 3. Here, we recall the notation $\bar{\varphi} := \varphi - \nu(\varphi)$. Regarding Theorem 2, whose proof is given in Appendix A:

(i) If $w = 1$ constant, in which case $\mu = \nu$, it reduces to [4, Lemma 32]. If also $(\underline{c}, \overline{c}) = (0, 1)$, it is the covariance ordering [43, Thm. 4.2], which is a Peskun [50, 61] type criterion based on the Dirichlet form [see also 61, Proof of Lem. 3].
(ii) The assumptions are the same as those of [37, Lem. 13.22] about comparison of mixing times in the countable state space context.
(iii) (5) holds even if we 'forget' $\overline{c}$, i.e. set $\overline{c} = \infty$ but also require $w\varphi \in L^2(\mu)$. In practice, (5) is usually useless since we can only assume $\underline{c} = 0$.

3.2. **Intermezzo: some simple comparison examples.** We show how Theorem 2 implies results in two simple and common settings before introducing the various machinery that occupies the remainder of this paper.

3.2.1. *With Metropolis-Hastings and delayed acceptance correction.* Let q be a probability kernel and ν a probability on $\mathbf{X}$. With $\mathbf{R}^{(\nu)} \subset \mathbf{X}^2$ the 'symmetric set' as in [61, Prop. 1], we let $0 < r^{(\nu)}(x, x') < \infty$ denote the Radon-Nikodým derivative

$$r^{(\nu)}(x, x') := \frac{\nu(\mathrm{d}x')q_{x'}(\mathrm{d}x)}{\nu(\mathrm{d}x)q_x(\mathrm{d}x')}$$

for $(x, x') \in \mathbf{R}^{(\nu)}$, and otherwise we define $r^{(\nu)}(x, x') := 0$. The Metropolis-Hastings chain $\mathrm{MH}(q \to \nu)$ with proposal q and target ν has kernel

$$P_x(\mathrm{d}x') := q_x(\mathrm{d}x') \min\left\{1, r^{(\nu)}(x, x')\right\} + [1 - \alpha_{\mathrm{MH}}(x)]\delta_x(\mathrm{d}x'), \tag{6}$$

where δ_x is the Dirac measure at x, $\alpha_{\mathrm{MH}}(x) := P_x(\mathbf{X}\backslash\{x\})$ [cf. 31, 40].

We show DA can take two forms: DA0 correction, and DA1 correction. In the following we define the two corrections. These corrections will turn the approximate chain into an exact chain, targeting directly the probability most of interest. Traditionally, DA has meant DA0 correction of PM [cf. 10, 17]. However, we will see that DA0 is applicable to the more general class of 'proposal-rejection' chains, while a different type of algorithm, which we call DA1 correction, is applicable to general reversible chains. DA1 has been considered in [40, 'surrogate transition method,' §9.4.3]. Both DA corrections, DA0 and DA1, will lend themselves to comparison with IS correction.

We call a kernel K a *μ-proposal-rejection* kernel if it is μ-reversible and can be written as

$$K_x(\mathrm{d}x') = q_x(\mathrm{d}x')\alpha(x, x') + \left(1 - \int q_x(\mathrm{d}y)\alpha(x, y)\right)\delta_x(\mathrm{d}x')$$

for some measurable function $\alpha : \mathbf{X} \times \mathbf{X} \to [0, 1]$. This obviously includes the case where K is a $\mathrm{MH}(q \to \mu)$ kernel, but also includes, for example, DA (leading to 'multi-stage' DA [cf. 10]) and 'MHAAR' [3]. Proposal-rejection kernels have also been considered in [64], where the abstraction arose from consideration of the marginal chain of a certain pseudomarginal chain arising from approximate Bayesian computation.

If (μ, ν, w) satisfies Assumption 1 and K is a μ-proposal-rejection kernel, then we define *DA0 correction* of the proposal-rejection kernel K to be the kernel

$$K_x^{\mathrm{DA0}}(\mathrm{d}x') := q_x(\mathrm{d}x')\alpha(x, x') \min\left\{1, w(x')/w(x)\right\} + [1 - \alpha_{\mathrm{DA0}}(x)]\delta_x(\mathrm{d}x'), \tag{7}$$

where $\alpha_{\mathrm{DA0}}(x) := K_x^{\mathrm{DA0}}(\mathbf{X}\backslash\{x\})$. It is straightforward to check that K^{DA0} is ν-reversible; this is the standard delayed acceptance kernel in the case K is $\mathrm{MH}(q \to \mu)$ (cf. [10, 40] and §6.1).

If (μ, ν, w) satisfies Assumption 1, and K is a μ-reversible kernel, we define the *DA1 correction* to be the chain with transition kernel given by

$$K_x^{\mathrm{DA1}}(\mathrm{d}x') := K_x(\mathrm{d}x') \min\left\{1, w(x')/w(x)\right\} + [1 - \alpha_{\mathrm{DA1}}(x)]\delta_x(\mathrm{d}x'), \tag{8}$$

where $\alpha_{\mathrm{DA1}}(x) := K_x^{\mathrm{DA1}}(\mathbf{X}\backslash\{x\})$; K^{DA1} is ν-reversible, as is straightforward to check.

It is a direct application of the covariance (or Peskun) ordering to see that the asymptotic variances of K^{DA0} and K^{DA1} are the same, where K is a μ-proposal-rejection chain. However, we will see in §6 that K^{DA0} is likely to be more computationally efficient in practice.

Corollary 4. *Suppose (μ, ν, w) satisfies Assumption 1, and that*

(I) $L = K^{DA0}$ (7), where K is a μ-proposal-rejection kernel,
(II) $L = K^{DA1}$ (8), where K is a μ-reversible kernel, or
(III) $L = P$ (6), and $K = MH(q \to \mu)$.

Assume K and L form Harris ergodic chains. The following statements hold.

(i) If $w \leq \bar{c}$ μ-a.e., then for all $\varphi \in L^2(\nu)$, (4) holds.
(ii) If $w \geq \underline{c}$ μ-a.e., then for all $\varphi \in L^2(\nu)$ with $w\varphi \in L^2(\mu)$, (5) holds.

The result follows from Theorem 2 and Lemma 21 of Appendix B.

3.2.2. *With Gibbs samplers and delayed acceptance correction.* Suppose ν is probability density on a product space $\mathbf{X} := \mathbf{X}_1 \times \cdots \times \mathbf{X}_m$, with $m \in \mathbb{N}$. Let $\mathcal{I}$ be a Markov kernel on the discrete set $\{1, \ldots, m\}$. For example, the 'scan' $\mathcal{I}$ could be a systematic scan: $\mathcal{I}_i(j) = \delta_{i+1}(j)$ for $i = 1, \ldots, m-1$, and $\mathcal{I}_m(j) = \delta_1(j)$. Or, $\mathcal{I}$ could be a random scan: $\mathcal{I}_i(j) = 1/m$ for all $j \in \{1, \ldots, m\}$. For each $i = 1, \ldots, m$, let $q^{(i)}_{x^{(-i)}}(x'^{(i)})$ be a transition density from $\mathbf{X}^{(-i)}$ to $\mathbf{X}^{(i)}$, which, to avoid technical problems, may be assumed strictly positive. We define a Markov transition density q on $\mathbf{X} \times \{1, \ldots, m\}$,

$$q_{x,i}(x', j) := \mathcal{I}_i(j)\, q^{(j)}_{x^{(-j)}}(x'^{(j)}).$$

The Metropolis-within-Gibbs with random scan, $\mathrm{MGrs}(q \to \nu)$, has kernel

$$P_{x,i}(x', j) := q_{x,i}(x', j) \min\left\{1, \frac{\nu(x') q^{(j)}_{x'^{(-j)}}(x^{(j)})}{\nu(x) q^{(j)}_{x^{(-j)}}(x'^{(j)})}\right\} + [1 - \alpha_{\mathrm{MGrs}}(x, i)]\delta_{x,i}(x', j),$$

where $\alpha_{\mathrm{MGrs}}(x, 1) := L_{x,i}(\mathbf{X} \times (1 : m) \backslash \{(x, i)\})$, which is reversible as an MH kernel, and targets ν marginally [cf. 40]. If $q^{(i)}_{x^{(-i)}}(x'^{(i)}) = \nu(x'^{(i)} | x^{(-i)})$ for all $i = 1, \ldots, m$, then the acceptance ratio is identically 1 and $\mathrm{MGrs}(q \to \nu)$ becomes the standard Gibbs sampler (without the Metropolis-Hastings step) [cf. 27, §11.3].

Suppose μ is a density on $\mathbf{X}$ with $\nu \ll \mu$. Because the MGrs may be viewed as a full-dimensional MH on $\mathbf{X} \times \{1, \ldots, m\}$, Corollary 4 applies, with $K = \mathrm{MGrs}(q \to \mu)$.

3.3. **Marginalisations and augmented importance sampling kernels.** Let $\mathbf{X} = \mathbf{T} \times \mathbf{Y}$, where $\mathbf{T}$ and $\mathbf{Y}$ are measurable spaces. For a probability μ on $\mathbf{X}$, denote by $\mu^*(\mathrm{d}\theta) = \mu(\mathrm{d}\theta, \mathbf{Y})$ its marginal probability. If (μ, ν, w) on $\mathbf{X}$ satisfies Assumption 1, then $\nu^* \ll \mu^*$, and with $w^*(\theta) := \frac{d\nu^*}{d\mu^*}(\theta)$, the triplet (μ^*, ν^*, w^*) satisfies Assumption 1 on $\mathbf{T}$.

Assumption 3. Assumption 2, with Assumption 2(a–b) replaced with
(a) $\underline{c}\,\mathcal{E}_K(g) \leq \mathcal{E}_L(g) \leq \bar{c}\,\mathcal{E}_K(g)$, for all $g \in L^2(\mu^*)$, and
(b) $\underline{c} \leq w^* \leq \bar{c}$, μ^*-a.e.

We introduce the notion of an augmented Markov kernel, as in [9, 65].

Definition 3. Let $\dot{\mu}$ be some probability on $\mathbf{T}$, let $\dot{K}$ be a $\dot{\mu}$-invariant Markov kernel on $\mathbf{T}$, and let $Q_\theta(\mathrm{d}y)$ be a probability kernel from $\mathbf{T}$ to $\mathbf{Y}$. The *Q-augmentation of $\dot{K}$*, or the *Q-augmented kernel K*, is a Markov kernel on $\mathbf{X}$, with transition K and invariant measure μ, given by

$$K_{\theta\, y}(\mathrm{d}\theta', \mathrm{d}y') = \dot{K}_\theta(\mathrm{d}\theta') Q_{\theta'}(\mathrm{d}y'), \quad \text{and} \quad \mu(\mathrm{d}\theta, \mathrm{d}y) = \dot{\mu}(\mathrm{d}\theta) Q_\theta(\mathrm{d}y). \tag{9}$$

Theorem 5. *Suppose Assumption 3 holds, and that K is an augmented kernel as in Definition 3. Let $\varphi \in L^2(\nu)$ with $w\varphi \in L^2(\mu)$. With $\mathscr{N}_K := 0$ if K is positive, and $\mathscr{N}_K := 1$ if not, the following bound holds:*

$$\operatorname{var}(K, w\varphi) \leq \overline{c}\big[\operatorname{var}(L, \varphi) + \operatorname{var}_\nu(\varphi)\big] + (1 + 2\mathscr{N}_K) \operatorname{var}_\mu(w\bar{\varphi}) \tag{10}$$

Moreover, if $w\varphi$ only depends on $\theta \in \mathbf{T}$, then (4) holds.

Remark 6. Regarding Theorem 5, whose proof is given in Appendix A:

(i) The function φ (and $w\varphi$) is allowed to depend on the auxiliary variable $y \in \mathbf{Y}$, unlike comparison results in the PM setting (cf. [8, Thm. 7] and [59, Thm. 1]) that are based on the convex order [9, Thm. 10].
(ii) K is positive iff $\dot{K}$ is positive (Lemma 22 of Appendix C). This is the case e.g. if $\dot{K}$ is a random walk Metropolis-Hastings kernel with normal proposals [13, Lem. 3.1]. See [23, Prop. 3] for more examples.
(iii) See also Remarks 17(iii–iv) in Appendix A about Assumption 3, which also hold for Assumption 2 by trivialising the space $\mathbf{Y}$ (Lemma 18(i)).

4. Jump chains and self-normalised importance sampling

4.1. **Jump chains.** We recall the notion of a jump chain [cf. 22], which is a Markov chain consisting of the accepted states of the original chain.

Definition 4. Let $(\Theta_k)_{k\geq 1}$ be a Markov chain with transition $K_\theta(\mathrm{d}\theta')$. The *jump chain* $(\tilde{\Theta}_k, \tilde{N}_k)_{k\geq 1}$ with transition $\tilde{K}_{\theta n}(\mathrm{d}\theta', \mathrm{d}n')$ and holding times

$$\tilde{N}_j := \min\Big\{i \geq 1 | \Theta_{\tilde{N}^*_{j-1}+i+1} \neq \Theta_{\tilde{N}^*_{j-1}+1}\Big\}, \qquad j \geq 1,$$

is given by $\tilde{\Theta}_1 := \Theta_1$ and $\tilde{\Theta}_{k+1} := \Theta_{\tilde{N}^*_k+1}$, where $\tilde{N}^*_k := \sum_{j=1}^k \tilde{N}_j$, $\tilde{N}^*_0 := 0$.

For a Harris ergodic chain K, $(\tilde{N}_k)_{k\geq 1}$ are independent random variables given $(\tilde{\Theta}_k)_{k\geq 1}$, where $\tilde{N}_k$ is geometrically distributed with parameter $\alpha(\tilde{\Theta}_k)$. Here, $\alpha(\theta) := K(\theta, \mathbf{T}\backslash\{\theta\})$ is the acceptance probability function of K at $\theta \in \mathbf{T}$. See [65, Prop. 27] for this as well as for proof of the following result.

Lemma 7. *Let K be a μ-invariant Markov chain with $\alpha > 0$. The marginal chain $\tilde{K}$ of the jump chain of K has transition $\tilde{K}(\theta, A) = K(\theta, A\backslash\{\theta\})/\alpha(\theta)$, for all $A \in \mathcal{B}(\mathbf{T})$, and is $\tilde{\mu}$-invariant, where $\tilde{\mu}(\mathrm{d}\theta) = \alpha(\theta)\mu(\mathrm{d}\theta)/\mu(\alpha)$. Moreover, K is μ-reversible iff $\tilde{K}$ is $\tilde{\mu}$-reversible, and K is μ-Harris ergodic iff $\tilde{K}$ is $\tilde{\mu}$-Harris ergodic.*

We note that $(\tilde{\Theta}_k, \tilde{N}_k)_{k\geq 1}$ has as its transition the $Q^{(N)}$-augmentation of $\tilde{K}$ (Definition 3), where $\tilde{K}$ is as in Lemma 7 and $Q^{(N)}_\theta(\cdot) \sim \mathrm{Geo}(\alpha(\theta))$ [23].

Different estimators can sometimes be used in place of $(\tilde{N}_k)$, which can lead to lower asymptotic variance of the related MCMC than when not using the jump chain, or when using the jump chain with standard $(\tilde{N}_k)$ [22].

4.2. **Self-normalised importance sampling.** Jump chains can be naturally used with IS estimators, and can lead to improved computational and statistical efficiency [cf. 65]. To avoid redundancy, we shall adhere to the following convention: when we write $(\Theta_k, \mathsf{N}_k, \mathsf{a}, \mu)$, it shall stand simultaneously for $(\tilde{\Theta}_k, \tilde{N}_k, \alpha, \tilde{\mu})$,

corresponding to an IS jump chain (denoted 'ISJ'), and for $(\Theta_k, 1, 1, \mu)$, corresponding to a non-jump IS chain (denoted 'IS0').

Suppose (μ, ν, w) satisfies Assumption 1 and that $(\Theta_k)_{k\geq 1}$ is μ-Harris ergodic. Often one can not evaluate $w(\theta)$. However, one can often evaluate an unnormalised version $w_u(\theta) = c_\xi \cdot w(\theta)$, with $c_\xi > 0$ a (unknown) constant. In this case, for $\varphi \in L^1(\nu)$, one can use the following SNIS estimator,

$$E_n^{SNIS}(\varphi) := \frac{\sum_{k=1}^n \mathsf{N}_k w_u(\Theta_k)\varphi(\Theta_k)}{\sum_{k=1}^n \mathsf{N}_k w_u(\Theta_k)} = \frac{\frac{1}{n}\sum_{k=1}^n \mathsf{N}_k w_u(\Theta_k)\varphi(\Theta_k)}{\frac{1}{n}\sum_{k=1}^n \mathsf{N}_k w_u(\Theta_k)}. \tag{11}$$

By Harris ergodicity, the SNIS estimator is a consistent estimator for $\nu(\varphi)$,

$$E_n^{SNIS}(\varphi) \xrightarrow[\text{a.s.}]{n\to\infty} \frac{\mu(\mathbf{E}[\mathsf{N}_k|\Theta_k]w_u\varphi)}{\mu(\mathbf{E}[\mathsf{N}_k|\Theta_k]w_u)} = \frac{\mu(w_u\varphi/\mathsf{a})}{\mu(w_u/\mathsf{a})} = \nu(\varphi).$$

Next we consider a framework on an extended space, from which a Peskun type ordering for SNIS will trivially follow (Remark 13(ii) of Theorem 12).

5. Unbiased estimators and exact approximation schemes

In an auxiliary variable framework, such as a latent variable model, joint inference involves expectations of the form

$$\nu(f) = \int f(\theta, z)\nu(\mathrm{d}\theta, \mathrm{d}z),$$

where $\theta \in \mathbf{T}$ is the model 'parameter' and $z \in \mathbf{Z}$ is the 'latent variable' or 'state.' The marginal inference case, i.e. when $f(\theta, \cdot) = f(\theta)$ only depends on $\theta \in \mathbf{T}$, is important for model parameter estimation [cf. 6]. State estimation (when θ is viewed as fixed) is possible in the state space model (SSM) setting using sequential Monte Carlo (SMC) [cf. 40], while particle MCMC [2], which uses a specialised SMC within an MCMC, allows for joint inference.

5.1. Exact approximation schemes.

The approximation schemes we consider rely on the existence of PM probability kernels, which represent the laws corresponding to draws from e.g. i.i.d. IS, or from SMC, and which are basic to the PM approach [6].

We associate to a probability kernel $Q_\theta^{(U)}(\mathrm{d}u)$ from $\mathbf{T}$ to a space $\mathbf{U}$ a function $\eta(1) := \eta(\theta, u)$ on $\mathbf{T} \times \mathbf{U}$. For example, if $U \sim Q_\theta^{(U)}(\cdot)$ and θ is fixed, then $\eta(1)$ is an unbiased estimator for the 'likelihood' at θ in the SSM setting [cf. 2]. Let $\mathbf{V}$ be the space

$$\mathbf{V} := \big\{(m, z^{(1:m)}, \zeta^{(1:m)}) : m \in \mathbb{N}, \text{ and } z^{(i)} \in \mathbf{Z}, \zeta^{(i)} \in [0, \infty) \text{ for } i = 1, \dots m\big\}.$$

We then similarly associate to a probability kernel $Q_{\theta u}^{(V)}(\mathrm{d}v)$ from $\mathbf{T} \times \mathbf{U}$ to $\mathbf{V}$, a function $\zeta(1)$ of $v \in \mathbf{V}$, given by

$$\zeta(1) := \sum_{i=1}^m \zeta^{(i)}, \qquad \text{if } v = (m, z^{(1:m)}, \zeta^{(1:m)}).$$

Assumption 4 (Pseudomarginal kernels)**.** The two kernels and two functions defined above determine probability measures μ on $\mathbf{T} \times \mathbf{U}$ and π on $\mathbf{T} \times \mathbf{U} \times \mathbf{V}$, given by

$$\mu(\mathrm{d}\theta, \mathrm{d}u) := c_\eta^{-1} \mathrm{d}\theta Q_\theta^{(U)}(\mathrm{d}u)\eta(1),$$
$$\pi(\mathrm{d}\theta, \mathrm{d}u, \mathrm{d}v) := c_\zeta^{-1} \mathrm{d}\theta Q_\theta^{(U)}(\mathrm{d}u) Q_{\theta u}^{(V)}(\mathrm{d}v)\zeta(1),$$

where c_η and c_ζ are finite normalising constants. For a target probability ν on a space $\mathbf{T} \times \mathbf{Z}$, with $\dot{\nu}(\mathrm{d}\theta) := \nu(\mathrm{d}\theta, \mathbf{Z})$ as before, we assume conditions:

(i) $\dot{\nu} \ll \dot{\mu}$, where $\dot{\mu}(\mathrm{d}\theta) := \mu(\mathrm{d}\theta, \mathbf{U})$, 'approx. marginal posterior'
(ii) $\dot{\nu} = \dot{\pi}$, where $\dot{\pi}(\mathrm{d}\theta) := \pi(\mathrm{d}\theta, \mathbf{U}, \mathbf{V})$, 'exact marginal posterior'
(iii) $\eta(1) = 0 \Longrightarrow \zeta(1) = 0$ on $\mathbf{T} \times \mathbf{U} \times \mathbf{V}$. '(IS estimator) support cond.'

If Assumption 4 holds, for $f \in L^1(\nu)$ we define the following functions of $(\theta, u, v) = (\Theta_k, U_k, V_k)$, where $V_k = (M_k, Z_k^{(1:M_k)}, \zeta_k^{(1:M_k)})$:

$$\zeta_k(f) := \sum_{i=1}^{M_k} \zeta_k^{(i)} f(\Theta_k, Z_k^{(i)}), \quad \xi_k(f) := \frac{\zeta_k(f)}{\eta_k(1)}, \quad \hat{\zeta}_k(f) := \frac{\zeta_k(f)}{\zeta_k(1)}. \tag{12}$$

We define the following subsets $\mathcal{L}_\pi^p(\nu) \subset L^p(\nu)$, $p = 1$ or 2, by

$$\mathcal{L}_\pi^1(\nu) := \{f \in L^1(\nu) : \pi(\hat{\zeta}(f)) = \nu(f) \text{ and } \pi(\hat{\zeta}(|f|)) < \infty\},$$
$$\mathcal{L}_\pi^2(\nu) := \{g \in \mathcal{L}_\pi^1(\nu) : g^2 \in \mathcal{L}_\pi^1(\nu)\}.$$

Remark 8. Regarding Assumption 4 and the above definitions:

(i) If $f \in L^1(\nu)$ satisfies $f(\theta, \cdot) = f(\theta)$, then $f \in \mathcal{L}_\pi^1(\nu)$. In many settings, e.g. SSMs where V_k is constructed from SMC as part of a particle MCMC, $\mathcal{L}_\pi^1(\nu)$ may be much larger, or all of $L^1(\nu)$ [cf. 65, Cor. 21].
(ii) Support condition (iii) holds quite generally, e.g. if $\eta(1) > 0$. In a latent variable model, where, given θ, $\eta(1)$ is an unbiased estimator for an approx. marginal posterior $\mathrm{pr}(\theta)L^{(U)}(\theta) \propto \dot{\mu}(\theta)$, this can be achieved by inflating the likelihood $L^{(U)}$ by a constant $\epsilon > 0$: $L^{(U)}(\theta) \mapsto L^{(U)}(\theta) + \epsilon$ [cf. 65, Prop. 17 and Rem. 18], renormalising μ accordingly.

The following concerns a PM type scheme targeting ν directly [cf. 6].

Proposition 9. *Suppose a Markov chain $(\Theta_k, U_k, V_k)_{k \geq 1}$ is π-reversible Harris ergodic, where Assumption 4 holds. Then, for all $f \in \mathcal{L}_\pi^1(\nu)$,*

$$E_n^{PM}(f) := \frac{1}{n} \sum_{k=1}^{n} \hat{\zeta}_k(f) \xrightarrow[a.s.]{n \to \infty} \nu(f). \tag{13}$$

Proof. Follows by Harris ergodicity, as $\pi(\hat{\zeta}(f)) = \nu(f)$, $f \in \mathcal{L}_\pi^1(\nu)$. ■

Consider now an IS scheme (Algorithm 1) as in [65]. Compared to [65], we additionally assume μ-reversibility of the base chain and nonnegativity of the estimators $\zeta^{(i)} \geq 0$. This is done to facilitate comparison with the previous PM type scheme corresponding to PM and DA algorithms, which are π-reversible and require $\zeta^{(i)} \geq 0$, as $\zeta(1)$ is present in their acceptance ratio (cf. §6). If Assumption

Algorithm 1 (Importance sampling scheme). Suppose Assumption 4 holds.

(Phase 1) Let $(\Theta_k, U_k)_{k\geq 1}$ be a μ-reversible Harris ergodic Markov chain.
(Phase 2) For each $k \geq 1$, let V_k be drawn as follows, for the IS0 and ISJ cases:

(IS0) $V_k \sim Q^{(V)}_{\Theta_k U_k}(\cdot)$. For $f \in \mathcal{L}^1_\pi(\nu)$, we define

$$\mathbf{m}_f(\theta, u) := \mathbf{E}[\xi_k(f)|\Theta_k = \theta, U_k = u]. \tag{14}$$

(ISJ) Form a jump chain $(\tilde{\Theta}_k, \tilde{U}_k, \tilde{N}_k)_{k\geq 1}$, and draw V_k from some kernel $V_k \sim Q^{(V|N)}_{\tilde{\Theta}_k \tilde{U}_k \tilde{N}_k}(\cdot)$ from $\mathbf{T} \times \mathbf{U} \times \mathbb{N}$ to $\mathbf{V}$ such that

$$\mathbf{E}[\xi_k(f)|\tilde{\Theta}_k = \theta, \tilde{U}_k = u, \tilde{N}_k = n] = \mathbf{m}_f(\theta, u)$$

for all $n \in \mathbb{N}$ and $f \in \mathcal{L}^1_\pi(\nu)$.

4 (PM kernels) holds, then for all $f \in \mathcal{L}^1_\pi(\nu)$,

$$\mu(\mathbf{m}_f) = \frac{1}{c_\eta} \int \mathrm{d}\theta Q^{(U)}_\theta(\mathrm{d}u)\eta(1)Q^{(V)}_{\theta u}(\mathrm{d}v)\frac{\zeta(f)}{\eta(1)} = c_\xi \nu(f)$$

where $c_\xi := c_\zeta / c_\eta$, and $\mathbf{m}_f$ is defined in (14). This motivates the following consistency result, an instance of [65, Prop. 15] for the $\mathsf{N}_k = 1$ case (IS0) and [65, Thm. 11] for the $\mathsf{N}_k = \tilde{N}_k$ case (ISJ).

Proposition 10. *Under Algorithm 1, for all $f \in \mathcal{L}^1_\pi(\nu)$,*

$$E^{\mathrm{IS}}_n(f) := \frac{\sum_{k=1}^n \mathsf{N}_k \xi_k(f)}{\sum_{k=1}^n \mathsf{N}_k \xi_k(1)} \xrightarrow[a.s.]{n\to\infty} \nu(f). \tag{15}$$

Remark 11. In the ISJ case, allowing for dependence on $\tilde{N}_k$ when drawing V_k in Algorithm 1 allows for variance reduction of $\xi_k(f)$ and hence of the resultant estimator (15) (cf. Proposition 19), by using larger M_k when $\tilde{N}_k$ is large. For example, M_k could correspond to the number of independent samples drawn from an instrumental, or, more generally, to the number of particles used in the SMC, if this is how V_k is generated.

5.2. **A Peskun type ordering for importance sampling schemes.** Under Assumption 5 below, the IS estimator $E^{\mathrm{IS}}_n(f)$ (15) satisfies a CLT

$$\sqrt{n}[E^{\mathrm{IS}}_n(f) - \nu(f)] \xrightarrow{n\to\infty} \mathcal{N}\big(0, \mathbb{V}^{\mathrm{IS}}_f\big), \qquad \text{in distribution.} \tag{16}$$

See [65] or Proposition 19 of Appendix A, with a formula for $\mathbb{V}^{\mathrm{IS}}_f$. In analogy with Definition 2 and (3), we refer to $\mathbb{V}^{\mathrm{IS}}_f$ as the *IS asymptotic variance*.

Assumption 5 (Importance sampling CLT)**.** Suppose Algorithm 1 (IS scheme) and that $(\Theta_k, \mathsf{U}_k, \mathsf{N}_k)_{k\geq 1}$ is aperiodic. Let $f \in \mathcal{L}^2_\pi(\nu)$ be a function such that $\mathrm{var}(K, \mathbf{m}_f) < \infty$, where $\mathbf{m}_f$ is defined in (14), and $\mathsf{v}_{\bar{f}}$ by

(IS0) $v_{\bar{f}}(\theta, u) := \mathrm{var}\big(\xi_k(\bar{f})|\Theta_k = \theta, U_k = u\big)$,
(ISJ) $\tilde{v}_{\bar{f}}(\theta, u) := \mathbf{E}[\tilde{N}_k^2 \mathrm{var}\big(\xi_k(\bar{f})|\tilde{\Theta}_k = \theta, \tilde{U}_k = u, \tilde{N}_k\big)|\tilde{\Theta}_k = \theta, \tilde{U}_k = u]$,

satisfies $\mu\big(\mathsf{av}_{\bar{f}}\big) < \infty$.

Let us denote the kernel and measure of the IS0 corrected chain of Algorithm 1 by $(\bar{K}, \bar{\mu})$ on the space $\mathbf{X} = (\mathbf{T} \times \mathbf{U}) \times \mathbf{V}$, where,

$$\bar{K}_{\theta uv}(\mathrm{d}\theta', \mathrm{d}u', \mathrm{d}v') := K_{\theta u}(\mathrm{d}\theta', \mathrm{d}u') Q^{(V)}_{\theta' u'}(\mathrm{d}v')$$
$$\bar{\mu}(\mathrm{d}\theta, \mathrm{d}u, \mathrm{d}v). = \mu(\mathrm{d}\theta, \mathrm{d}u) Q^{(V)}_{\theta u}(\mathrm{d}v). \tag{17}$$

Note that $\bar{K} = K^{(V)}$ is an augmented kernel (Definition 3).

With definitions as in Assumption 5, we define a 'difference' constant $\mathsf{D}_{\bar{f}}$, for the IS0 and ISJ cases, respectively, by $D_{\bar{f}} := 0$ and

$$\tilde{D}_{\bar{f}} := \mu(a) c_\xi^{-2} \mu(a\tilde{v}_{\bar{f}} - v_{\bar{f}}).$$

Theorem 12. *Suppose Algorithm 1 (IS scheme) and Assumption 5 (IS CLT) hold.*

(i) If $(\bar{\mu}, \pi, w, \bar{K}, L, \underline{c}, \overline{c})$ satisfies Assumption 2 on $\mathbf{X}$, then

$$\mathbb{V}^{IS}_f + \mu(\mathsf{a}) \mathrm{var}_{\bar{\mu}}\big(w\hat{\zeta}(\bar{f})\big) \le \overline{c}\, \mu(\mathsf{a}) \big\{ \mathrm{var}\big(L, \hat{\zeta}(f)\big) + \mathrm{var}_\pi\big(\hat{\zeta}(f)\big) \big\} + \mathsf{D}_{\bar{f}}$$
$$\mathbb{V}^{IS}_f + \mu(\mathsf{a}) \mathrm{var}_{\bar{\mu}}\big(w\hat{\zeta}(\bar{f})\big) \ge \underline{c}\, \mu(\mathsf{a}) \big\{ \mathrm{var}\big(L, \hat{\zeta}(f)\big) + \mathrm{var}_\pi\big(\hat{\zeta}(f)\big) \big\} + \mathsf{D}_{\bar{f}}.$$

(ii) If $(\bar{\mu}, \pi, w, \bar{K}, L, \underline{c}, \overline{c})$ satisfies Assumption 3 on $\mathbf{X}$, then

$$\mathbb{V}^{IS}_f \le \overline{c}\, \mu(\mathsf{a}) \big\{ \mathrm{var}\big(L, \hat{\zeta}(f)\big) + \mathrm{var}_\pi\big(\hat{\zeta}(f)\big) \big\} + (1 + 2\mathscr{N}_K) \mu(\mathsf{a}) \mathrm{var}_{\bar{\mu}}\big(w\hat{\zeta}(\bar{f})\big) + \mathsf{D}_{\bar{f}}$$

where $\mathscr{N}_K := 0$ if K is positive, and $\mathscr{N}_K := 1$ if not.

Remark 13. Regarding Theorem 12, whose proof is in Appendix A:

(i) Note that $0 \le \mu(\mathsf{a}) \le 1$, with a as in §4.2, and that $w = c_\xi^{-1} \xi(1)$ and $w^* = c_\xi^{-1} \mathbf{m}_1$, with $\mathbf{m}_f(\theta, u)$ defined in (14).

(ii) As a trivialisation, when $\eta(\Theta_k, U_k) := \eta(1) = \dot{\mu}(\Theta_k)$ a.s., $\mathbf{Z} = \{0\}$, and $\xi_k(f) = w_u(\Theta_k) f(\Theta_k)$ a.s., we obtain a Peskun type ordering for SNIS (11). Here, the simplifications are $\bar{K} \leftrightarrow K$, $\hat{\zeta}(\bar{f}) \leftrightarrow \bar{f}$ and $\xi(\bar{f}) \leftrightarrow c_\xi w \bar{f}$.

6. Pseudomarginal and delayed acceptance MCMC

We define PM and DA type algorithms in the setting of the auxiliary variable framework of §5, where PM could be the 'particle marginal Metropolis-Hastings' [2]; a DA type variant of this algorithm has been implemented e.g. in [30, 52, 65]. After defining the corresponding kernels, we then compare the asymptotic variances of PM/DA with IS (Theorem 14).

6.1. **Algorithms.** Let $q_\theta(\mathrm{d}\theta') = q_\theta(\theta')\mathrm{d}\theta'$ be a proposal kernel on $\mathbf{T}$. Assume the setup of Assumption 4 (recall that $\eta(1) \ge 0$ and $\zeta(1) \ge 0$). Whenever the denominators are not zero we define the following 'acceptance ratios' for $x, x' \in \mathbf{X} := \mathbf{T} \times \mathbf{U} \times \mathbf{V}$, where $x = (\theta, u, v)$,

$$r^{(U)}(x, x') := \frac{\eta'(1) q_{\theta'}(\theta)}{\eta(1) q_\theta(\theta')}, \quad \text{and} \quad r^{(V)}(x, x') := \frac{\zeta'(1) q_{\theta'}(\theta)}{\zeta(1) q_\theta(\theta')}. \tag{18}$$

Consider Algorithm 2 ('PM parent,' following the terminology of [57]), Algorithm 3 ('DA0'), and Algorithm 4 ('DA1'), with transition kernels given later and which are π-invariant [cf. 2, 6, 10]. Under Assumption 4 (PM kernels) and

Algorithm 2 (Pseudomarginal parent). Suppose Assumption 4 (PM kernels) holds. Initialise $X_0 \in \mathbf{X}$ with $\zeta_0(1) > 0$. For $k = 1, \dots n$, do:

(1) Draw $\Theta'_k \sim q_{\Theta_{k-1}}(\cdot)$ and $U'_k \sim Q^{(U)}_{\Theta'_k}(\cdot)$ and $V'_k \sim Q^{(V)}_{\Theta'_k U'_k}(\cdot)$. With probability $\min\left\{1, r^{(V)}(X_{k-1}, X'_k)\right\}$ accept X'_k; otherwise, reject.

Algorithm 3 (Delayed acceptance ('DA0')). Suppose Assumption 4 (PM kernels) holds, and K is a μ-proposal-rejection kernel of the form (20). Initialise $X_0 \in \mathbf{X}$ with $\zeta_0(1) > 0$. For $k = 1, \dots, n$, do:

(1) Draw $\Theta'_k \sim q_{\Theta_{k-1}}(\cdot)$. Construct $U'_k \sim Q^{(U)}_{\Theta'_k}(\cdot)$. With probability $\alpha(\Theta_{k-1}, U_{k-1}; \Theta'_k, U'_k)$, proceed to step (2). Otherwise, reject.
(2) Construct $V'_k \sim Q^{(V)}_{\Theta'_k, U'_k}(\cdot)$. With probability $\min\left\{1, \xi'_k(1)/\xi_k(1)\right\}$, accept (Θ'_k, U'_k, V'_k); otherwise, reject.

Algorithm 4 (Delayed acceptance ('DA1')). Suppose Assumption 4 (PM kernels) holds. Initialise $X_0 \in \mathbf{X}$ with $\zeta_0(1) > 0$. For $k = 1, \dots, n$, do:

(1) Draw $(\Theta'_k, U'_k) \sim K_{\Theta_{k-1}, U_{k-1}}(\cdot)$.
(2) Construct $V'_k \sim Q^{(V)}_{\Theta'_k, U'_k}(\cdot)$. With probability $\min\left\{1, \xi'_k(1)/\xi_k(1)\right\}$, accept (Θ'_k, U'_k, V'_k); otherwise, reject.

the assumption that the resultant chains are π-Harris ergodic, by construction Algorithms (2-4) produce output as in Proposition 9 (PM type scheme). In PM parent (Algorithm 2) and DA1 (Algorithm 4), the computationally expensive V_k-variable is drawn whenever U_k is drawn. This is the essential difference with DA0 (Algorithm 3). The separation of sampling steps can substantially reduce computational cost in DA0 [cf. 17], even though the asymptotic variance of DA0 is more than PM parent in the case K is the approximate PM kernel (22) [cf. 10], and more than DA1 in the case K is a μ-proposal-rejection chain (see Section 6.2 below, these are the cases when the chains are comparable).

6.2. **Kernels.** Let K be the transition kernel of a μ-reversible Harris ergodic IS0 base chain $(\Theta_k, U_k)_{k\geq 1}$, with definitions as in Assumption 4 (PM kernels). The *DA1 correction* of K is the π-reversible kernel K^{DA1} corresponding to Algorithm 4, given by,

$$K^{\mathrm{DA1}}_{\theta uv}(\mathrm{d}\theta', \mathrm{d}u', \mathrm{d}v') = K_{\theta u}(\mathrm{d}\theta', \mathrm{d}u') Q^{(V)}_{\theta' u'}(\mathrm{d}v') \min\left\{1, \xi'(1)/\xi(1)\right\} + [1 - \alpha_{\mathrm{DA1}}(\theta, u, v)] \delta_{\theta uv}(\mathrm{d}\theta', \mathrm{d}u', \mathrm{d}v'), \tag{19}$$

where $\alpha_{\mathrm{DA1}}(\theta, u, v) := \int K_{\theta u}(\mathrm{d}\theta', \mathrm{d}u') Q^{(V)}_{\theta' u'}(\mathrm{d}v') \min\left\{1, \xi'(1)/\xi(1)\right\}$.

If K is in particular a μ-proposal-rejection kernel (see §3.2.1) of the form

$$K_{\theta u}(\mathrm{d}\theta', \mathrm{d}u') = q_\theta(\mathrm{d}\theta') Q^{(U)}_{\theta'}(\mathrm{d}u') \alpha(\theta, u; \theta', u') + \left(1 - \int q_\theta(\mathrm{d}\theta'') Q^{(U)}_{\theta''}(\mathrm{d}u'') \alpha(\theta, u; \theta'', u'')\right) \delta_{\theta, u}(\mathrm{d}\theta', \mathrm{d}u'), \tag{20}$$

then *DA0 correction* of K is

$$K_x^{\mathrm{DA0}}(\mathrm{d}x') = q_\theta(\mathrm{d}\theta')Q_{\theta'}^{(U)}(\mathrm{d}u')\alpha(\theta,u;\theta',u')Q_{\theta'u'}^{(V)}(\mathrm{d}v')\min\left\{1,\xi'(1)/\xi(1)\right\} + [1-\alpha_{\mathrm{DA0}}(x)]\delta_{\theta uv}(\mathrm{d}\theta',\mathrm{d}u',\mathrm{d}v'), \tag{21}$$

where $\alpha_{\mathrm{DA0}}(x) = K_x^{\mathrm{DA0}}(\mathbf{X}\backslash\{x\})$, and $\mathbf{X} := \mathbf{T}\times\mathbf{U}\times\mathbf{V}$, $x\in\mathbf{X}$, $x := (\theta,u,v)$.

Decreasing the variability of $\xi'(1) = \zeta'(1)/\eta'(1)$ by coupling the u' and v' variables can lead to improved mixing of (19), and is similar in idea to recently proposed 'correlated PM' [19] and 'MHAAR' [3] chains. The mere requirement of reversibility allows the kernel K to be taken to be approximate versions of the two chains listed above, or an approximate DA or 'multi-stage DA' [10]. Regardless, the most straightforward choice for K is the (approximate) PM kernel targeting μ with proposal q, given by,

$$K_{\theta u}(\mathrm{d}\theta',\mathrm{d}u') = q_\theta(\mathrm{d}\theta')Q_{\theta'}^{(U)}(\mathrm{d}u')\min\left\{1,r^{(U)}(x,x')\right\} + [1-\alpha(\theta,u)]\delta_{\theta u}(\mathrm{d}\theta',\mathrm{d}u'), \tag{22}$$

where $\alpha(\theta,u) := \int q_\theta(\mathrm{d}\theta')Q_{\theta'}^{(U)}(\mathrm{d}u')\min\left\{1,r^{(U)}(x,x')\right\}$.

We remark that by the covariance (or Peskun) ordering, we have $\mathrm{var}(K^{\mathrm{DA1}},f)\leq \mathrm{var}(K^{\mathrm{DA0}},f)$ for all $f\in L^2(\pi)$, where K is a μ-proposal-rejection kernel. However, for the reason discussed in Section 6.1, DA0 is likely more computationally efficient than DA1 in practice.

We define the PM parent kernel P of K^{DA1} to be given by

$$P_{\theta uv}(\mathrm{d}\theta',\mathrm{d}u',\mathrm{d}v') = q_\theta(\mathrm{d}\theta')Q_{\theta'}^{(U)}(\mathrm{d}u')Q_{\theta'u'}^{(V)}(\mathrm{d}v')\min\left\{1,r^{(V)}(x,x')\right\} + [1-\alpha_{\mathrm{PMP}}(\theta,v)]\delta_{\theta uv}(\mathrm{d}\theta',\mathrm{d}u',\mathrm{d}v'), \tag{23}$$

where $\alpha_{\mathrm{PMP}}(\theta,v) := \int q_\theta(\mathrm{d}\theta')Q_{\theta'}^{(U)}(\mathrm{d}u')Q_{\theta'u'}^{(V)}(\mathrm{d}v')\min\left\{1,r^{(V)}(x,x')\right\}$.

We define a probability kernel from $\mathbf{T}$ to $\mathbf{V}$ by

$$\hat{Q}_\theta^{(V)}(\mathrm{d}v) := \int_{\mathbf{U}} Q_\theta^{(U)}(\mathrm{d}u)Q_{\theta u}^{(V)}(\mathrm{d}v) \tag{24}$$

We then define the following PM kernel with proposal q targeting π,

$$M_{\theta v}(\mathrm{d}\theta',\mathrm{d}v') = q_\theta(\mathrm{d}\theta')\hat{Q}_{\theta'}^{(V)}(\mathrm{d}v')\min\left\{1,r^{(V)}(x,x')\right\} + [1-\alpha_{\mathrm{PM}}(\theta,v)]\delta_{\theta v}(\mathrm{d}\theta',\mathrm{d}v'), \tag{25}$$

where $\alpha_{\mathrm{PM}}(\theta,v) := \int q_\theta(\mathrm{d}\theta')\hat{Q}_{\theta'}^{(V)}(\mathrm{d}v')\min\left\{1,r^{(V)}(x,x')\right\}$.

When U_k and V_k are independent given θ, i.e.

$$Q_{\theta u}^{(V)}(\mathrm{d}v) = Q_\theta^{(V)}(\mathrm{d}v), \tag{26}$$

then M (25) is the standard PM with proposal q and target π, since,

$$\hat{Q}_\theta^{(V)}(\mathrm{d}v) = Q_\theta^{(V)}(\mathrm{d}v).$$

6.3. **Comparison with importance sampling correction.**

Theorem 14. *Suppose Assumption 4 (PM kernels) holds, and that one of the following conditions for pairs of kernels holds:*

(I) $L = K^{DA0}$ *is DA0 correction* (21)*, and* K *is* μ*-proposal-rejection* (20)*,*
(II) $L = K^{DA1}$ *is DA1 correction* (19)*, and* K *is* μ*-reversible,*
(III) $L = P$ *is the PM parent* (23)*, and* K *is the approx. PM* (22)*, or*

(IV) $L = M$ is the PM kernel (25)*, and K is the approx. PM* (22)*.*

Assume K and L are Harris ergodic, and a function $f \in \mathcal{L}^2_\pi(\nu)$ is such that Assumption 5 (IS CLT) holds. The following statements hold:

(i) The IS asymptotic variance (16) *satisfies, with $\underline{c} := \bar{\mu}$-ess inf w,*

$$\mathbb{V}_f^{IS} + \mu(\mathsf{a})\mathrm{var}_{\bar{\mu}}\big(w\hat{\zeta}(\bar{f})\big) \le \mu(\mathsf{a})\,\|w\|_\infty \left\{\mathrm{var}\big(L, \hat{\zeta}(f)\big) + \mathrm{var}_\pi\big(\hat{\zeta}(f)\big)\right\} + \mathsf{D}_{\bar{f}}$$

$$\mathbb{V}_f^{IS} + \mu(\mathsf{a})\mathrm{var}_{\bar{\mu}}\big(w\hat{\zeta}(\bar{f})\big) \ge \mu(\mathsf{a}) \cdot \underline{c} \cdot \left\{\mathrm{var}\big(L, \hat{\zeta}(f)\big) + \mathrm{var}_\pi\big(\hat{\zeta}(f)\big)\right\} + \mathsf{D}_{\bar{f}}.$$

(ii) With $\mathscr{N}_K := 0$ if K is positive and $\mathscr{N}_K := 1$ if not, the following holds:

$$\mathbb{V}_f^{IS} \le \mu(\mathsf{a})\,\|w^*\|_\infty \left\{\mathrm{var}\big(L, \hat{\zeta}(f)\big) + \mathrm{var}_\pi\big(\hat{\zeta}(f)\big)\right\} + (1 + 2\mathscr{N}_K)\mu(\mathsf{a})\mathrm{var}_{\bar{\mu}}\big(w\hat{\zeta}(\bar{f})\big) + \mathsf{D}_{\bar{f}}.$$

See Remark 13(i) for w and w^*. See Appendix B for the proof of Theorem 14, which follows from Theorem 12, after bounding the Dirichlet forms.

7. Discussion

In this section we discuss various issues of stability (§7.1), computational efficiency (§7.2), and approximation strategies (§7.3).

7.1. **Importance sampling weight and stability considerations.** A necessary condition for a successful implementation of an IS or PM scheme is a simple support condition, Assumption 4(iii), that can often be easily ensured by Remark 8(ii). On the other hand, Theorem 14 depends on a uniform bound on the marginal weight $w^* \propto \mathbf{m}_1$, with $\mathbf{m}_f(\theta, u)$ as in (14). This bound is much weaker than a bound on w, and can often be ensured. For example, assuming that $\eta(1)\mathbf{m}_1$ is bounded, one can often inflate $\eta(1)$ as in Remark 8(ii) to obtain an uniform bound on w^*. Other techniques may be applicable if a bounded w^* is particularly desired, such as a combination of cutoff functions, approximations, or tempering [cf. 48, 65].

When considering a PM/DA implementation, the issue of boundedness of the full weight $w \propto \zeta(1)/\eta(1)$ takes particular importance, more so than in the case with IS. This is because PM and DA are more liable to be poorly mixing, while IS is less affected by noisy estimators, as described below.

We claim that if $\zeta(1)$ is not bounded, then PM and K^{DA0}, with K as in (22), are not geometrically ergodic. This is [6, Thm. 8] for PM chains. To prove that result for PM chains, or in particular for the PM parent chain (23), [6] show that for all $\epsilon > 0$,

$$\nu\big(\mathbf{1}\{\alpha_{\mathrm{PMP}} \le \epsilon\}\big) > 0. \tag{27}$$

By [55, Thm. 5.1], one concludes that the PM parent is not geometrically ergodic [6]. Moreover, with K as in (22) and $L = K^{\mathrm{DA0}}$ as in (19), from

$$\min\{1, r^{(U)}(x, x')\}\min\{1, w(x')/w(x)\} \le \min\{1, r^{(V)}(x, x')\}, \tag{28}$$

it follows that $\alpha_{\mathrm{DA0}}(x) \le \alpha_{\mathrm{PMP}}(x)$. By (27), one concludes that K^{DA0} also is not geometrically ergodic.

On the other hand, the IS chain may converge fine, even in the case of unbounded $\zeta(1)$. For example, if K is a random walk Metropolis-Hastings chain, then K is geometrically ergodic essentially if μ has exponential or lighter tails and

a certain contour regularity condition holds [32, 55], where we have said nothing about the exact level estimator $\zeta(1)$. We then apply Lemma 22(v), which says that whenever K is geometrically ergodic then so is $\bar{K}$, to conclude that the IS chain is geometrically ergodic, even in the case of unbounded $\zeta(1)$. This may be beneficial if adaptation is used [5, 7, 54].

Of course, high variability affects also the IS estimator, but we believe this noise to be a smaller issue in IS, as the noise is in the IS output estimator rather than in the acceptance ratio as in PM/DA. This can make a significant difference in the evolution and ergodicity of the chains, as described above.

7.2. **Computational aspects of the importance sampling correction.** The finite-size perturbation bounds for the asymptotic variance of IS versus PM/DA (Corollary 4 or Theorem 14) show that IS can not do much worse than PM/DA in terms of statistical efficiency. On the other hand, the flexibility of the IS implementation allows for the use of many potentially substantial computational efficiency enhancements [65], which we briefly mention.

Thinning, where only every kth value of a chain is kept [cf. 47], may be applied to the IS base chain, which may be e.g. adaptive [5, 7, 54], approximated [33, 46], correlated [19], 'MHAAR' [3], or nonreversible [63]. The thinning can be performed before any calculations of weights. The weights also need not be calculated in the burn-in phase. The use of a jump chain estimator can further decrease the number of necessary weight calculations, and shows the strength of IS in relation to PM/DA using 'early rejection' [60], which is computational efficiency enhancement for PM/DA applicable when the posterior is factorisable and the subposteriors are monotonically decreasing [cf. 60, §4], but may involve expensive calculations for all innovations, unlike ISJ. Real-valued IS type weight estimators also allow for multilevel Monte Carlo [cf. 20]. Also, the IS correction, which is based on independent 'post-processing' correction of the approximate chain output, allows for separation of approximate and exact phases, leading to easy process management, output analysis, and parallelisation.

7.3. **Finding an approximation.** A necessity of the IS approach compared to a direct PM approach is finding a suitable approximate Markov chain; see [33, 58, 65] for suggestions. We remark that this problem simplifies when there is a clear grading of approximate models, for then one can use a PM chain targeting a coarse-level distribution and then IS correct to the fine-level. The grading could be based on the tolerance size in approximate Bayesian computation as in [51] or on the discretisation size of a discretely observed diffusion as in [65], who both show performance gains over a direct approach.

The grading could also come from the order of the Taylor [17] or Fourier [18, 24, 58] series approximation needed for the posterior density, a multilevel [20], multiscale [24], or dimension reduction [18] framework, the amount of subsampled data in a big data setting [10, 52], the size of introduced noise in a perturbed problem strategy [11], the subfactor length of a factorisable likelihood [60], or the number of nearest neighbours used in a local approximation [56]. The cited works are just a few of the many that use the DA implementation, which may alternatively be run as an IS implementation by a simple rearrangement of the

algorithm. The two-phase IS method may lead to performance gains over a direct MCMC, especially with massive parallelisation [35].

Acknowledgments

The authors have been supported by the Academy of Finland (grants 274740, 284513 and 312605). JF thanks also the organisers of the 2017 SMC course and workshop in Uppsala for hosting a great event.

Appendix A. Proofs for the Peskun type orderings

A.1. **Subprobability kernels.** Let K be a μ-reversible Markov kernel on $\mathbf{X}$. For all $\lambda \in (0,1]$, λK is a *subprobability kernel*: $\lambda K(x,\mathbf{X}) \leq 1$ for all $x \in \mathbf{X}$. The *Dirichlet form* $\mathcal{E}_{\lambda K}(f)$ of the subprobability kernel λK is

$$\mathcal{E}_{\lambda K}(f) := \langle f, (1-\lambda K) f \rangle_\mu = \lambda \mathcal{E}_K(f) + (1-\lambda) \|f\|_\mu^2, \tag{29}$$

defined for $f \in L^2(\mu)$. For $f \in L^2_0(\mu)$, if $(1-K)^{-1} f$ exists in $L^2(\mu)$, then by (2), $\operatorname{var}(K,f) = 2\langle f, (1-K)^{-1} f\rangle_\mu - \mu(f^2)$ [cf. 9]. Following [9, 61], we then (formally) extend Definition 2 of the asymptotic variance to subprobability kernels: for $\lambda \in (0,1)$, the operator $(1-\lambda K)$ is always invertible, and we define

$$\operatorname{var}(\lambda K, f) := 2\left\langle f, (1-\lambda K)^{-1} f\right\rangle_\mu - \mu(f^2). \tag{30}$$

Moreover, (1) and (29) imply for $\lambda \in (0,1]$ that $1-\lambda K$ is a positive operator, i.e. $\mathcal{E}_{\lambda K}(f) \geq 0$ for all $f \in L^2(\mu)$. By a result attributed to Bellman [14, Eq. 14], for positive self-adjoint operators, and used e.g. in [1, 9, 15, 44, 43], we have another asymptotic variance representation: for all $\lambda \in (0,1)$ and $f \in L^2_0(\mu)$,

$$\operatorname{var}(\lambda K, f) = 2 \sup_{g \in L^2(\mu)} \left\{ 2\langle f, g\rangle_\mu - \mathcal{E}_{\lambda K}(g) \right\} - \mu(f^2). \tag{31}$$

Here, the supremum is attained with $g := (1-\lambda K)^{-1} f$, in which case (31) simplifies to (30). For $\lambda \in (0,1)$, equalities (30–31) hold and are finite for any $f \in L^2_0(\mu)$. The function $\lambda \mapsto \operatorname{var}(\lambda K, f)$ has a limit as $\lambda \uparrow 1$ on the extended real numbers $[0,\infty]$, and $\operatorname{var}(K,f)$ equals this limit [61].

A.2. **Normalised importance sampling ordering.** We set

$$\mathscr{N}_K := -\inf_{\mu(g)=0, \mu(g^2)=1} \langle g, Kg\rangle_\mu \tag{32}$$

for a μ-reversible kernel K, so that the *left spectral gap* of K is $1-\mathscr{N}_K$ [cf. 9].

Lemma 15. *Suppose $(\mu,\nu,w,K,L,\underline{c},\overline{c})$ satisfies Assumption 3 on $\mathbf{X} := \mathbf{T}\times\mathbf{Y}$. Let $\varphi \in L^2_0(\nu)$ be such that $w\varphi \in L^2(\mu)$. Define $u_\lambda := (1-\lambda K)^{-1}(w\varphi)$ and $\breve{u}_\lambda := u_\lambda - w\varphi$, in $L^2(\mu)$ for all $\lambda \in (0,1)$. The following hold:*

(i) If $u_\lambda(\theta,y) = u_\lambda(\theta)$, $\lambda \in (0,1)$, then (4) holds.
(ii) If $\breve{u}_\lambda(\theta,y) = \breve{u}_\lambda(\theta)$, $\lambda \in (0,1)$, then (10) holds, with $\mathscr{N}_K$ as in (32).

Proof. Note that $L^2(\mu^*) \subset L^2(\nu^*)$ by Assumption 3(b). For $g \in L^2(\mu^*)$,

$$\mathcal{E}_{\lambda L}(g) = \lambda \mathcal{E}_L(g) + (1-\lambda)\nu^*(g^2) \leq \overline{c}\lambda \mathcal{E}_K(g) + (1-\lambda)\nu^*(g^2),$$

by Assumption 3(a). From the above first equality, now for λK and μ^*,

$$\begin{aligned}\mathcal{E}_{\lambda L}(g) &\le \overline{c}\big[\mathcal{E}_{\lambda K}(g) - (1-\lambda)\mu^*(g^2)\big] + (1-\lambda)\nu^*(g^2)\\ &= \overline{c}\mathcal{E}_{\lambda K}(g) - (1-\lambda)\mu^*\big(g^2[\overline{c} - w^*]\big) \le \overline{c}\mathcal{E}_{\lambda K}(g),\end{aligned} \tag{33}$$

by Assumption 3(b). Since $1-\lambda K$ is self-adjoint on $L^2(\mu)$, we also note that

$$\mathcal{E}_{\lambda K}(\check{u}_\lambda) = \mathcal{E}_{\lambda K}(u_\lambda - w\varphi) = \mathcal{E}_{\lambda K}(u_\lambda) + \mathcal{E}_{\lambda K}(w\varphi) - 2\left\|w\varphi\right\|_\mu^2,$$

as $\langle v_\lambda, (1-\lambda K)w\varphi\rangle_\mu = \left\|w\varphi\right\|_\mu^2$. Regardless of $\lambda \in (0,1)$, $1-\lambda K$ has support of its spectral measure contained in $[0, 1+\mathscr{N}_K]$ (cf. Remark 17(ii) below). Hence, $\mathcal{E}_{\lambda K}(w\varphi) \le (1+\mathscr{N}_K)\left\|w\varphi\right\|_\mu^2$, so

$$\mathcal{E}_{\lambda K}(\check{u}_\lambda) \le \mathcal{E}_{\lambda K}(u_\lambda) + (\mathscr{N}_K - 1)\left\|w\varphi\right\|_\mu^2. \tag{34}$$

We now compare the asymptotic variances. By (30),

$$LS := \operatorname{var}(\lambda K, w\varphi) + \left\|w\varphi\right\|_\mu^2 = 2\big[2\left\langle w\varphi, u_\lambda\right\rangle_\mu - \mathcal{E}_{\lambda K}(u_\lambda)\big].$$

With $\psi := u_\lambda$ for (i), and with $\psi := \check{u}_\lambda$ for (ii) using (34),

$$LS \le 2\big[2\left\langle w\varphi, \psi\right\rangle_\mu - \mathcal{E}_{\lambda K}(\psi)\big] + E_\psi,$$

where $E_\psi := 0$ if $\psi = u_\lambda$ and $E_\psi := 2(1+\mathscr{N}_K)\left\|w\varphi\right\|_\mu^2$ if $\psi = \check{u}_\lambda$. Hence,

$$LS \le 2\big[2\left\langle \varphi, \psi\right\rangle_\nu - \mathcal{E}_{\lambda K}(\psi)\big] + E_\psi \le 2\big[2\left\langle \varphi, \psi\right\rangle_\nu - (\overline{c})^{-1}\mathcal{E}_{\lambda L}(\psi)\big] + E_\psi,$$

where we have used (33). Since $\psi \in L^2(\mu^*) \subset L^2(\nu)$,

$$\begin{aligned}LS &\le \frac{1}{\overline{c}}\Big(2\sup_{g\in L^2(\nu)}\big\{2\left\langle \overline{c}\varphi, g\right\rangle_\nu - \mathcal{E}_{\lambda L}(g)\big\} - \left\|\overline{c}\varphi\right\|_\nu^2\Big) + \overline{c}\left\|\varphi\right\|_\nu^2 + E_\psi\\ &= \overline{c}\big(\operatorname{var}(\lambda L, \varphi) + \left\|\varphi\right\|_\nu^2\big) + E_\psi,\end{aligned}$$

by (31). We then take the limit $\lambda \uparrow 1$ [61]. Noting that $\left\|w\varphi\right\|_\mu^2 = \operatorname{var}_\mu(w\varphi)$ since $\mu(w\varphi) = \nu(\varphi) = 0$, we conclude. ∎

Lemma 16. *Suppose the assumptions of Lemma 15 hold, where $\overline{c}$ may be also ∞. If $v_\lambda := (1-\lambda L)^{-1}(\varphi)$ satisfies $v_\lambda(\theta, y) = v_\lambda(\theta)$, then* (5) *holds.*

Proof. The lower bound (5) is trivial if $\underline{c} = 0$. Assume $\underline{c} > 0$. Then $\mu \ll \nu$, $w^{-1} \le \underline{c}^{-1}$ (implying $L^2(\nu) \subseteq L^2(\mu)$), and $\mathcal{E}_K(g) \le \underline{c}^{-1}\mathcal{E}_L(g)$ for all $g \in L^2(\nu)$. The result follows by applying Lemma 15(i). ∎

Remark 17. Regarding Lemma 15 and Lemma 16:

(i) The solution v_λ to the Poisson eq. [cf. 42], $(1-\lambda L)g = \varphi$ in $L^2(\mu)$, is also used in [9, Thm. 17] as a lemma for the proof of the convex order criterion Peskun type ordering for PM chains [9, Thm. 10].

(ii) We have $\mathscr{N}_K \in [-1, 1]$ in general, but $\mathscr{N}_K \in [-1, 0]$ if K is positive.

(iii) It is reasonable to use a single constant $\overline{c}$ in Assumptions 3(a–b). If one replaces Assumption 3(b) with $w^* \le \overline{c}'\ \mu^*-a.e.$, then, if $\overline{c}' < \overline{c}$, one obtains the same result after bounding a nonpositive quantity by zero in (33). If $\overline{c}' > \overline{c}$, then one would need to impose the unappealing condition that $\sup_{\lambda\in(0,1)}\left\|u_\lambda\right\|_{\mu^*}^2 < \infty$ and add a positive constant involving this bound to the final results. Anyways, for the the application in this paper, we have $\overline{c} = \overline{c}'$ (Lemma 21).

(iv) Assumption 3(a) can be replaced with the weaker assumption that $\mathcal{E}_L(g) \leq \bar{c}\,\mathcal{E}_K(g)$ for all $g \in \mathcal{G} \subset L^2(\mu^*)$, where $\mathcal{G} := \{u_\lambda : \lambda \in (0,1)\}$ for (i) and $\mathcal{G} := \{\check{u}_\lambda : \lambda \in (0,1)\}$ for (ii).

Lemma 18. *Let K be a μ-reversible chain on $\mathbf{X} = \mathbf{T} \times \mathbf{Y}$. For $h \in L^2(\mu)$ and $\lambda \in (0,1)$, set $h_\lambda := (1-\lambda K)^{-1}h$ and $\check{h}_\lambda := h_\lambda - h$, which are in $L^2(\mu)$.*

(i) If $\mathbf{Y} = \{y_0\}$ is the trivial space, then $h_\lambda(\theta, y) = h_\lambda(\theta)$.
(ii) If K is an augmented kernel, then $\check{h}_\lambda(\theta, y) = \check{h}_\lambda(\theta)$. Moreover, if also $h(\theta, y) = h(\theta)$, then $h_\lambda(\theta, y) = h_\lambda(\theta)$.

Proof. (i) is clear. For (ii), we write the series representation for the inverse of an invertible operator and use Lemma 22(iii), to get that,

$$h_\lambda(\theta, y) = \sum_{n=0}^{\infty} \lambda^n K^n h(\theta, y) = h(\theta, y) + \sum_{n=1}^{\infty} \lambda^n \dot{K}^n (Qh)(\theta).$$

The result then follows. ∎

Proof of Theorem 2. The upper bound (4) follows from Lemma 15(i) and Lemma 18(i), while (5) follows from Lemma 16 and Lemma 18(i). ∎

Proof of Theorem 5. Follows by Lemma 15 and Lemma 18(ii). ∎

A.3. **Importance sampling schemes.** The following CLT, based on Proposition 1, and asymptotic variance formula, are [65, Theorem 7 & 13].

Proposition 19. *Under Assumption 5, the IS estimator* (15) *satisfies the CLT* (16)*, with limiting variance $\mathbb{V}_f^{IS} = \mu(\mathsf{a})\big[\mathrm{var}(K, \mathbf{m}_f) + \mu(\mathsf{av}_{\bar{f}})\big]/c_\xi^2$.*

Proof of Theorem 12. We first note that

$$\xi(f) := \frac{\zeta(f)}{\eta(1)} = \frac{c_\zeta}{c_\eta} \cdot \frac{c_\eta}{c_\zeta} \frac{\zeta(1)}{\eta(1)} \cdot \frac{\zeta(f)}{\zeta(1)} = c_\xi w \hat{\zeta}(f).$$

By Slutsky's Theorem applied to (15) in the IS0 case,

$$\mathbb{V}_f^{IS0} = \mathrm{var}\big(\bar{K}, \xi(f)\big)/c_\xi^2 = \mathrm{var}\big(\bar{K}, w\hat{\zeta}(f)\big).$$

Then (i) follows by Theorem 2, and (ii) by Theorem 5, for the IS0 case. To prove the result for the ISJ case, we first note the relationship

$$\mathbb{V}_f^{ISJ} = \mu(\alpha) c_\xi^{-2} \Big[\mathrm{var}(K, \mathbf{m}_f) + \mu(v_{\bar{f}}) + \mu(\alpha \tilde{v}_{\bar{f}} - v_{\bar{f}})\Big] = \mu(\alpha)\mathbb{V}_f^{IS0} + \tilde{D}_{\bar{f}},$$

from Proposition 19. The result then follows from the IS0 case. ∎

Appendix B. Proofs for main comparison application

Lemma 20. *Let (K, L) be the pair of kernels as in* (II)*,* (I)*, or* (III) *of Theorem 14, where we assume that $(\bar{\mu}, \nu, w)$ satisfies Assumption 1, with $(\bar{K}, \bar{\mu})$ the $Q^{(V)}$-augmentation of K* (17)*. Then, the following hold:*

(i) If $\|w\|_\infty < \infty$, then $\mathcal{E}_L(g) \leq \|w\|_\infty \mathcal{E}_{\bar{K}}(g)$ for all $g \in L^2(\bar{\mu})$.
If $\underline{c} := \bar{\mu}\text{-ess inf}\, w$, then $\mathcal{E}_L(g) \geq \underline{c}\,\mathcal{E}_{\bar{K}}(g)$ for all $g \in L^2(\bar{\mu})$.
(ii) If $\|w^*\|_\infty < \infty$, then $\mathcal{E}_L(g) \leq \|w^*\|_\infty \mathcal{E}_{\bar{K}}(g)$ for all $g \in L^2(\mu)$.

Proof. This is done separately below for the cases $L \in \{P, K^{\mathrm{DA0}}, K^{\mathrm{DA1}}\}$. Set $G := [g(x) - g(x')]^2$, $g \in L^2(\bar{\mu})$, with $x, x' \in \mathbf{X} := \mathbf{T} \times \mathbf{U} \times \mathbf{V}$. Then,

$$\begin{aligned}\mathcal{E}_P(g) &= \frac{1}{2}\int \pi(\mathrm{d}x) q_\theta(\mathrm{d}\theta') Q^{(U)}_{\theta'}(\mathrm{d}u') Q^{(V)}_{\theta' u'}(\mathrm{d}v') \min\left\{1, r^{(V)}(x,x')\right\} G \\ &= \frac{1}{2}\int \bar{\mu}(\mathrm{d}x) q_\theta(\mathrm{d}\theta') Q^{(U)}_{\theta'}(\mathrm{d}u') Q^{(V)}_{\theta' u'}(\mathrm{d}v') \min\left\{w(x), w(x) r^{(V)}(x,x')\right\} G \\ &= \frac{1}{2}\int \bar{\mu}(\mathrm{d}x) q_\theta(\mathrm{d}\theta') Q^{(U)}_{\theta'}(\mathrm{d}u') Q^{(V)}_{\theta' u'}(\mathrm{d}v') \min\left\{w(x), w(x') r^{(U)}(x,x')\right\} G,\end{aligned}$$

because $w(x) r^{(V)}(x,x') = w(x') r^{(U)}(x,x')$, well-defined on the set of interest. We then use the uniform bounds $\underline{c} \le w \le \|w\|_\infty$ to conclude (i) for $L = P$.

Now assume $g \in L^2(\mu)$, so $G = [g(\theta,u) - g(\theta',u')]^2$. By Jensen's inequality and concavity of $(x,x') \mapsto \min\{x,x'\}$ when one of x, $x' \ge 0$ is held fixed,

$$\begin{aligned}\mathcal{E}_P(g) &= \frac{1}{2}\int \bar{\mu}(\mathrm{d}x) q_\theta(\mathrm{d}\theta') Q^{(U)}_{\theta'}(\mathrm{d}u') G \int Q^{(V)}_{\theta' u'}(\mathrm{d}v') \min\left\{w(x), w(x') r^{(U)}(x,x')\right\} \\ &\le \frac{1}{2}\int \bar{\mu}(\mathrm{d}x) q_\theta(\mathrm{d}\theta') Q^{(U)}_{\theta'}(\mathrm{d}u') G \min\left\{w(x), w^*(\theta',u') r^{(U)}(x,x')\right\}.\end{aligned}$$

Here, we have used that $r^{(U)}(x,x')$ does not depend on $v' \in \mathbf{V}$, and that

$$\int w(x) Q^{(V)}_{\theta u}(\mathrm{d}v) = \frac{c_\eta}{c_\zeta}\frac{1}{\eta(1)}\int \zeta(1) Q^{(V)}_{\theta u}(\mathrm{d}v) = \frac{\pi^*(\mathrm{d}\theta, \mathrm{d}u)}{\mu(\mathrm{d}\theta, \mathrm{d}u)} = w^*(\theta,u).$$

We then apply Jensen again, this time integrating out $v \in \mathbf{V}$, to get,

$$\begin{aligned}&\mathcal{E}_P(g) \\ &\le \frac{1}{2}\int \mathrm{d}\theta Q^{(U)}_\theta(\mathrm{d}u)\frac{\eta(1)}{c_\eta} q_\theta(\mathrm{d}\theta') Q^{(U)}_{\theta'}(\mathrm{d}u') G \int Q^{(V)}_{\theta u}(\mathrm{d}v) \min\left\{w(x), w^*(x') r^{(U)}(x,x')\right\} \\ &\le \frac{1}{2}\int \mathrm{d}\theta Q^{(U)}_\theta(\mathrm{d}u)\frac{\eta(1)}{c_\eta} q_\theta(\mathrm{d}\theta') Q^{(U)}_{\theta'}(\mathrm{d}u') \min\left\{w^*(\theta,u), w^*(\theta',u') r^{(U)}(x,x')\right\} G.\end{aligned}$$

We then apply the uniform bound $w^* \le \|w^*\|_\infty$ and use the fact that $\mathcal{E}_K(g) = \mathcal{E}_{\bar{K}}(g)$ for all $g \in L^2(\mu)$ to conclude (ii) for $L = P$.

Now consider the case $L = K^{\mathrm{DA0}}$. With $G := [g(x) - g(x')]^2$ on $\mathbf{X}^2$,

$$\begin{aligned}\mathcal{E}_{K^{\mathrm{DA0}}}(g) &= \frac{1}{2}\int \pi(\mathrm{d}x) q_\theta(\mathrm{d}\theta') Q^{(U)}_{\theta'}(\mathrm{d}u') \alpha(\theta,u;\theta',u') Q^{(V)}_{\theta' u'}(\mathrm{d}v') \min\left\{1, \frac{w(x')}{w(x)}\right\} G \\ &= \frac{1}{2}\int \bar{\mu}(\mathrm{d}x) q_\theta(\mathrm{d}\theta') Q^{(U)}_{\theta'}(\mathrm{d}u') \alpha(\theta,u;\theta',u') Q^{(V)}_{\theta' u'}(\mathrm{d}v') \min\left\{w(x), w(x')\right\} G,\end{aligned}$$

for all $g \in L^2(\bar{\mu})$. As before, this allows us to conclude (i) for $L = K^{\mathrm{DA0}}$.

Now assume $g \in L^2(\mu)$, with $G := [g(\theta,u) - g(\theta',u')]^2$. By Jensen,

$$\begin{aligned}\mathcal{E}_{K^{\mathrm{DA0}}}(g) &\le \frac{1}{2}\int \bar{\mu}(\mathrm{d}x) q_\theta(\mathrm{d}\theta') Q^{(U)}_{\theta'}(\mathrm{d}u') \alpha(\theta,u;\theta',u') G \min\left\{w(x), w^*(\theta',u')\right\} \\ &\le \frac{1}{2}\int \mu(\mathrm{d}\theta, \mathrm{d}u) q_\theta(\mathrm{d}\theta') Q^{(U)}_{\theta'}(\mathrm{d}u') \alpha(\theta,u;\theta',u') G \min\left\{w^*(\theta,u), w^*(\theta',u')\right\},\end{aligned}$$

which allows us to conclude (ii) as before.

Now consider the case $L = K^{\mathrm{DA1}}$. With $G := [g(x) - g(x')]^2$ on $\mathbf{X}^2$,

$$\begin{aligned}\mathcal{E}_{K^{\mathrm{DA1}}}(g) &= \frac{1}{2}\int \pi(\mathrm{d}x)K_{\theta u}(\mathrm{d}\theta', \mathrm{d}u')Q^{(V)}_{\theta' u'}(\mathrm{d}v')\min\Big\{1, \frac{w(x')}{w(x)}\Big\}G \\ &= \frac{1}{2}\int \bar{\mu}(\mathrm{d}x)K_{\theta u}(\mathrm{d}\theta', \mathrm{d}u')Q^{(V)}_{\theta' u'}(\mathrm{d}v')\min\big\{w(x), w(x')\big\}G,\end{aligned}$$

for all $g \in L^2(\bar{\mu})$. As before, this allows us to conclude (i) for $L = K^{\mathrm{DA1}}$.

Now assume $g \in L^2(\mu)$, with $G := [g(\theta, u) - g(\theta', u')]^2$. By Jensen,

$$\begin{aligned}\mathcal{E}_{K^{\mathrm{DA1}}}(g) &\le \frac{1}{2}\int \bar{\mu}(\mathrm{d}x)K_{\theta u}(\mathrm{d}\theta', \mathrm{d}u')G\min\big\{w(x), w^*(\theta', u')\big\} \\ &\le \frac{1}{2}\int \mu(\mathrm{d}\theta, \mathrm{d}u)K_{\theta u}(\mathrm{d}\theta', \mathrm{d}u')G\min\big\{w^*(\theta, u), w^*(\theta', u')\big\},\end{aligned}$$

which allows us to conclude (ii) as before. ■

Lemma 21. *With assumptions as in Lemma 20, and additionally assuming that K and L determine Harris ergodic chains, the following hold:*

(i) If $\|w\|_\infty < \infty$, then $(\bar{\mu}, \pi, w, \bar{K}, L, \underline{c}, \|w\|_\infty)$ satisfies Assumption 2.
(ii) If $\|w^*\|_\infty < \infty$, then $(\bar{\mu}, \pi, w, \bar{K}, L, 0, \|w^*\|_\infty)$ satisfies Assumption 3.

Proof. Lemma 20(i) and (ii) imply respectively (i) and (ii). ■

Proof of Theorem 14. The support condition Assumption 4(iii) implies that $(\bar{\mu}, \pi, w)$ satisfies Assumption 1. Under conditions (II), (I), or (III), the result follows by Lemma 21 and Theorem 12.

Assume condition (IV). Because $g := \hat{\zeta}(f)$ is a function on $\mathbf{X} = \mathbf{T} \times \mathbf{U} \times \mathbf{V}$ which does not depend on the second coordinate, $P^k g(\theta, u, v) = M^k g(\theta, v)$ for all $(\theta, u, v) \in \mathbf{X}$ and $k \ge 1$. Therefore, $\mathrm{var}(M, g) = \mathrm{var}(P, g)$. ■

Appendix C. Properties of augmented kernels

For measurable functions $V : \mathbf{X} \to [1, \infty)$ and $f : \mathbf{X} \to \mathbb{R}$, we set

$$\|\nu\|_V := \sup_{f:|f|\le V} \nu(f), \qquad \text{and} \qquad \|f\|_V := \sup_{x\in\mathbf{X}} \frac{|f(x)|}{V(x)}$$

for any finite signed measure ν on $\mathbf{X}$.

Definition 5. A μ-invariant Markov chain K on $\mathbf{X}$ is said to be

(i) *V-geometrically ergodic* if there is a function $V : \mathbf{X} \to [1, \infty)$ such that

$$\|K^n(x, \cdot) - \mu(\cdot)\|_V \le RV(x)\rho^n$$

for all $n \ge 1$, where $R < \infty$ and $\rho \in (0, 1)$ are constants.
(ii) *uniformly ergodic* if K is 1-geometrically ergodic.

Lemma 22. *Let $K_{\theta y}(\mathrm{d}\theta', \mathrm{d}y') = \dot{K}_\theta(\mathrm{d}\theta')Q_{\theta'}(\mathrm{d}y')$ be an augmented kernel on $\mathbf{T} \times \mathbf{Y}$.*

(i) The invariant measures of K and $\dot{K}$ satisfy $(\mu K = \mu \implies \mu^* \dot{K} = \mu^*)$, and $(\dot{\mu}\dot{K} = \dot{\mu} \implies \mu K = \mu)$, where $\mu(\mathrm{d}\theta, \mathrm{d}y) := \dot{\mu}(\mathrm{d}\theta)Q_\theta(\mathrm{d}y)$. These implications hold with invariance replaced with reversibility.
(ii) K is μ-Harris ergodic $\iff$ $\dot{K}$ is $\dot{\mu}$-Harris ergodic.

(iii) For all $f \in L^1(\mu)$ and $n \geq 1$, $K^n f(\theta, y) = \dot{K}^n(Qf)(\theta)$.
(iv) K is aperiodic $\iff$ $\dot{K}$ is aperiodic. K is positive $\iff$ $\dot{K}$ is positive.
(v) K is geometrically ergodic $\iff$ $\dot{K}$ is geometrically ergodic.
(vi) K is uniformly ergodic $\iff$ $\dot{K}$ is uniformly ergodic.

Proof. (i–iii) are [65, Lem. 24]. Proof of (iv) is straightforward.

For (v), consider first the case that $\dot{K}$ is $\dot{V}$-geometrically ergodic:

$$\sup_{|f| \leq \dot{V}} |\dot{K}^n(f)(\theta) - \dot{\mu}(f)| \leq R\dot{V}(\theta)\rho^n, \qquad n \geq 1,$$

with $\dot{V} : \mathbf{T} \to [1, \infty)$ and constants R and ρ. Define $V(\theta, y) := \dot{V}(\theta)$. By (iii),

$$\sup_{|f| \leq V} |K^n f(\theta, y) - \mu(f)| = \sup_{|f| \leq V} |\dot{K}^n(Qf)(\theta) - \dot{\mu}(Qf)|. \tag{35}$$

Since $Qf(\theta, y) \leq QV(\theta, y) = \dot{V}(\theta)$, we get that K is V-geometrically ergodic.

Assume now that K is V-geometrically ergodic. Using (35), we have,

$$\sup_{|f| \leq V} |K^n f(\theta, y) - \mu(f)| = \sup_{g = Qf : |f| \leq V} |\dot{K}^n g(\theta) - \dot{\mu}(g)|, \tag{36}$$

for $n \geq 1$. Define $\dot{V}(\theta) := \inf_y V(\theta, y)$. For all g such that $|g(\theta)| \leq \dot{V}(\theta)$, set $f(\theta, y) := g(\theta)$. Then $|f| \leq V$ and $Qf = g$. By (36), $\dot{K}$ is $\dot{V}$-geometrically ergodic. This proves (v), and (vi) follows from the form of $\dot{V}$ and V. ∎

Appendix D. Toy examples of two extremes

Figure 3. Mass allocations for μ, ν, and f on $\mathbf{X} = \{0, 1, 2\}$, $a \in [\frac{1}{2}, 1)$.

$$\begin{array}{rcl} \mu & = & \left(\begin{array}{ccc} \frac{1-a}{2} & \frac{1-a}{2} & a \end{array}\right) \\ \nu & = & \left(\begin{array}{ccc} 1/2 & 1/2 & 0 \end{array}\right) \\ f & = & \left(\begin{array}{ccc} 1 & -1 & 0 \end{array}\right) \end{array}$$

(A) 'MH/DA better' case

$$\begin{array}{rcl} \mu & = & \left(\begin{array}{ccc} 1/3 & 1/3 & 1/3 \end{array}\right) \\ \nu & = & \left(\begin{array}{ccc} \frac{a}{2} & \frac{1-a}{2} & 1/2 \end{array}\right) \\ f & = \frac{\sqrt{2}}{\sqrt{a+a^2}} & \left(\begin{array}{ccc} 1 & 0 & -a \end{array}\right) \end{array}$$

(B) 'IS better' case

Let $\mathbf{X} := \{0, 1, 2\}$ and consider the two mass allocations for probabilities μ and ν on $\mathbf{X}$ and function $f \in L^2_0(\nu)$ given pictorially in Figure 1 and precisely in Figure 3. Denote by $q^{(r)}$ the (reflected) random walk proposal on $\mathbf{X}$, given by $q_0^{(r)}(x) = \delta_1(x)$, $q_1^{(r)}(x) = \frac{1}{2}[\delta_0(x) + \delta_2(x)]$, and $q_2^{(r)}(x) = \delta_1(x)$, and by $q_x^{(u)}(x')$ the uniform proposal on $\mathbf{X}$. We set $K :=\mathrm{MH}(q \to \mu)$ and let L be the MH (6) or DA0 (7) kernels, using proposals $q^{(r)}$ or $q^{(u)}$, and targeting ν. We use a parameter $a \in [\frac{1}{2}, 1)$ to allow for continuous intensity shifts in the mass allocations in our examples. Because μ is constant on the support of ν, one can check that the MH and DA0 kernels coincide for $a \in [\frac{1}{2}, 1)$.

The resulting IS and MH/DA asymptotic variances, $\mathrm{var}(K, wf)$ and $\mathrm{var}(L, f)$, are listed in Table 1, and plotted in Figure 5. Here,

$$\mathrm{UB}_a(f) := \max(w)\mathrm{var}(L, f) + \nu(f^2[\max(w) - w]). \tag{37}$$

is the upper bound on $\mathrm{var}(K, wf)$ from Corollary 4.

TABLE 1. Asymptotic variance as a function of $a \in [1/2, 1)$

Proposal	$\mathrm{var}(L,f) \leq \mathrm{var}(K,wf)$	$\mathrm{var}(L,f) \geq \mathrm{var}(K,wf)$
RW $q^{(r)}$	1	$\frac{1}{1-a}$	$\frac{-1+8a+a^2}{a^2-1}$	$\frac{9a}{1+a}$
uniform $q^{(u)}$	2	$\frac{1}{1-a}$	$\frac{-1+10a-a^2}{(1+a)^2}$	$\frac{15a}{4(1+a)}$

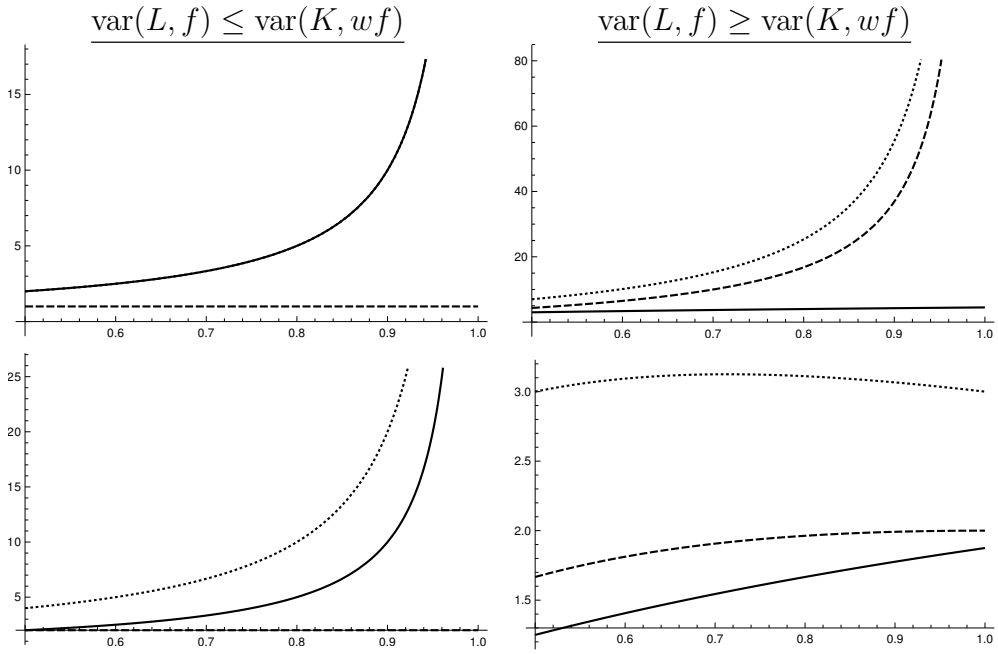


FIGURE 5. Plots from Table 1: $\mathrm{var}(K,wf)$ '—', $\mathrm{var}(L,f)$ '– –', and $\mathrm{UB}_a(f)$ '$\cdots$', vs. $a \in [\frac{1}{2}, 1)$. Here, in the top left, $\mathrm{UB}_a(f)$ exactly coincides with $\mathrm{var}(K,wf)$.

The code used to calculate the asymptotic variances can be found in the earlier preprint [26, App. C]. It is based on a straightforward diagonalisation of 3×3 matrices and a discrete version of a spectral formula [34, Cor. 1.5] for $\mathrm{var}(K,f)$ of Proposition 1.

REFERENCES

[1] C. Andrieu. On random- and systematic-scan samplers. *Biometrika*, 103(3):719–726, 2016.

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

[3] C. Andrieu, A. Doucet, S. Yıldırım, and N. Chopin. On the utility of Metropolis-Hastings with asymmetric acceptance ratio. Preprint arXiv:1803.09527, 2018.

[4] C. Andrieu, A. Lee, and M. Vihola. Uniform ergodicity of the iterated conditional SMC and geometric ergodicity of particle Gibbs samplers. *Bernoulli*, 24(2), 2018.

[5] C. Andrieu and É. Moulines. On the ergodicity properties of some adaptive MCMC algorithms. *J. Appl. Probab.*, 16(3):1462–1505, 2006.
[6] C. Andrieu and G. Roberts. The pseudo-marginal approach for efficient Monte Carlo computations. *Ann. Statist.*, 37(2):697–725, 2009.
[7] C. Andrieu and J. Thoms. A tutorial on adaptive MCMC. *Statist. Comput.*, 18(4):343–373, 2008.
[8] C. Andrieu and M. Vihola. Convergence properties of pseudo-marginal Markov chain Monte Carlo algorithms. *Ann. Appl. Probab.*, 25(2):1030–1077, 04 2015.
[9] C. Andrieu and M. Vihola. Establishing some order amongst exact approximations of MCMCs. *Ann. Appl. Probab.*, 2016. arXiv:1404.6909.
[10] M. Banterle, C. Grazian, A. Lee, and C. Robert. Accelerating Metropolis-Hastings algorithms by delayed acceptance. Preprint arXiv:1503.00996, 2015.
[11] J. Bardsley, A. Solonen, H. Haario, and M. Laine. Randomize-then-optimize: A method for sampling from posterior distributions in nonlinear inverse problems. *SIAM J. Sci. Comput.*, 36(4):A1895–A1910, 2014.
[12] F. Bassetti and P. Diaconis. Examples comparing importance sampling and the Metropolis algorithm. *Illinois J. Math.*, 50(1-4):67–91, 2006.
[13] P. Baxendale. Renewal theory and computable convergence rates for geometrically ergodic Markov chains. *Ann. Appl. Probab.*, 15(1B):700–738, 2005.
[14] R. Bellman. Some inequalities for the square root of a positive definite matrix. *Linear Algebra Appl.*, 1(3):321–324, 1968.
[15] S. Caracciolo, A. Pelissetto, and A. Sokal. Nonlocal Monte Carlo algorithm for self-avoiding random walks with fixed endpoints. *J. Stat. Phys.*, 60:1–53, 1990.
[16] N. Chopin, P. Jacob, and O. Papaspiliopoulos. SMC^2: A sequential Monte Carlo algorithm with particle Markov chain Monte Carlo updates. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 75(3):397–426, 2013.
[17] J. Christen and C. Fox. Markov chain Monte Carlo using an approximation. *J. Comput. Graph. Statist.*, 14(4), 2005.
[18] T. Cui, Y. Marzouk, and K. Willcox. Scalable posterior approximations for large-scale Bayesian inverse problems via likelihood-informed parameter and state reduction. *J. Comput. Phys.*, 315:363–387, 2016.
[19] G. Deligiannidis, A. Doucet, M. K. Pitt, and R. Kohn. The correlated pseudo-marginal method. Preprint arXiv:1511.04992, 2015.
[20] T. Dodwell, C. Ketelsen, R. Scheichl, and A. Teckentrup. A hierarchical multilevel Markov chain Monte Carlo algorithm with applications to uncertainty quantification in subsurface flow. Preprint arXiv:1303.7343, 2013.
[21] H. Doss. Discussion: Markov chains for exploring posterior distributions. *Ann. Statist.*, 22(4):1728–1734, 1994.
[22] R. Douc and C. Robert. A vanilla Rao-Blackwellization of Metropolis-Hastings algorithms. *Ann. Statist.*, 39(1):261–277, 2011.
[23] A. Doucet, M. Pitt, G. Deligiannidis, and R. Kohn. Efficient implementation of Markov chain Monte Carlo when using an unbiased likelihood estimator. *Biometrika*, 102(2):295–313, 2015.
[24] Y. Efendiev, T. Hou, and W. Luo. Preconditioning Markov chain Monte Carlo simulations using coarse-scale models. *SIAM J. Sci. Comput.*, 28(2):776–803,

2006.
[25] J. M. Flegal and G. L. Jones. Batch means and spectral variance estimators in Markov chain Monte Carlo. *Ann. Statist.*, 38(2):1034–1070, 2010.
[26] J. Franks and M. Vihola. Importance sampling and delayed acceptance via a Peskun type ordering. Preprint arXiv:1706.09873v1, 2017.
[27] A. Gelman, J. Carlin, H. Stern, and D. Rubin. *Bayesian Data Analysis*. Chapman & Hall/CRC, 1995.
[28] W. Gilks and G. Roberts. Strategies for improving MCMC. In *Markov chain Monte Carlo in practice*, volume 6, pages 89–114. 1996.
[29] P. Glynn and D. Iglehart. Importance sampling for stochastic simulations. *Management Sci.*, 35(11):1367–1392, 1989.
[30] A. Golightly, D. Henderson, and C. Sherlock. Delayed acceptance particle MCMC for exact inference in stochastic kinetic models. *Statist. Comput.*, 25, 2015.
[31] W. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109, Apr. 1970.
[32] S. Jarner and E. Hansen. Geometric ergodicity of Metropolis algorithms. *Stochastic Process. Appl.*, 85(2):341–361, 2000.
[33] J. Johndrow, J. Mattingly, S. Mukherjee, and D. Dunson. Optimal approximating Markov chains for Bayesian inference. Preprint arXiv:1508.03387v3, 2017.
[34] C. Kipnis and S. Varadhan. Central limit theorem for additive functionals of reversible Markov processes and applications to simple exclusions. *Comm. Math. Phys.*, 104(1):1–19, 1986.
[35] A. Lee, C. Yau, M. Giles, A. Doucet, and C. Holmes. On the utility of graphics cards to perform massively parallel simulation of advanced Monte Carlo methods. *J. Comput. Graph. Statist.*, 19(4):769–789, 2010.
[36] F. Leisen and A. Mira. An extension of Peskun and Tierney orderings to continuous time Markov chains. *Statist. Sinica*, 18:1641–1651, 2008.
[37] D. Levin, Y. Peres, and E. Wilmer. *Markov chains and mixing times*. American Mathematical Society, 2009.
[38] L. Lin, K. Liu, and J. Sloan. A noisy Monte Carlo algorithm. *Phys. Rev. D*, 61, 2000.
[39] J. Liu. Metropolized independent sampling with comparisons to rejection sampling and importance sampling. *Statist. Comput.*, 6(2):113–119, 1996.
[40] J. Liu. *Monte Carlo Strategies in Scientific Computing*. Springer, New York, 2003.
[41] F. Maire, R. Douc, and J. Olsson. Comparison of asymptotic variances of inhomogeneous Markov chains with application to Markov chain Monte Carlo methods. *Ann. Statist.*, 42(4):1483–1510, 2014.
[42] S. Meyn and R. Tweedie. *Markov Chains and Stochastic Stability*. Cambridge University Press, second edition, 2009.
[43] A. Mira and C. Geyer. Ordering Monte Carlo Markov Chains. Technical report, School of Statistics, University of Minnesota, 1999.
[44] A. Mira and F. Leisen. Covariance ordering for discrete and continuous time Markov chains. *Statist. Sinica*, pages 651–666, 2009.

[45] R. Neal. Annealed importance sampling. *Statist. Comput.*, 11(2):125–139, 2001.
[46] J. Negrea and J. Rosenthal. Error bounds for approximations of geometrically ergodic Markov chains. Preprint arXiv:1702.07441, 2017.
[47] A. Owen. Statistically efficient thinning of a Markov chain sampler. *J. Comput. Graph. Statist.*, 26(3), 2017.
[48] A. Owen and Y. Zhou. Safe and effective importance sampling. *J. Amer. Statist. Assoc.*, 95(449):135–143, 2000.
[49] P. Parpas, B. Ustun, M. Webster, and Q. K. Tran. Importance sampling in stochastic programming: A Markov chain Monte Carlo approach. *INFORMS J. Comput.*, 27(2):358–377, 2015.
[50] P. Peskun. Optimum Monte-Carlo sampling using Markov chains. *Biometrika*, 60(3):607–612, 1973.
[51] D. Prangle. Lazy ABC. *Statist. Comput.*, 26(1-2):171–185, 2016.
[52] M. Quiroz, M.-N. Tran, M. Villani, and R. Kohn. Speeding up MCMC by delayed acceptance and data subsampling. *J. Comput. Graph. Statist.*, 2017. To appear.
[53] G. Roberts and J. Rosenthal. Harris recurrence of Metropolis-within-Gibbs and trans-dimensional Markov chains. *Ann. Appl. Probab.*, 16(4):2123–2139, 2006.
[54] G. Roberts and J. Rosenthal. Coupling and ergodicity of adaptive Markov chain Monte Carlo algorithms. *J. Appl. Probab.*, 44(2):458–475, 2007.
[55] G. Roberts and R. Tweedie. Geometric convergence and central limit theorems for multidimensional Hastings and Metropolis algorithms. *Biometrika*, 83(1):95–110, 1996.
[56] C. Sherlock, A. Golightly, and D. Henderson. Adaptive, delayed-acceptance MCMC for targets with expensive likelihoods. *J. Comput. Graph. Statist.*, 26(2), 2017.
[57] C. Sherlock and A. Lee. Variance bounding of delayed-acceptance kernels. Preprint arXiv:1706.02142, 2017.
[58] C. Sherlock, A. Thiery, and A. Golightly. Efficiency of delayed-acceptance random walk Metropolis algorithms. Preprint arXiv:1506.08155, 2015.
[59] C. Sherlock, A. Thiery, and A. Lee. Pseudo-marginal Metropolis-Hastings using averages of unbiased estimators. Preprint arXiv:1610.09788, 2016.
[60] A. Solonen, P. Ollinaho, M. Laine, H. Haario, J. Tamminen, and H. Järvinen. Efficient MCMC for climate model parameter estimation: parallel adaptive chains and early rejection. *Bayesian Anal.*, 7(3):715–736, 2012.
[61] L. Tierney. A note on Metropolis-Hastings kernels for general state spaces. *Ann. Appl. Probab.*, 8(1):1–9, 1998.
[62] M.-N. Tran, M. Scharth, M. Pitt, and R. Kohn. Importance sampling squared for Bayesian inference in latent variable models. *arXiv:1309.3339v3*, 2014.
[63] P. Vanetti, A. Bouchard-Côté, G. Deligiannidis, and A. Doucet. Piecewise deterministic Markov chain Monte Carlo. Preprint arXiv:1707.05296, 2017.
[64] M. Vihola and J. Franks. On the use of ABC-MCMC with inflated tolerance and post-correction. Preprint arXiv:1902.00412, 2019.
[65] M. Vihola, J. Helske, and J. Franks. Importance sampling type estimators based on approximate marginal MCMC. Preprint arXiv:1609.02541v5, 2016.

Department of Mathematics and Statistics, University of Jyväskylä P.O.Box 35, FI-40014 Univ. of Jyväskylä, Finland
Email address: jordan.j.franks@jyu.fi, matti.vihola@iki.fi

Article [C]

Unbiased inference for discretely observed hidden Markov model diffusions

Jordan Franks, Ajay Jasra, Kody J.H. Law and Matti Vihola

Preprint arXiv:1807.10259v4, 2018.

UNBIASED INFERENCE FOR DISCRETELY OBSERVED HIDDEN MARKOV MODEL DIFFUSIONS

JORDAN FRANKS, AJAY JASRA, KODY J. H. LAW & MATTI VIHOLA

Abstract. We develop a Bayesian inference method for diffusions observed discretely and with noise, which is free of discretisation bias. Unlike existing unbiased inference methods, our method does not rely on exact simulation techniques. Instead, our method uses standard time-discretised approximations of diffusions, such as the Euler–Maruyama scheme. Our approach is based on particle marginal Metropolis–Hastings, a particle filter, randomised multilevel Monte Carlo, and importance sampling type correction of approximate Markov chain Monte Carlo. The resulting estimator leads to inference without a bias from the time-discretisation as the number of Markov chain iterations increases. We give convergence results and recommend allocations for algorithm inputs. Our method admits a straightforward parallelisation, and can be computationally efficient. The user-friendly approach is illustrated on two examples, where the underlying diffusion is an Ornstein–Uhlenbeck process or a geometric Brownian motion.

1. Introduction

Hidden Markov models (HMMs) are widely used in real applications, for example, for financial and physical systems modeling [cf. 5]. We focus on the case where the hidden Markov chain arises from a diffusion process that is observed with noise at some number of discrete points in time [cf. 30]. The parameters associated to the model are static and assigned a prior density. Bayesian inference involves expectations with respect to (w.r.t.) the joint posterior distribution of parameters and states, and is important in problems of model calibration and uncertainty quantification. A difficult part of Bayesian inference for these models is simulation of the diffusion dynamics. Except for some special cases where the transition probability is explicitly known [cf. 22, Section 4.4] or exact simulation [3] type methods can be applied [cf. 3, 4, 10, 33], one must time-discretise the diffusion dynamics with an approximation scheme in order to facilitate tractable inference. This is despite the fact that one is ideally interested when there is no time-discretisation: *unbiased inference*.

Our goal is unbiased inference for HMM diffusions. As previously mentioned, one approach to unbiased inference is based on exact simulation type methods [3, 4, 10, 33]. At the present point in time, exact simulation type methods are mostly only applicable to one-dimensional models where the Lamperti transformation [cf. 24] can be applied (cf. [25, 28, 33] for reviews). In contrast, we proceed with an Euler–Maruyama [cf. 22] (referred henceforth as *Euler*) or similar time-discretisation of the diffusion, which is generally applicable.

Traditional inference approaches based on time-discretisations face a trade-off between bias and computational cost. Once the user has decided on a suitably fine discretisation size, one can run, for example, the particle marginal Metropolis-Hastings (PMMH) [2]. This algorithm uses a particle filter (PF) [cf. 7], where proposals between time points are generated by the approximation scheme, and ultimately accepted or rejected according to a

2010 *Mathematics Subject Classification.* 65C05 (primary); 60H35, 65C35, 65C40 (secondary).

Key words and phrases. Diffusion, importance sampling, Markov chain/multilevel/sequential Monte Carlo.

Metropolis-Hastings type acceptance ratio [cf. 16]. As the discretisation size adopted must be quite fine, a PMMH algorithm can be computationally intensive.

To deal with the computational cost of PMMH, [19] develop a PMMH based method which uses (deterministic) multilevel Monte Carlo (dMLMC) [14, 17]. The basic premise of MLMC is to introduce a telescoping sum representation of the posterior expectation associated to the most precise time discretisation. Then, given an appropriate coupling of posteriors with 'consecutive' time discretisations, the cost associated to a target mean square error is reduced, relative to exact sampling from the most precise (time-discretised) posterior. In the HMM diffusion context, the standard MLMC method is not possible, so based upon a PF coupling approach and PMMH, an MLMC method is devised in [19, 20], which achieves fine-level, though biased, inference.

1.1. **Method.** The unbiased and computationally efficient inference method suggested in this paper is built firstly on PMMH, using Euler type discretisations, but using a PMMH targeting a coarse-level model, which is less computationally expensive. This does not yield unbiased inference yet, but it can be achieved by an importance sampling (IS) type correction [cf. 32].

We suggest an IS type correction that is based on a single-term (randomised) MLMC type estimator [23, 26] and the PF coupling approach of [19]. The rMLMC correction is based on randomising the running level in the multilevel context of a certain PF, which we refer to as the 'delta PF (ΔPF)' (Algorithm 3). In short, the ΔPF uses the PF coupling introduced in [19], but here an estimator is used for unbiased estimation of the difference of *unnormalised* integrals corresponding to two consecutive discretisation levels, over the latent states with parameter held fixed (cf. Section 2), rather than to the difference of self-normalised PMMH averages.

The resulting IS type estimator leads to unbiased inference over the joint posterior distribution, and is highly parallelisable, as the more costly (randomised) ΔPF corrections may be performed independently *en masse* given the PMMH base chain output. We are also able to suggest optimal choices for algorithm inputs in a straightforward manner (Recommendation 1 and Figure 1). This is because there is no bias, and therefore the difficult cost–variance–bias trade-off triangle associated with dMLMC is not present. Besides being unbiased and efficient, our method is user-friendly, as it is a combination of well-known and relatively straightforward components: PMMH, Euler approximations, PF, rMLMC, and an IS type estimator. For more about the strengths of the method, see Remark 10 later, as well as [12, 32] for more discussion about IS (type) estimators based on approximate Markov chain Monte Carlo (MCMC).

Key to verifying consistency of the method is a finite variance assumption for the rΔPF estimator. We verify a parameter-uniform bound for the variance under a simple set of HMM diffusion conditions in Section 3. Note, however, that consistency of our method is likely to hold more generally. This is in contradistinction to methods based on exact simulation, which require analytically tractable transformations to unit covariance diffusion term and computable bounds in the rejection sampler, in order to even apply the method (see for example the review in the recent preprint [33]).

If an exact simulation method is applicable, the obvious question arises whether our method or the exact simulation method should be applied. The efficiency of exact simulation type methods is dependent upon several and different factors than our method. These factors for exact simulation include proper tuning and tight computable bounds for the rejection sampler. In an ideal scenario for exact simulation, a method based on exact simulation is likely to perform better than our method. However, in the reverse case, our

method can perform better, if the efficiency of exact simulation is poor. For instance, the efficiency of exact simulation decreases to zero as the analytically computed upper bound of the IS weight used in the rejection sampler increases to infinity.

Although we have mostly in mind the case of Euler approximation schemes for the diffusion dynamics approximation, which are generally implementable, other schemes could be possibly be used as well [cf. 13]. However, suitable couplings for these schemes in dimensions greater than one may not be trivial. For the sake of theory and proof of consistency, ideally these would have also known weak and strong order convergence rates [cf. 22]. Indeed, assuming a coupling exists, such higher-order schemes can improve convergence of our method (see Sections 5 and 6). More generally, our approach based on PMMH or other approximate MCMC, increasingly fine families of approximations, MLMC, and IS correction, could be applied beyond the HMM diffusion context, for example, to HMM jump-diffusions [cf. 21].

1.2. **Outline.** Section 2 introduces the aforementioned ΔPF (Algorithm 2) and subsequently discusses some applications of randomisation techniques. The theoretical properties of the ΔPF in the HMM diffusion context are summarised in Section 3. Section 4 presents the suggested IS type estimator (Algorithm 4), based on PMMH with rMLMC (i.e. rΔPF) correction, and details its consistency and a corresponding central limit theorem (CLT). Section 5 suggests suitable allocations in the ΔPF based on rMLMC efficiency considerations. The numerical experiments in Section 6 illustrate our method in practice in the setting of an Ornstein–Uhlenbeck process and geometric Brownian motion. Proofs for the technical results of Sections 3, 4 and 5 are given in Appendix A, B and C, respectively.

1.3. **Notation.** Let $(E_n, \mathcal{E}_n)$ be a measurable space. Functions $\varphi : E_n \to \mathbb{R}$ will be assumed measurable. We denote by $\mathscr{P}(E_n)$ the collection of probability measures on $(E_n, \mathcal{E}_n)$, and by $\mathcal{B}_b(E_n)$ the set of $\varphi : E_n \to \mathbb{R}$ with $\|\varphi\| := \sup_{x \in \mathsf{X}} |\varphi(x)| < \infty$. For a measure μ on $(E_n, \mathcal{E}_n)$, we set $\mu(\varphi) := \int_{E_n} \varphi(x)\mu(\mathrm{d}x)$ whenever well-defined. For $K : E_n \times E_n \to [0, 1]$ a Markov kernel and $\mu \in \mathscr{P}(E_n)$, we set $\mu K(\mathrm{d}y) := \int_{E_n} \mu(\mathrm{d}x)K(x, \mathrm{d}y)$, and $K(\varphi)(x) := \int_{E_n} \varphi(y)K(x, \mathrm{d}y)$, whenever well-defined. We use the convention $\prod_{\emptyset} := 1$, and $p{:}q := \{r \in \mathbb{Z} : p \leq r \leq q\}$.

2. Delta particle filter for unbiased estimation of level differences

Consider the (Itô) diffusion process

$$\mathrm{d}X_t = a_\theta(X_t)\mathrm{d}t + b_\theta(X_t)\mathrm{d}W_t, \qquad t \geq 0, \tag{1}$$

with $X_t \in \mathsf{X} := \mathbb{R}^d$, model parameter $\theta \in \mathsf{T}$, and $\{W_t\}_{t \geq 0}$ a Brownian motion of appropriate dimension. We suppose that there are data $\{Y_p = y_p\}_{p=0}^n$, $y_p \in \mathbb{R}^m$, which are observed at equally spaced discrete times, $p = 0{:}n$ for simplicity. The Markov transition between X_{p-1} and X_p is given by some kernel $M_p^{(\theta,\infty)}(x_{p-1}, \mathrm{d}x_p)$. It is assumed that conditional on X_p, Y_p is independent of random variables $\{X_i, Y_i\}_{i \neq p}$ and has density $g_\theta(y_p|x_p) =: G_p^{(\theta)}(x_p)$. The resulting pair $(M_p^{(\theta,\infty)}, G_p^{(\theta)})$ defines the HMM diffusion, and is an example of a so-called *Feynman-Kac model* [cf. 7] described below. As the results of this section can just as easily be stated in terms of Feynman-Kac models, we do so in the following, which shows the generality of our approach.

2.1. **Particle filters.** A Feynman–Kac model (M_n, G_n) on spaces $(E_n, \mathcal{E}_n)$ arises when

(i) $M_n(x_{0:n-1}, \mathrm{d}x_n)$ are (regular) probability 'transition' kernels from $E_{0:n-1}$ to E_n for $n \geq 1$, and $M_0(x_{-1:0}, \mathrm{d}x_0) := \eta_0(\mathrm{d}x_0) \in \mathscr{P}(E_0)$, and
(ii) $G_n(x_{0:n})$ are $[0, \infty)$-valued (measurable) 'potential' functions for $n \geq 0$.

Particle filter (Algorithm 1) [cf. 7] generates sets of samples and weights corresponding to the Feynman–Kac model, which for $\varphi : E_{0:n} \to \mathbb{R}$ lead to an unbiased estimator for the (unnormalised) *smoother* $\boldsymbol{\gamma}_n(G_n\varphi)$, defined here in terms of the (unnormalised) *predictor*

$$\boldsymbol{\gamma}_n(\varphi) := \int \varphi(x_{0:n}) \Big(\prod_{t=0}^{n-1} G_t(x_{0:t}) \Big) \eta_0(\mathrm{d}x_0) \prod_{t=1}^{n} M_t(x_{0:t-1}, \mathrm{d}x_t). \tag{2}$$

Algorithm 1 Particle filter for model $(M_{0:n}, G_{0:n}) := (M_t, G_t)_{t=0:n}$ with N particles.

In each line, i takes values $1{:}N$. Do:

(i) Sample $x_0^{(i)} \sim \eta_0(\,\cdot\,)$ and set $\boldsymbol{x}_0^{(i)} := x_0^{(i)}$.

(ii) Compute $\omega_0^{(i)} := G_0(\boldsymbol{x}_0^{(i)})$ and set $\bar{\omega}_0^{(i)} := \omega_0^{(i)}/\omega_0^*$ where $\omega_0^* = \sum_{j=1}^N \omega_0^{(j)}$.

For $t = 1{:}n$, do:

(iii) Given $\bar{\omega}_{t-1}^{(1:N)}$, sample $A_{t-1}^{(1:N)}$ satisfying $\mathbb{E}\big[\sum_{j=1}^N \mathbf{1}\{A_{t-1}^{(j)} = k\}\big] = N\bar{\omega}_{t-1}^{(k)}$.

(iv) Sample $x_t^{(i)} \sim M_t(\boldsymbol{x}_{t-1}^{A_{t-1}^{(i)}}, \,\cdot\,)$ and set $\boldsymbol{x}_t^{(i)} = (\boldsymbol{x}_{t-1}^{(A_{t-1}^{(i)})}, x_t^{(i)})$.

(v) Compute $\omega_t^{(i)} := G_t(\boldsymbol{x}_t^{(i)})$ and set $\bar{\omega}_t^{(i)} := \omega_t^{(i)}/\omega_t^*$ where $\omega_t^* := \sum_{j=1}^N \omega_t^{(j)}$.

Report $(V^{(1:N)}, \mathbf{X}^{(1:N)})$ where $V^{(i)} := \bar{\omega}_n^{(i)} \prod_{t=0}^n \frac{1}{N}\omega_t^*$ and $\mathbf{X}^{(i)} := \boldsymbol{x}_n^{(i)}$.
(In case $\omega_t^* = 0$, the algorithm is terminated with $V^{(i)} = 0$ and with arbitrary $\mathbf{X}^{(i)} \in E_{0:n}$.)

Proposition 1. *Suppose that $\varphi : E_{0:n} \to \mathbb{R}$ is such that $\boldsymbol{\gamma}_n(G_n\varphi) < \infty$. Then, the output of Algorithm 1 satisfies*

$$\mathbb{E}\Big[\sum_{i=1}^N V^{(i)} \varphi(\mathbf{X}^{(i)}) \Big] = \boldsymbol{\gamma}_n(G_n\varphi).$$

Proposition 1 is a restatement of [7, Theorem 7.4.2] in case $A_{t-1}^{(i)}$ are sampled independently ('multinomial resampling'). The extension to the general unbiased case, which covers popular residual, stratified and systematic resampling schemes [cf. 5, 8], is straightforward [cf. 32].

2.2. **Level difference estimation.** Suppose that we have two Feynman–Kac models (M_n^F, G_n^F) and (M_n^C, G_n^C) defined on common spaces $(E_n, \mathcal{E}_n)$. The models correspond to 'finer' and 'coarser' Euler type discretised HMM diffusions. We are interested in estimating (unbiasedly) the difference

$$\boldsymbol{\gamma}_n^F(G_n^F\varphi) - \boldsymbol{\gamma}_n^C(G_n^C\varphi). \tag{3}$$

If the models are close to each other, as they will be in the multilevel (diffusion) context, we would like the estimator also to be typically small. In many contexts, if one can estimate the difference using a coupling, it is possible to obtain a variance reduction. The particular coupling approach we use here is based on using a combined Feynman–Kac model as in [19], which provides a simple, general and effective coupling of PFs, and which we will use to estimate the level difference of unnormalised smoother (3).

Hereafter, we denote $\check{x}_n = (\check{x}_n^F, \check{x}_n^C) \in E_n \times E_n$, and for $\check{x}_{0:n} = (\check{x}_0, \dots, \check{x}_n) \in E_0^2 \times \dots E_n^2$, we set $\check{x}_{0:n}^s := (\check{x}_0^s, \dots, \check{x}_n^s) \in E_{0:n}$ for $s \in \{F, C\}$.

Assumption 2. Suppose that $(\check{M}_t, \check{G}_t)$ is a Feynman–Kac model on the product spaces $(E_t \times E_t, \mathcal{E}_t \otimes \mathcal{E}_t)$, such that:

(i) $\check{M}_t$ is a coupling of the probabilities M_t^F and M_t^C, i.e. for all $A \in \mathcal{E}_t$, we have

$$\int_{A\times E_t} \check{M}_t(\check{x}_{0:t-1}, \mathrm{d}\check{x}_t) = M_t^F(\check{x}^F_{0:t-1}, A), \qquad \int_{E_t\times A} \check{M}_t(\check{x}_{0:t-1}, \mathrm{d}\check{x}_t) = M_t^C(\check{x}^C_{0:t-1}, A),$$

and for $A \in \mathcal{E}_0$, we have $\check{\eta}_0(A \times E_0) = \eta_0^F(A)$ and $\check{\eta}_0(E_0 \times A) = \eta_0^C(A)$.

(ii) $\check{G}_t(\check{x}_{0:t}) := \frac{1}{2}\big[G_t^F(\check{x}^F_{0:t}) + G_t^C(\check{x}^C_{0:t})\big]$.

Algorithm 2 Delta particle filter (ΔPF) for unbiased estimation of level differences.

(i) Run Algorithm 1 with $(\check{M}_{0:n}, \check{G}_{0:n}, N)$, outputting $(\check{V}^{(1:N)}, \check{\mathbf{X}}^{(1:N)})$.

(ii) Report $(V^{(1:2N)}, \mathbf{X}^{(1:2N)})$ where

$$\big(V^{(i)}, \quad \mathbf{X}^{(i)}\big) := \begin{cases} \Big(\check{V}^{(i)} w^F(\check{\mathbf{X}}^{(i)}), \quad \check{\mathbf{X}}^{(i)F}\Big) & i = 1{:}N, \\ \Big(-\check{V}^{(i-N)} w^C(\check{\mathbf{X}}^{(i-N)}), \quad \check{\mathbf{X}}^{(i-N)C}\Big) & i = (N+1){:}2N, \end{cases}$$

and where $w^F(\check{x}_{0:n}) := \frac{\prod_{t=0}^n G_t^F(\check{x}^F_{0:t})}{\prod_{t=0}^n \check{G}_t(\check{x}_{0:t})}$ and $w^C(\check{x}_{0:n}) := \frac{\prod_{t=0}^n G_t^C(\check{x}^C_{0:n})}{\prod_{t=0}^n \check{G}_t(\check{x}_{0:t})}$.

Proposition 3. *Under Assumption 2, the output of Algorithm 2 satisfies*

$$\mathbb{E}\Big[\sum_{i=1}^{2N} V^{(i)}\varphi(\mathbf{X}^{(i)})\Big] = \boldsymbol{\gamma}_n^F(G_n^F\varphi) - \boldsymbol{\gamma}_n^C(G_n^C\varphi)$$

whenever both integrals on the right are well-defined and finite.

Proof. By the unbiasedness property of PF Algorithm 1, we have

$$\begin{aligned}\mathbb{E}\Big[\sum_{i=1}^{N} V^{(i)}\varphi(\mathbf{X}^{(i)})\Big] &= \int w^F(\check{x}_{0:n})\varphi(\check{x}^F_{0:n})\Big(\prod_{t=0}^n \check{G}_t(\check{x}_{0:t})\Big)\check{\eta}(\mathrm{d}x_0)\prod_{t=1}^n \check{M}_t(\check{x}_{0:t-1}, \mathrm{d}\check{x}_t) \\ &= \int \varphi(\check{x}^F_{0:n})\Big(\prod_{t=0}^n G_t^F(\check{x}^F_{0:t})\Big)\check{\eta}(\mathrm{d}x_0)M_t^F(\check{x}^F_{0:t-1}, \mathrm{d}\check{x}^F_t) = \boldsymbol{\gamma}_n^F(G_n^F\varphi),\end{aligned}$$

where Assumption 2(ii) guarantees $\check{G}_t > 0$ whenever $G_t^F > 0$, and (i) implies the marginal law of $\prod_{t=0}^n \check{M}_t$ is $\prod_{t=0}^n M_t^F$. Similarly, $\mathbb{E}\big[\sum_{i=N+1}^{2N} V^{(i)}\varphi(\mathbf{X}^{(i)})\big] = -\boldsymbol{\gamma}_n^C(G_n^C\varphi)$. □

Remark 4. Regarding Algorithm 2:

(i) Typically, in the discretisation of diffusions context [14, 26], the couplings $\check{M}_t$ would be based on using the same underlying Brownian motion [cf. 22]. That is, if

$$\begin{aligned}X^F_{t+h^F} &= X^F_t + a_\theta(X^F_t)h^F + b_\theta(X^F_t)\delta W^F_{t+h^F} \\ X^F_{t+2h^F} &= X^F_{t+h^F} + a_\theta(X^F_{t+h^F})h^F + b_\theta(X^F_{t+h^F})\delta W^F_{t+2h^F}\end{aligned}$$

with $\delta W^F_{t+kh^F} \sim N(0, h^F)$, $k = 1, 2, 3, \ldots$, corresponds to two steps of an Euler discretisation with step-size h^F, then we can use

$$X^C_{t+h^C} = X^C_t + a_\theta(X^C_t)h^C + b_\theta(X^C_t)\big(\delta W^F_{t+h^F} + \delta W^F_{t+2h^F}\big)$$

with $h^C := 2h^F$ for the coarser Euler discretisation. The kernels $\check{M}_t$ on the joint space then move N particles according to the fine-level discretisation, and N according to the coarse-level discretisation, both based on the same underlying sequence of standard normals $(\delta W^F_{t+kh^F})_{k\geq 1}$.

(ii) The choice of $\check{G}_t$ in Assumption 2(ii) provides a safe 'balance' in between the approximations, as w^F and w^C are upper bounded by 2^{n+1}. Indeed, the coupled Feynman–Kac model can be thought as an 'average' of the two extreme cases—with the choice $\check{G}_t(x_{0:t}) = G_t^F(\check{x}_{0:t}^F)$ the coupled PF would coincide marginally with the Feynman–Kac model with dynamics M_t^F. What is the optimal choice for $\check{G}_t$ is an interesting question.

(iii) Clearly, the choice of $\check{G}_{0:t}$ can be made also in other ways. It is sufficient for unbiasedness to choose $\check{G}_t(\check{x}_{0:t})$ such that it is strictly positive whenever either the $G_t^F(\check{x}_{0:t}^F)$ or $G_t^C(\check{x}_{0:t}^C)$ product is positive, but choices which make w^F and w^C bounded are safer, for instance $\check{G}_{0:t}(\check{x}_{0:t}) = \max\{G_t^F(\check{x}_{0:t}^F), G_t^C(\check{x}_{0:t}^C)\}$. This was the original choice made in [19] for approximation of normalised smoother differences. This PF coupling approach based on change of measure and weight corrections w^F and w^C, has been further used also, for example, in [20].

(iv) Later, in the HMM diffusion context, we set $G_t^F = G_t^C$, corresponding to common observational densities, but the method is also of interest with differing potentials.

2.3. **Unbiased latent inference.** We show here how the randomisation techniques of [23, 26] can be used with the output of Algorithms 1 and 2 to provide an unbiased estimator according to the true model, even though the PFs are only run according to approximate models. Let us index the transitions $M_p^{(\ell)}$ and potentials $G_p^{(\ell)}$ by $\ell \geq 0$. They are assumed throughout to be increasingly refined approximations, in the (weak) sense that

$$\boldsymbol{\gamma}_n^{(\ell)}(G_n^{(\ell)}\varphi) \longrightarrow \boldsymbol{\gamma}_n^{(\infty)}(G_n^{(\infty)}\varphi), \qquad \text{as} \quad \ell \to \infty, \tag{4}$$

for all $\varphi \in \mathcal{B}_b(E_{0:n})$, where

$$\boldsymbol{\gamma}_n^{(\ell)}(\varphi) := \int \varphi(x_{0:n}) \Big(\prod_{t=0}^{n-1} G_t^{(\ell)}(x_{0:t}) \Big) \eta_0^{(\ell)}(\mathrm{d}x_0) \prod_{t=1}^{n} M_t^{(\ell)}(x_{0:t-1}, \mathrm{d}x_t).$$

In Assumption 2 we set symbols (F, C) to be $(\ell, \ell-1)$ for $\ell \geq 1$. Algorithm 3 can then provide unbiased estimation of $\boldsymbol{\gamma}_n^{(\infty)}(G_n^{(\infty)}\varphi)$ (Lemma 6), leading to unbiased inference w.r.t. the normalised smoother

$$\varphi \mapsto \frac{\boldsymbol{\gamma}_n^{(\infty)}(G_n\varphi)}{\boldsymbol{\gamma}_n^{(\infty)}(G_n)},$$

which is stated as Proposition 7 below.

Algorithm 3 Unbiased estimator based on PF and rΔPF; N particles, probability $\mathbf{p} = (p_\ell)_{\ell \in \mathbb{N}}$.

(i) Run Algorithm 1 with $(M_{0:n}^{(0)}, G_{0:n}^{(0)}, N)$, outputting $(V^{(1:N)'}, \mathbf{X}^{(1:N)'})$.
(ii) Sample $L \sim \mathbf{p}$ (independently from the other random variables).
(iii) Run Algorithm 2 with $(M_{0:n}^{(L)}, G_{0:n}^{(L)}, N)$, outputting $(V^{(1:2N)}, \mathbf{X}^{(1:2N)})$.

Report $\big((V^{(1:N)'}, \mathbf{X}^{(1:N)'}), (V^{(1:2N)}, \mathbf{X}^{(1:2N)}), L\big)$.

Assumption 5. Assumption 2 holds, $\mathbf{p} = (p_\ell)_{\ell \in \mathbb{N}}$ is a probability on $\mathbb{N} := \mathbb{Z}_{\geq 1}$ with $p_\ell > 0$ for all $\ell \geq 1$, $g : E_{0:n} \to \mathbb{R}$ is a function, and

$$s_g := \sum_{\ell \geq 0} \frac{\mathbb{E}\Delta_\ell^2(g)}{p_\ell} < \infty, \tag{5}$$

where

$$\Delta_\ell(g) := \sum_{i=1}^{2N} V^{(i)} g(\mathbf{X}^{(i)}) \tag{6}$$

is formed from the output $(V^{(1:2N)}, \mathbf{X}^{(1:2N)})$ of Algorithm 2 with $(\check{M}_{0:n}^{(\ell)}, \check{G}_{0:n}^{(\ell)}, N)$.

Lemma 6. *Under Assumption 5, the estimator*

$$\zeta(g) := \sum_{i=1}^{N} V^{(i)'} g(\mathbf{X}^{(i)'}) + \frac{1}{p_L} \Delta_L(g) \tag{7}$$

formed from the output of Algorithm 3 satisfies

$$\mathbb{E}[\zeta(g)] = \boldsymbol{\gamma}_n^{(\infty)}(G_n^{(\infty)} g),$$

whenever $\boldsymbol{\gamma}_n^{(0)}(G_n g)$ *and* $\boldsymbol{\gamma}_n^{(\infty)}(G_n g)$ *are both finite.*

Proof. Under Assumption 5, we have [cf. 26, 31]

$$\mathbb{E}[p_L^{-1} \Delta_L(g)] = \boldsymbol{\gamma}_n^{(\infty)}(G_n^{(\infty)} g) - \boldsymbol{\gamma}_n^{(0)}(G_n^{(0)} g),$$

so the result follows by Proposition 1 and linearity of the expectation. □

The following suggests a fully parallelisable algorithm for unbiased inference over the normalised smoother, and is an unbiased alternative to the particle independent Metropolis-Hastings (PIMH) [2] run at some fine level of discretisation.

Proposition 7. *Suppose* $\mathbf{p}$ *on* $\mathbb{N}$ *satisfies Assumption 5 for functions* $g \in \{1, \varphi\}$, *with* $\boldsymbol{\gamma}_n^{(0)}(G_n^{(0)} g)$ *and* $\boldsymbol{\gamma}_n^{(\infty)}(G_n^{(\infty)} g)$ *finite, and* $\boldsymbol{\gamma}^{(\infty)}(G_n^{(\infty)}) > 0$. *For each* $k \in \{1{:}m\}$, *if one runs independently Algorithm 3, forming* $\zeta_k(g)$ *from the output as in (7) for each* k, *then*

$$E_{m,N,\mathbf{p}}(\varphi) := \frac{\sum_{k=1}^m \zeta_k(\varphi)}{\sum_{k=1}^m \zeta_k(1)} \xrightarrow{m\to\infty} p^{(\infty)}(\varphi) \qquad \textit{almost surely.}$$

Moreover, with $\bar{\varphi} := \varphi - p^{(\infty)}(\varphi)$,

$$\sqrt{m}[E_{m,N,\mathbf{p}}(\varphi) - p^{(\infty)}(\varphi)] \xrightarrow{m\to\infty} \mathcal{N}(0, \sigma^2) \qquad \textit{in distribution,}$$

where

$$\sigma^2 = \frac{s_{\bar{\varphi}} - \left(\boldsymbol{\gamma}^{(\infty)}(G_n^{(\infty)} \bar{\varphi}) - \boldsymbol{\gamma}^{(0)}(G_n^{(0)} \bar{\varphi})\right)^2}{[\boldsymbol{\gamma}^{(\infty)}(G_n^{(\infty)})]^2}.$$

The above result follows directly from the results of Section 4. It can also be seen as a multilevel version of [32, Proposition 23], with straightforward estimators for σ^2. See Section 5 for suggested choices for $\mathbf{p}$ and N_ℓ.

3. A variance bound for the delta particle filter

In this section we give theoretical results for the ΔPF (Algorithm 2) in the setting of HMM diffusions, which can be used to verify finite variance and therefore consistency of related estimators.

3.1. **Hidden Markov model diffusions.** We consider an HMM diffusion and corresponding Feynman-Kac model as in Section 2. We omit θ from the notation in the following, which is allowed as the remaining conditions and results in this Section 3 will hold uniformly in θ (i.e. any constants are independent of θ). The following will be assumed throughout.

Condition (D). The coefficients $a^j, b^{j,k}$ are twice differentiable for $j,k = 1,\dots,d$, and

(i) **uniform ellipticity**: $b(x)b(x)^T$ is uniformly positive definite;
(ii) **globally Lipschitz**: there is a $C > 0$ such that $|a(x)-a(y)|+|b(x)-b(y)| \le C|x-y|$ for all $x, y \in \mathbb{R}^d$;
(iii) **boundedness**: $\mathbb{E}|X_0|^p < \infty$ for all $p \ge 1$.

Let $M^{(\infty)}(x, \mathrm{d}y) =: M_p^{(\infty)}(x, \mathrm{d}y)$ for $p = 0{:}n$ denote the Markov transition of the unobserved diffusion (1), i.e. the distribution of the solution X_1 of (1) started at $X_0 = x$. With similar setup from Section 2, with $E_n := \mathsf{X}^{n+1}$, we have that (2) takes the form

$$\gamma_n^{(\infty)}(\varphi) = \int \varphi(x_{0:n}) \Big(\prod_{p=0}^{n-1} G_p(x_p) \Big) \eta_0(\mathrm{d}x_0) \prod_{p=1}^{n} M^{(\infty)}(x_{p-1}, \mathrm{d}x_p).$$

In practice one usually must approximate the true dynamics $M^{(\infty)}(x, \mathrm{d}y)$ of the underlying diffusion with a simpler transition $M^{(\ell)}(x, \mathrm{d}y)$, based on some Euler type scheme using a discretisation parameter $h_\ell = 2^{-\ell}$ for $\ell \ge 0$ [cf. 22]. The scheme allows for a coupling of the diffusions $(X_t^{(\ell)}, X_t^{(\ell-1)})_{t \ge 0}$ running at discretisation levels ℓ and $\ell - 1$ (based on using the same Brownian path W_t), such that for some $\beta \in \{1, 2\}$, we have

$$\mathbb{E}_{(x,y)}[|X_1^{(\ell)} - X_1^{(\ell-1)}|^2] \le M(|x-y|^2 + h_\ell^\beta), \tag{8}$$

where $M < \infty$ does not depend on $\ell \ge 1$. In particular, if the diffusion coefficient $b(X_t)$ in (1) is constant or if a Milstein scheme can be applied otherwise, then $\beta = 2$; otherwise $\beta = 1$ [cf. 18, Proposition D.1.].

3.2. **Variance bound.** Assume we are in the above HMM diffusion setting, and that the coupling of Assumption 2 holds, with symbols (F, C) equal to $(\ell, \ell-1)$ for $\ell \ge 1$, and $G_p^{(\ell)} = G_p^{(\ell-1)} := G_p$ for $p = 0{:}n$. Running Algorithm 2, we recall that $\Delta_\ell(\varphi)$, defined in (6), satisfies, by Proposition 3,

$$\mathbb{E}[\Delta_\ell(\varphi)] = \gamma_n^{(\ell)}(G_n\varphi) - \gamma_n^{(\ell-1)}(G_n\varphi),$$

regardless of the number $N \ge 1$ of particles.

Recall that a (measurable) function $\varphi : \mathsf{X} \to \mathbb{R}$ is Lipschitz, denoted $\varphi \in \mathrm{Lip}(\mathsf{X})$, if for some $C' < \infty$, $|\varphi(x) - \varphi(y)| \le C'|x - y|$ for all $x, y \in \mathsf{X}$.

Condition (A). The following conditions hold for the model (M_n, G_n):

(**A1**) (i) $\|G_n\| < \infty$ for each $n \ge 0$.
(ii) $G_n \in \mathrm{Lip}(\mathsf{X})$ for each $n \ge 0$.
(iii) $\inf_{x \in \mathsf{X}} G_n(x) > 0$ for each $n \ge 0$.

(**A2**) For every $n \ge 1$, $\varphi \in \mathrm{Lip}(\mathsf{X}) \cap \mathcal{B}_b(\mathsf{X})$ there exist a $C' < \infty$ such that for $s \in \{F, C\}$, we have for every $(x, y) \in \mathsf{X} \times \mathsf{X}$ that $|M_n^s(\varphi)(x) - M_n^s(\varphi)(y)| \le C'|x - y|$.

In the following results for $\Delta_\ell(\varphi)$, the constant $M < \infty$ may change from line-to-line. It will not depend upon N or ℓ (or θ), but may depend on the time-horizon n or the function φ. $\mathbb{E}$ denotes expectation w.r.t. the law associated to the ΔPF started at (x, x), with $x \in \mathsf{X}$. Below we only consider multinomial resampling in the ΔPF for simplicity, though Theorem 8 and Corollary 9 can be proved also assuming other resampling schemes.

Theorem 8. *Assume (A1-2). Then for any $\varphi \in \mathcal{B}_b(\mathsf{X}^{n+1}) \cap \mathrm{Lip}(\mathsf{X}^{n+1})$, there exists a $M < \infty$ such that*

$$\mathbb{E}\Big[\Big(\Delta_\ell(\varphi) - \mathbb{E}[\Delta_\ell(\varphi)]\Big)^2\Big] \leq \frac{M h_\ell^{2\wedge\beta}}{N}, \qquad \text{with } \beta \text{ as in (8).}$$

Corollary 9. *Assume (A1-2). Then for any $\varphi \in \mathcal{B}_b(\mathsf{X}^{n+1}) \cap \mathrm{Lip}(\mathsf{X}^{n+1})$, there exists a $M < \infty$ such that*

$$\mathbb{E}\Big[\Big(\Delta_\ell(\varphi)\Big)^2\Big] \leq M\Big(\frac{h_\ell^{2\wedge\beta}}{N} + h_\ell^2\Big), \qquad \text{with } \beta \text{ as in (8).}$$

The proofs are given in Appendix A.

Based on Corollary 9, Recommendation 1 of Section 5 suggests allocations for $\mathbf{p}$ and N_ℓ in the ΔPF (Algorithm 2) to optimally use resources and minimise variance (5).

4. Unbiased joint inference for hidden Markov model diffusions

We are interested in unbiased inference for the Bayesian model posterior

$$\pi(\mathrm{d}\theta, \mathrm{d}x_{0:n}) \propto \mathrm{pr}(\mathrm{d}\theta) G_n^{(\theta)}(x_n) \gamma_n^{(\theta,\infty)}(\mathrm{d}x_{0:n}),$$

where $\mathrm{pr}(\mathrm{d}\theta) = \mathrm{pr}(\theta)\mathrm{d}\theta$ is the prior on the model parameters, and

$$\gamma_n^{(\theta,\infty)}(\mathrm{d}x_{0:n}) = \Big(\prod_{t=0}^{n-1} G_t^{(\theta)}(x_t)\Big)\eta_0^{(\theta)}(\mathrm{d}x_0)\prod_{t=1}^{n} M_t^{(\theta,\infty)}(x_{t-1}, \mathrm{d}x_t).$$

Here, $M_t^{(\theta,\infty)}$ corresponds to the transition density of the diffusion model of interest. The dependence of the HMM on θ is made explicit in this section. As in Section 3, we assume the transition densities $M_t^{(\theta,\infty)}$ cannot be simulated, but that there are increasingly refined discretisations $M_t^{(\theta,\ell)}$ approximating $M_t^{(\theta,\infty)}$ in the sense of (4) (with $E_{0:n} := \mathsf{X}^{n+1}$).

4.1. Randomised MLMC IS type estimator based on coarse-model PMMH.

We now consider Algorithm 4 for joint inference w.r.t. the above Bayesian posterior. Algorithm 4 uses the following ingredients:

(i) $\check{M}_{0:n}^{(\theta,\ell)}$ satisfying Assumption 2(i) with $M_{0:n}^F = M_{0:n}^{(\theta,\ell)}$, and $M_{0:n}^C = M_{0:n}^{(\theta,\ell-1)}$.
(ii) $\check{G}_{0:n}^{(\theta)}$ defined as in Assumption 2(ii), with $G_{0:n}^F = G_{0:n}^C = G_{0:n}^{(\theta)}$.
(iii) Metropolis–Hastings proposal distribution $q(\cdot \mid \theta)$ for parameters.
(iv) Algorithm constant $\epsilon \geq 0$.
(v) Number of MCMC iterations $m_{\mathrm{iter}} \in \mathbb{N}$ and number of particles $N \in \mathbb{N}$.
(vi) Probability mass $\mathbf{p} = (p_\ell)_{\ell\in\mathbb{N}}$ on $\mathbb{N}$ with $p_\ell > 0$ for all $\ell \in \mathbb{N}$.

Remark 10. Before stating consistency and central limit theorems, we briefly discuss various aspects of this approach, which are appealing from a practical perspective, and we also mention certain algorithmic modifications which could be further considered.

(i) The first phase (P1) of Algorithm 4 implements a PMMH type algorithm [2]. If $\epsilon = 0$, this is exactly PMMH targeting the model $\pi^{(0)}(\mathrm{d}\theta, \mathrm{d}x_{0:n}) \propto \mathrm{pr}(\mathrm{d}\theta) G_n^{(\theta)}(x_n) \gamma_n^{(\theta,0)}(\mathrm{d}x_{0:n})$. It is generally safer to choose $\epsilon > 0$ [32], which ensures that the IS type correction in phase (P2) will yield consistent inference for the ideal model $\pi^{(\infty)}(\mathrm{d}\theta, \mathrm{d}x_{0:n}) \propto \mathrm{pr}(\mathrm{d}\theta) G_n^{(\theta)}(x_n) \gamma_n^{(\theta,\infty)}(x_{0:n})$ (Theorem 11). Setting $\epsilon > 0$ may be helpful otherwise in terms of improved mixing, as the PMMH will target marginally an averaged probability between a 'flat' prior and a 'multimodal' $\ell = 0$ marginal posterior.

Algorithm 4 Randomised multilevel importance sampling type estimator.

(P1) Let $(\Theta_0, V_0^{(1:N)}, \mathbf{X}_0^{(1:N)})$ such that $\sum_{i=1}^N V_0^{(i)} > 0$, and for $k = 1{:}m_{\text{iter}}$, iterate:

(i) Propose $\hat{\Theta}_k \sim q(\cdot \mid \Theta_{k-1})$.

(ii) Run Algorithm 1 with $(M_{0:n}^{(\hat{\Theta}_k,0)}, G_{0:n}^{(\hat{\Theta}_k)}, N)$ and call the output $(\hat{V}_k^{(1:N)}, \hat{\mathbf{X}}_k^{(1:N)})$.

(iii) With probability

$$\min\left\{1, \frac{\mathrm{pr}(\hat{\Theta}_k) q(\Theta_{k-1} \mid \hat{\Theta}_k) \big(\sum_{i=1}^N \hat{V}_k^{(i)} + \epsilon\big)}{\mathrm{pr}(\Theta_{k-1}) q(\hat{\Theta}_k \mid \Theta_{k-1}) \big(\sum_{j=1}^N V_{k-1}^{(j)} + \epsilon\big)}\right\},$$

accept and set $(\Theta_k, V_k^{(1:N)}, \mathbf{X}_k^{(1:N)}) \leftarrow (\hat{\Theta}_k, \hat{V}_k^{(1:N)}, \hat{\mathbf{X}}_k^{(1:N)})$; otherwise set $(\Theta_k, V_k^{(1:N)}, \mathbf{X}_k^{(1:N)}) \leftarrow (\Theta_{k-1}, V_{k-1}^{(1:N)}, \mathbf{X}_{k-1}^{(1:N)})$.

(P2) For all $k \in \{1{:}m_{\text{iter}}\}$, independently conditional on $(\Theta_k, V_k^{(1:N)}, \mathbf{X}_k^{(1:N)})$:

(i) Set $\mathbf{X}_{k,0}^{(1:N)} := \mathbf{X}_k^{(1:N)}$, and set $W_{k,0}^{(i)} := V_k^{(i)} / \big(\sum_{j=1}^N V_k^{(j)} + \epsilon\big)$.

(ii) Sample $L_k \sim \mathbf{p}$ (independently from the other random variables).

(iii) Run ΔPF (Algorithm 2) with $(\check{M}_{0:n}^{(\Theta_k, L_k)}, \check{G}_{0:n}^{(\Theta_k)}, N)$, and call the output $(V_{k,L_k}^{(1:2N)}, \mathbf{X}_{k,L_k}^{(1:2N)})$. Set $W_{k,L_k}^{(i)} := V_{k,L_k}^{(i)} / \big[p_{L_k}\big(\sum_{j=1}^N V_k^{(j)} + \epsilon\big)\big]$.

Report the estimator

$$E_{m_{\text{iter}},N,\mathbf{p}}(f) := \frac{\sum_{k=1}^{m_{\text{iter}}} \big[\sum_{i=1}^N W_{k,0}^{(i)} f(\Theta_k, \mathbf{X}_{k,0}^{(i)}) + \sum_{i=1}^{2N} W_{k,L_k}^{(i)} f(\Theta_k, \mathbf{X}_{k,L_k}^{(i)})\big]}{\sum_{k=1}^{m_{\text{iter}}} \big[\sum_{i=1}^N W_{k,0}^{(i)} + \sum_{i=1}^{2N} W_{k,L_k}^{(i)}\big]} \approx \pi^{(\infty)}(f).$$

(ii) It is only necessary to implement PMMH for the coarsest level. This is typically relatively cheap, and therefore allows for a relatively long MCMC run. Consequently, relative cost of burn-in is small, and if the proposal q is adapted [cf. 1], it has time to converge.

(iii) The (potentially costly) rΔPFs are applied independently for each Θ_k, which allows for efficient parallelisation.

(iv) We suggest that the number of particles 'N_0' used in the PMMH be chosen based on [9, 29], while the number of particles 'N_ℓ' (and p_ℓ) can be optimised for each level ℓ based on Recommendation 1 of Section 5, or kept constant. One can also afford to increase the number of particles when a 'jump chain' representation is used (see the following remark).

(v) The rΔPF corrections may be calculated only once for each accepted state [32]. That is, suppose $(\tilde{\Theta}_k, \tilde{V}_k^{(1:N)}, \tilde{\mathbf{X}}_k^{(1:N)})_{k=1}^{m_{\text{iter}}^{\text{jump}}}$ are the accepted states, $(D_k)_{k=1}^{m_{\text{iter}}^{\text{jump}}}$ are the corresponding holding times, and $(\tilde{V}_{k,L_k}^{(1:2N_{L_k})}, \tilde{\mathbf{X}}_{k,L_k}^{(1:2N_{L_k})})_{k=1}^{m_{\text{iter}}^{\text{jump}}}$ are corresponding ΔPF outputs, then the estimator is formed as in Algorithm 4 using $(\tilde{\Theta}_k, \tilde{V}_k^{(1:N)}, \tilde{\mathbf{X}}_k^{(1:N)})$, and accounting for the holding times in the weights defined as $W_k^{(i)} := D_k \tilde{V}_k^{(i)} / \big(\sum_{j=1}^N \tilde{V}_k^{(j)} + \epsilon\big)$ and $W_{k,L_k}^{(i)} := \tilde{V}_{k,L_k}^{(i)} / \big[p_{L_k}\big(\sum_{j=1}^N \tilde{V}_k^{(j)} + \epsilon\big)\big]$.

(vi) In case the Markov chain in (P1) phase is slow mixing, (further) thinning may be applied (to the jump chain) before the (P2) phase.

(vii) In practice, Algorithm 4 may be implemented in an on-line fashion w.r.t. the number of iterations m_{iter}, and by progressively refining the estimator $E_{m_{\text{iter}},N,\mathbf{p}}(f)$. The r$\Delta$PF corrections may be calculated in parallel with the Markov chain.

(viii) In Algorithm 4, the rΔPFs depend only on Θ_k. They could depend also on $V_k^{(i)}$ and $\mathbf{X}_k^{(i)}$, but it is not clear how such dependence could be used in practice to achieve better

performance. Likewise, the 'zeroth level' estimate in Algorithm 4 is based solely on particles in (P1), but it could also be based on (additional) new particle filter output.

(ix) In order to save memory, it is possible also to 'subsample' only one trajectory $\mathbf{X}_k^*$ from $\mathbf{X}_k^{(1:N)}$, such that $\mathbb{P}[\mathbf{X}_k^* = \mathbf{X}_k^{(i)}] \propto V_k^{(i)}$, and set $W_{k,0}^* := \sum_{i=1}^N W_{k,0}^{(i)}$, and similarly in Algorithm 2 find $\check{\mathbf{X}}^*$ such that $\mathbb{P}[\check{\mathbf{X}}^* = \check{\mathbf{X}}^{(i)}] \propto \check{V}^{(i)}$, setting $\mathbf{X}_{k,L_k}^{*(1:2)} := \check{\mathbf{X}}^*$, and defining from the usual output of Algorithm 2, $W_{k,L_k}^{*(1)} := \sum_{i=1}^N W_{k,L_k}^{(i)}$ and $W_{k,L_k}^{*(2)} := \sum_{i=N+1}^{2N} W_{k,L_k}^{(i)}$. The subsampling output estimator then takes the form,

$$E_{m_{\mathrm{iter}},N,\mathbf{p}}^{\mathrm{subsample}}(f) := \frac{\sum_{k=1}^{m_{\mathrm{iter}}} \left[W_{k,0}^* f(\Theta_k, \mathbf{X}_k^*) + \sum_{i=1}^2 W_{k,L_k}^{*(i)} f(\Theta_k, \mathbf{X}_{k,L_k}^{*(i)})\right]}{\sum_{k=1}^{m_{\mathrm{iter}}} \left[W_{k,0}^* + \sum_{i=1}^2 W_{k,L_k}^{*(i)}\right]}.$$

Note, however, that the asymptotic variance of this estimator is higher, because $E_{m_{\mathrm{iter}},N,\mathbf{p}}(f)$ may be viewed as a Rao-Blackwellised version of $E_{m_{\mathrm{iter}},N,\mathbf{p}}^{\mathrm{subsample}}(f)$.

4.2. **Consistency and central limit theorem.**

Theorem 11. *Assume that the algorithm constant $\epsilon \geq 0$ is chosen positive, and that the Markov chain $(\Theta_k, X_k^{(1:N)}, V_k^{(1:N)})_{k\geq 1}$ is ψ-irreducible, and that $\pi^{(0)}(f)$ and $\pi^{(\infty)}(f)$ are finite. For each $\theta \in \mathsf{T}$, suppose Assumption 5 holds for $g \equiv 1$ and $g = f^{(\theta)} := f(\theta, \cdot)$, with $M_{0:n}^{(\ell)} := M_{0:n}^{(\theta,\ell)}$ and $G_{0:n}^{(\ell)} := G_{0:n}^{(\theta)}$. Assume*

$$\int \mathrm{pr}(\theta)\big(\sqrt{s_1(\theta)} + \sqrt{s_{f^{(\theta)}}(\theta)}\big)\mathrm{d}\theta < \infty.$$

Then, the estimator of Algorithm 4 is strongly consistent:

$$E_{m_{\mathrm{iter}},N,\mathbf{p}}(f) \xrightarrow{m_{\mathrm{iter}}\to\infty} \int \pi^{(\infty)}(\mathrm{d}\theta, \mathrm{d}x_{0:n}) f(\theta, x_{0:n}) \qquad \textit{(a.s.)}$$

Remark 12. Regarding Theorem 11, whose proof is given in Appendix B:

(i) If all potentials G_t are strictly positive, the algorithm constant ϵ may be taken to be zero. However, if $\epsilon = 0$ and Algorithm 1 with $(M_{0:n}^{(\hat{\Theta}_k,0)}, G_{0:n}^{(\hat{\Theta}_k)}, N)$ can produce an estimate with $\sum_{i=1}^N V^{(i)} = 0$ with positive probability, the consistency may be lost [32].

Proposition 13. *Suppose that the conditions of Theorem 11 hold. Suppose additionally that $\pi^{(\infty)}(f^2) < \infty$ and that the base chain $(\Theta_k, V_k^{(1:N)}, \mathbf{X}_k^{(1:N)})_{k\geq 1}$ is aperiodic, with transition probability denoted by P. Then,*

$$\sqrt{m_{\mathrm{iter}}}\big[E_{m_{\mathrm{iter}},N,\mathbf{p}}(f) - \pi^{(\infty)}(f)\big] \xrightarrow{m_{\mathrm{iter}}\to\infty} \mathcal{N}(0, \sigma^2), \qquad \textit{in distribution,}$$

whenever the asymptotic variance

$$\sigma^2 = \frac{\mathrm{var}(P, \mu_{\bar{f}}) + \Pi(\sigma_\xi^2)}{c^2} \tag{9}$$

is finite. Here, $\bar{f} := f - \pi^{(\infty)}(f)$, $c > 0$ is a constant (equal to $\Pi(\mu_1)$), and

$$\begin{aligned}\sigma_\xi^2(\theta, v^{(1:N)}, \mathbf{x}^{(1:N)}) &:= \mathrm{var}\big(\xi_k(\bar{f}) \mid \Theta_k = \theta, V_k^{(1:N)} = v^{(1:N)}, \mathbf{X}_k^{(1:N)} = \mathbf{x}^{(1:N)}\big) \\ &= \frac{s_{\bar{f}^{(\theta)}}(\theta) - \big(\gamma_n^{(\theta,\infty)}(G_n \bar{f}^{(\theta)}) - \gamma_n^{(\theta,0)}(G_n \bar{f}^{(\theta)})\big)^2}{\big(\sum_{i=1}^N v^{(i)} + \epsilon\big)^2}.\end{aligned}$$

Remark 14. Proposition 13 follows from [32, Theorem 7]. We suggest that $N = N_0$ for (P1) be chosen based on [9, 29] to minimise $\mathrm{var}(P, \mu_{\bar{f}})$, and that (p_ℓ) and $N = N_\ell$ in (P2) for the rΔPF be chosen as in Recommendation 1 of Section 5, to minimise σ_ξ^2, subject to cost constraints, in order to jointly minimise σ^2. However, the question of the optimal choice for N_0 in the IS context is not yet settled.

5. Asymptotic efficiency and randomised multilevel considerations

We summarise the results of this section by suggesting the following safe allocations for probability $\mathbf{p} = (p_\ell)_{\ell\in\mathbb{N}}$ and number $N = N_\ell$ of particles at level $\ell \geq 1$ in the ΔPF (Algorithm 2) used in Algorithm 3 and 4, and Proposition 7, with β given in (8) in the HMM diffusion context of Section 3, or, indeed, with β given in the abstract framework under Assumption 18 given later. See also Figure 1 for the recommended allocations.

Recommendation 1. With strong error convergence rate β given in (8), we suggest the following for $\mathbf{p} = (p_\ell)_{\ell\in\mathbb{N}}$ and $N_\ell \in \mathbb{N}$ in ΔPF (Algorithm 2):
($\beta = 1$) (e.g. Euler scheme). Choose $p_\ell \propto (\frac{1}{2})^\ell$ and $N_\ell \propto 1$ constant.
($\beta = 2$) (e.g. Milstein scheme). Choose $p_\ell \propto 2^{-1.5\ell} \approx (\frac{1}{3})^\ell$ and $N_\ell \propto 1$ constant.

The suggestions are based on Corollary 9 of Section 3, and Propositions 20 ($\beta = 2$) and 26 ($\beta = 1$) given below (with weak convergence rate $\alpha = 1$; see Figure 1 for general α). In the Euler case, although the theory below gives the same computational complexity order by choosing any $\rho \in [0, 1]$ and setting $p_\ell \propto 2^{-(1+\rho)\ell}$ and $N_\ell \propto 2^{\rho\ell}$, the experiment in Section 6 gave a better result using simply $\rho = 0$, corresponding to no scaling.

5.1. Efficiency framework.

The asymptotic efficiency of Monte Carlo was considered theoretically in [15]; see [14] in the dMLMC context. The developments of this section follow [26] for rMLMC (originally in the i.i.d. setting), while also giving some extensions (also applicable to that setting). We will see that the basic rMLMC results carry over to our setting involving MCMC and randomised estimators based on PF outputs. Proofs are given in Appendix C.

We are interested in modeling the computational costs involved in running Algorithm 4; the algorithm of Proposition 7 is recovered with $\mathsf{T} = \{\theta\}$. Let τ_{Θ_k, L_k} represent the combined *cost at iteration* k of the base Markov chain and weight calculation in Algorithm 4, so that the *total cost* $\mathscr{C}(m)$ of Algorithm 4 with m iterations is

$$\mathscr{C}(m) := \sum_{k=1}^{m} \tau_{\Theta_k, L_k}.$$

The following assumption seems natural in our setting.

Assumption 15. For $\Theta_k \in \mathsf{T}$, a family $\{\tau_{\Theta_k,\ell}\}_{k,\ell\geq 1}$ consists of positive-valued random variables that are independent of $\{L_k\}_{k\geq 1}$, where $L_k \sim \mathbf{p}$ i.i.d., and that are conditionally independent given $\{\Theta_k\}_{k\geq 1}$, such that $\tau_{\Theta_k,\ell}$ depends only on $\Theta_k \in \mathsf{T}$ and $\ell \in \mathbb{N}$.

Under a budget constraint $\kappa > 0$, the *realised length of the chain* is $\mathscr{L}(\kappa)$ iterations, where

$$\mathscr{L}(\kappa) := \max\{m \geq 1 : \mathscr{C}(m) \leq \kappa\}.$$

Under a budget constraint, the CLT of Proposition 13 takes the following altered form, where here $\Pi_m(\mathrm{d}\theta)$ denotes the θ-marginal of the invariant probability measure (given as (25) in Appendix B) of the base Markov chain (equal to the θ-marginal posterior of the $\ell = 0$ model).

Proposition 16. *If the assumptions of Theorem 13 hold with* $\sigma^2 < \infty$*, and if* $\mathbb{E}[\boldsymbol{\tau}] := \mathbb{E}_{\Pi_m \otimes \mathbf{p}}[\boldsymbol{\tau}] < \infty$ *with* $\boldsymbol{\tau}(\theta, \ell) := \mathbb{E}[\tau_{\Theta_k, L_k} | \Theta_k = \theta, L_k = \ell]$*, then*

$$\sqrt{\kappa}\big[E_{\mathscr{L}(\kappa),N,\mathbf{p}}(f) - \pi^{(\infty)}(f)\big] \xrightarrow{\kappa\to\infty} \mathcal{N}(0, \mathbb{E}[\boldsymbol{\tau}]\sigma^2), \qquad \text{in distribution.} \tag{10}$$

Remark 17. The quantity $\mathbb{E}[\boldsymbol{\tau}]\sigma^2$ is called the 'inverse relative efficiency' by [15], and is considered a more accurate quantity than the asymptotic variance (σ^2 here) for comparison of Monte Carlo algorithms run on the same computer, as it takes into account also the average computational time.

In the following we consider (possibly) variance reduced (if $\rho > 0$) versions of $\Delta_\ell(g)$ of Assumption 5, denoted Δ_ℓ, where $g = f^{(\theta)}$, based on running the ΔPF (Algorithm 2) with parameters θ, ℓ fixed. The constant $C < \infty$ may change line-to-line, but does not depend on N, ℓ, or θ, but may depend on the time-horizon n and function f.

Assumption 18. Assumption 15 holds, and constants $2\alpha \geq \beta > 0$, $\gamma > 0$, and $\rho \geq 0$ are such that the following hold:

(i) (Mean cost) $\mathbb{E}[\tau_{\theta,\ell}] \leq C 2^{\gamma\ell(1+\rho)}$
(ii) (Strong order) $\mathbb{E}[\Delta_\ell^2] \leq C 2^{-\ell(\beta+\rho)} + C 2^{-2\alpha\ell}$
(iii) (Weak order) $|\mathbb{E}\Delta_\ell| \leq C 2^{-\alpha\ell}$

Remark 19. Regarding Assumption 18:

(i) We only assume bounded mean cost in (i), rather than the almost sure cost bound commonly used. This generalisation allows for the setting where occasional algorithmic runs may take a long time.
(ii) In the original MLMC setting, the cost scaling γ in (i) is taken to be $\gamma = 1$ [14, 26]. However, in settings involving uncertainty quantification, and where the forward solver may involve non-sparse matrix inversions, often $\gamma \geq 1$ [6, 18, 20].
(iii) We assume in (i) that the mean cost to form Δ_ℓ is bounded by the γ-scaled product of the number of samples or particles N_ℓ times the number of Euler time steps $2^\ell + 2^{\ell-1}$ together with the $O(N_\ell)$-resampling cost, where there are $N_\ell \propto 2^{\rho\ell}$ particles at level ℓ. Here, we recall that the stratified, systematic, and residual resampling algorithms have $O(N_\ell)$ cost, as does an improved implementation of multinomial resampling [cf. 5, 8].
(iv) With $\rho = 0$, by Jensen's inequality one sees why $\alpha \geq \beta/2$ can be assumed, and that (ii) becomes $\mathbb{E}\Delta_\ell^2 \leq C 2^{-\ell\beta}$.
(v) $\rho \geq 0$ in (i) and (ii) corresponds to using an average of $N_\ell := \lceil 2^{\rho\ell} \rceil$ i.i.d samples of $\Delta_\ell^{(1)}$, i.e. $\Delta_\ell = \frac{1}{N_\ell} \sum_{i=1}^{N_\ell} \Delta_\ell^{(i)}$, or, of more present interest to us, to increasing the number of particles used in a PF by a factor of N_ℓ instead of the default lower number. The former leads to $\mathbb{E}\Delta_\ell^2 = \frac{1}{N_\ell}\mathrm{var}(\Delta_\ell^{(1)}) + \mathbb{E}[\Delta_\ell^{(1)}]^2$, justifying (ii), as does Corollary 9, with $\beta \in \{1, 2\}$ and $\alpha = 1$, for the ΔPF (Algorithm 2) in the HMM diffusion context (Section 3).

Proposition 20. *Suppose Assumption 18 and the assumptions of Proposition 13 hold, with* $\mathrm{var}(P, \mu_{\bar{f}}) < \infty$*. If* $p_\ell \propto 2^{-r\ell}$ *for some* $r \in \big(\gamma(1+\rho), \min(\beta+\rho, 2\alpha)\big)$*, then (10) holds, i.e.*

$$\sqrt{\kappa}\big[E_{\mathscr{L}(\kappa),N,\mathbf{p}}(f) - \pi^{(\infty)}(f)\big] \xrightarrow{\kappa\to\infty} \mathcal{N}(0, \mathbb{E}[\boldsymbol{\tau}]\sigma^2), \qquad \text{in distribution.}$$

Remark 21. Regarding Proposition 20, in the common case $\gamma = 1$ for simplicity:

(i) If $\beta > 1$ ('canonical convergence regime') and $\rho = 0$, then a choice for $r \in (1, \beta)$ exists. See also [26, Theorem 4] for a discussion of the theoretically optimal $\mathbf{p}$.
(ii) If $\beta \leq 1$ ('subcanonical convergence regime'), then $\beta + \rho \leq 1 + \rho$ and so no choice for r exists.

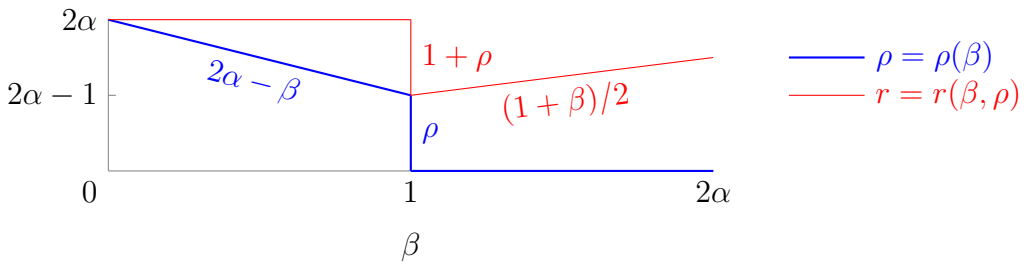


FIGURE 1. Recommendations for number of particles $N_\ell \propto 2^{\rho\ell}$ and probability $p_\ell \propto 2^{-r\ell}$. Here, $\gamma = 1$ always, and $\rho \in [0, 2\alpha - 1]$ when $\beta = 1$ provides a line of choices with the same *order* of computational complexity. In our particular experiment in Section 6, however, the simple choice $\rho = 0$, corresponding to no scaling in the particles, will turn out to be optimal.

5.2. **Subcanonical convergence.** When $\beta > 1$, within the framework above we have seen that a canonical convergence rate holds (Proposition 20) because $\mathbb{E}[\boldsymbol{\tau}] < \infty$ and $\sigma^2 < \infty$. When $\beta \leq 1$, this is no longer the case, and one must choose between a finite asymptotic variance and infinite expected cost, or vice versa. Assuming the former, and that a CLT holds (Proposition 13), for $\epsilon > 0$ and $0 < \delta < 1$ the Chebyshev inequality implies that the number of iterations of Algorithm 4 so that

$$\mathbb{P}[|E_{m,N,\mathbf{p}}(f) - \pi^{(\infty)}(f)| \leq \epsilon] \geq 1 - \delta \tag{11}$$

holds implies that m must be of the order $O(\epsilon^{-2})$. The question is then how to minimise the total cost $\mathscr{C}(m)$, or *computational complexity*, involved in obtaining the m samples. This will involve optimising for (p_ℓ) and N_ℓ to minimise $\mathscr{C}(m)$, while keeping the asymptotic variance finite.

Proposition 22. *Suppose that the assumptions of Proposition 13 hold with $\sigma^2 < \infty$, and Assumption 18 holds with $\mathbb{E}[\tau_{\Theta_{k_0}, L_{k_0}}] = \infty$ for some $k_0 \geq 1$. If*

$$\sum_{k \geq 1} \sup_{j \geq 1} \mathbb{P}[\tau_{\Theta_j, L_j} > a_k] < \infty$$

with $a_k = O\big(k^{c_1} (\log_2 k)^{c_0}\big)$ for some constants $c_0 > 0$ and $c_1 \geq 1$, then (11) can be obtained with computational complexity

$$O\big(\epsilon^{-2c_1} |\log_2 \epsilon|^{c_0}\big) \qquad \text{as} \quad \epsilon \to 0.$$

Remark 23. The above result shows that even for costs with unbounded tails, reasonable confidence intervals and complexity order may be possible. This may be the case for example when a rejection sampler or adaptive resampling mechanism is used within Algorithm 1 or 4, which may lead to large costs for some Θ_k, for example a cost with a geometric tail.

The next results are as in [26, Proposition 4 and 5] in the standard rMLMC setting, and shows how one can choose $\mathbf{p}$, assuming an additional almost sure cost bound, so that $\sigma^2 < \infty$, with reasonable complexity.

Proposition 24. *Suppose that the assumptions of Proposition 13 hold with $\mathrm{var}(P, \mu_{\bar{f}}) < \infty$, and that Assumption 18 holds with $\beta \leq 1$, where moreover $\tau_{\theta,\ell} \leq C 2^{\gamma\ell(1+\rho)}$ almost surely, uniformly in $\Theta_k = \theta \in \mathsf{T}$. For all $q > 2$ and $\eta > 1$, the choice of probability*

$$p_\ell \propto 2^{-2b\ell} \ell [\log_2(\ell + 1)]^\eta$$

where $b := \min((\beta+\rho)/2, \alpha)$*, leads to* $\sigma^2 < \infty$*, and* (11) *can be obtained with computational complexity*

$$O\Big(\epsilon^{-\gamma\frac{(1+\rho)}{b}}|\log_2 \epsilon|^{q\gamma\frac{(1+\rho)}{2b}}\Big) \qquad as \quad \epsilon \to 0.$$

Remark 25. Regarding Proposition 24, with $\gamma = 1$:

(i) Under Assumption 18 with $\rho = 0$, the usual setup in MLMC before variance reduced estimators are used, the above proposition shows that finite variance and (11) can be obtained without increasing the number of particles at the higher levels, even in the subcanonical regime. We have in this case $b = \beta/2 \leq \alpha$ and complexity $O\Big(\epsilon^{-\frac{2}{\beta}}|\log_2 \epsilon|^{\frac{q}{\beta}}\Big)$. When $\beta = 1$ (borderline case), dMLMC gives complexity $O(\epsilon^{-2}|\log_2 \epsilon|^2)$ [14, 18], which is negligibly better (recall $q > 2$), but is biased inference.

(ii) When $\alpha > \beta/2$, which is the usual case in the subcanonical regime ($\beta \leq 1$) [cf. 22], a more efficient use of resources can be obtained by increasing the number of particles (see Proposition 26 below).

Proposition 26. *Suppose the assumptions of Proposition 24 hold, where moreover* $\rho \geq 0$ *may vary as a free parameter without changing the constant* $C > 0$*. Then, for all* $q > 2$*,* $\eta > 1$ *constants, the choice* $\rho = 2\alpha - \beta$ *and probability*

$$p_\ell \propto 2^{-2\alpha\ell}\ell[\log_2(\ell+1)]^\eta$$

leads to $\sigma^2 < \infty$*, and* (11) *can be obtained with computational complexity*

$$O\Big(\epsilon^{\gamma[-2-\frac{(1-\beta)}{\alpha}]}|\log_2 \epsilon|^{\gamma[q+\frac{(1-\beta)}{2\alpha}]}\Big) \qquad as \quad \epsilon \to 0.$$

6. Numerical simulations

Now the theoretical results relating to the method herein introduced will be demonstrated on two examples. We will consider one example in the canonical regime, and one in the sub-canonical, both of which have likelihoods that can be computed exactly, so that the ground truth $\pi^{(\infty)}(f)$ can be easily calculated to arbitrary precision. We run each example with 100 independent replications, and calculate the MSE when the chain is at length m as

$$\mathrm{MSE}(m) = \frac{1}{100}\sum_{i=1}^{100}\left|E^{(i)}_{m,N,\mathbf{p}}(f) - \pi^{(\infty)}(f)\right|^2,$$

which is depicted as the thick red line, average of the thin lines, in Figure 2 below. The error decays with the optimal rate of cost^{-1} and $\log(\mathrm{cost})\mathrm{cost}^{-1}$ in the canonical and sub-canonical cases, respectively, where cost is the realised cost of the run, $\mathscr{C}(m)$ from Section 5, measured in seconds, with m iterations of the Markov chain.

6.1. Ornstein–Uhlenbeck process.

Consider the Ornstein–Uhlenbeck (OU) process

$$\mathrm{d}X_t = -aX_t\mathrm{d}t + b\mathrm{d}W_t\,, \qquad t \geq 0, \tag{12}$$

with initial condition $x_0 = 0$, model parameter $\theta = (\theta_1, \theta_2) \sim N(0, \sigma^2 I)$, and $a := a_\theta = \exp(\theta_1)$ and $b := b_\theta = \exp(\theta_2)$. The process is discretely observed for $k = 1, \ldots, n$,

$$Y_k = X_k + \xi_k\,, \tag{13}$$

where $\xi_k \sim N(0, \gamma^2)$ i.i.d. Therefore,

$$G_k(x) = \exp(-\frac{1}{2\gamma^2}|x - y_k|^2)\,.$$

The marginal likelihood is given by

$$\mathbb{P}[y_{1:n}|\theta] = \prod_{k=1}^{n} \mathbb{P}[y_k|y_{1:k-1},\theta]\,,$$

and each factor can be computed as the marginal of the joint on the prediction and current observation, i.e.

$$\mathbb{P}[y_k|y_{1:k-1},\theta] = \int_{\mathbb{R}} \mathbb{P}[y_k|x_k,\theta]\mathbb{P}[x_k|y_{1:k-1},\theta]\mathrm{d}x_k\,. \tag{14}$$

In this example the ground truth can be computed exactly via the Kalman filter. In particular, the solution of (12) is given by

$$X_1 = e^{-a}X_0 + W_1\,, \quad W_1 \sim \mathcal{N}\left(0, \frac{b^2}{2a}(1-e^{-2a})\right)\,.$$

The filter at time k is given by the following simple recursion

$$m_k = c_k\left(\frac{y_k}{\gamma^2} + \frac{\hat{m}_k}{\hat{c}_k}\right)\,, \quad c_k = (\gamma^{-2}+\hat{c}_k^{-1})^{-1}\,, \quad \hat{m}_k = e^{-a}m_{k-1}\,, \quad \hat{c}_k = e^{-2a}c_{k-1}+\frac{b^2}{2a}(1-e^{-2a})\,.$$

Additionally, the incremental marginal likelihoods (14) can be computed exactly

$$\mathbb{P}[y_k|y_{1:k-1},\theta] = \sqrt{\frac{c_k}{2\pi\hat{c}_k\gamma^2}}\exp\left\{-\frac{1}{2}\left[\frac{y_k^2}{\gamma^2} + \frac{\hat{m}_k^2}{\hat{c}_k} - c_k\left(\frac{y_k}{\gamma^2}+\frac{\hat{m}_k}{\hat{c}_k}\right)^2\right]\right\}\,.$$

The parameters are chosen as $\gamma = 1$, $\sigma^2 = 0.1$, $n = 5$, and the data is generated with $\theta = (0,0)^T$. Our aim is to compute $\mathbb{E}(\theta|y_{1:n})$ (or $\mathbb{E}[(a,b)^T|y_{1:n}]$, etc., but we will content ourselves with the former). This is done via a brute force random walk MCMC for $m = 10^8$ steps using the exact likelihood $\mathbb{P}[y_{1:n}|\theta]$ as above. The IACT is around 10, so this gives a healthy limit for MSE computations.

For the numerical experiment, we use Euler-Maruyama method at resolution $h_\ell = 2^{-\ell}$ to solve (12) as follows

$$X_{k+1} = (1-ah_\ell)X_k + bB_{k+1}\,,\ B_{k+1} \sim \mathcal{N}(0,h_\ell)\ i.i.d. \tag{15}$$

for $k = 1,\ldots,K_\ell = h_\ell^{-1}$. Levels ℓ and $\ell-1$ are coupled in the simulation of Δ_ℓ by defining $B^C_{1:K_\ell/2} = B^F_{1:2:K_\ell-1} + B^F_{2:2:K_\ell}$ Algorithm 2 is then run using the standard bootstrap particle filter (Algorithm 1) with $N = 20$ particles and $O(N)$-complexity multinomial resampling [cf. 5]. Theorem 8 provides a rate of $\beta = 2$ for Algorithm 2, because the diffusion coefficient is constant, which implies we are essentially running a Milstein scheme (cf. (8) and [22]). Recommendation 1 (or Proposition 20) of Section 5 suggests arbitrary precision can be obtained by Algorithm 4 with $p_\ell \propto 2^{-3\ell/2}$ and no scaling of particle numbers based on ℓ in this canonical $\beta = 2$ regime (with weak rate $\alpha = 1$). We choose a positive PMMH algorithm constant $\epsilon = 10^{-6}$ (cf. Remark 10(i)). We run Algorithm 4 for 10^4 steps, with 100 replications. The results are presented in Figure 2, where it is clear that the theory holds and the MSE decays according to 1/cost. The variance of the run-times is very small over replications.

6.2. **Geometric Brownian motion.** We next consider the following stochastic differential equation

$$\mathrm{d}X_t = aX_t\mathrm{d}W_t, \tag{16}$$

with initial condition $x_0 = 1$, and $a := a_\theta = \exp(\theta)$ with $\theta \sim \mathcal{N}(0,\sigma^2)$. This equation is

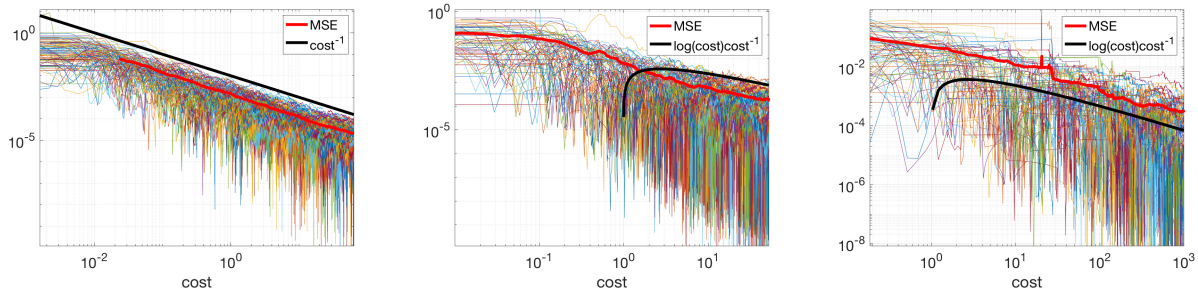


FIGURE 2. The MSE of PMMH rMLMC IS Algorithm 4 applied to the problem of parameter inference for the discretely observed OU process (left plot) and GBM process (middle plot with $\rho = 0$, right plot with $\rho = 1$). Replications are given by the thin curves, while the bold curves give the average MSE over replications as well as the lines cost^{-1} (left plot) and $\log(\text{cost})\text{cost}^{-1}$ (middle and right plots) to guide the eye.

analytically tractable as well, and the solution of the transformed equation $Z = \log X$ is given via Itô's formula by

$$\mathrm{d}Z_t = -\frac{a^2}{2}\mathrm{d}t + a\,\mathrm{d}W_t.$$

Defining $W_k \sim \mathcal{N}(0,1)$ i.i.d., one has that

$$Z_{k+1} = Z_k + -\frac{a^2}{2} + a\,W_k\,, \quad \text{with} \quad z_0 = \log x_0 = 0,$$

and the solution of (16) can be obtained via exponentiation: $X_k = e^{Z_k}$. Moreover, noisy observations are introduced on the form

$$Y_k = \log(X_k) + \xi_k = Z_k + \xi_k\,,$$

where $\xi_k \sim \mathcal{N}(0, \gamma^2)$ i.i.d. as above. Therefore we have

$$G_k(z) = \exp(-\frac{1}{2\gamma^2}|z - y_k|^2)\,.$$

Again $\mathbb{P}[y_{1:n}|\theta]$ can be computed analytically by integrating over $(z_1, \ldots, z_n)$.

In order to investigate the theoretical sub-canonical rate, we return to (16) and approximate this directly using Euler-Maruyama method (15), which introduces artificial approximation error. This problem suffers from stability problems when $X < 0$, so we take $h_\ell = 2^{-5-\ell}$. Algorithm 1 is then used along with the selection functions

$$G_k(x) = \exp(-\frac{1}{2\gamma^2}|\log(x) - y_k|^2)\,.$$

Here the diffusion coefficient is not constant, and Euler-Maruyama method yields a rate of $\beta = 1 = \alpha$, the borderline case, which is expected to give a logarithmic penalty. Based on Recommendation 1 (or Proposition 26) of Section 5, we consider scaling the particles as $2^{\rho\ell}$ with $\rho = 2\alpha - \beta = 1$ and $\rho = 0$, with $p_\ell \propto 2^{-2\ell}\ell\log(\ell)^2$ in both cases. Again we let $\epsilon = 10^{-6}$. Again the standard bootstrap particle filter is used, with $N = 20 \times 2^{\rho\ell}$ particles. Algorithm 4 is run for 10^5 steps, with 100 replications. The results are presented in Figure 2, and they show good agreement with the theory, in terms of rate. On the other hand, the cost for $\rho = 0$ is apparently smaller than that of $\rho = 1$ by a factor of approximately 100.

Acknowledgments

The authors have received support from the Academy of Finland (274740, 312605, 315619), as well as from the Institute for Mathematical Sciences, Singapore, during the 2018 programme 'Bayesian Computation for High-Dimensional Statistical Models.' In addition, AJ has received support from the Singapore Ministry of Education (R-155-000-161-112) and KL from the University of Manchester (School of Mathematics).

Appendix A. Analysis of the delta particle filter

We now give our analysis that is required for the proofs of Theorem 8 and Corollary 9 of Section 3 regarding the ΔPF (Algorithm 2) for HMM diffusions. The structure of the appendix is as follows. In Section A.1 we introduce some more Feynman–Kac notations, following [7, 18], emphasising that here we consider standard HMMs that can be coupled. In Section A.2 we recall the ΔPF stated earlier. A general variance bound for quantities such as $\Delta_\ell(\varphi)$ is given in Section A.3. This is particularised to the HMM diffusion case in Section A.4, where we supply the proofs for the results of Section 3.

A.1. **Models.** Let $(\mathsf{X}, \mathcal{X})$ be a measurable space and $\{G_n\}_{n\geq 0}$ a sequence of non-negative, bounded and measurable functions such that $G_n : \mathsf{X} \to \mathbb{R}_+$. Let $\eta_0^F, \eta_0^C \in \mathscr{P}(\mathsf{X})$ and $\{M_n^F\}_{n\geq 1}$, $\{M_n^C\}_{n\geq 1}$ be two sequences of Markov kernels, i.e. $M_n^F : \mathsf{X} \to \mathscr{P}(\mathsf{X})$, $M_n^C : \mathsf{X} \to \mathscr{P}(\mathsf{X})$. Set $\mathsf{E}_n := \mathsf{X}^{n+1}$ for $n \geq 0$, and for $x_{0:n} \in \mathsf{E}_n$,

$$\boldsymbol{G}_n(x_{0:n}) = G_n(x_n)$$

and for $n \geq 1$, $s \in \{F, C\}$, $x_{0:n-1} \in \mathsf{E}_{n-1}$

$$\boldsymbol{M}_n^s(x_{0:n-1}, \mathrm{d}x'_{0:n}) = \delta_{\{x_{0:n-1}\}}(\mathrm{d}x'_{0:n-1}) M_n^s(x'_{n-1}, \mathrm{d}x'_n).$$

Define for $s \in \{F, C\}$, $\varphi \in \mathcal{B}_b(\mathsf{E}_n)$, $u_n \in \mathsf{E}_n$

$$\boldsymbol{\gamma}_n^s(\varphi) = \int_{\mathsf{E}_0 \times \cdots \times \mathsf{E}_n} \varphi(u_n) \Big(\prod_{p=0}^{n-1} \boldsymbol{G}_p^s(u_p) \Big) \eta_0^s(\mathrm{d}u_0) \prod_{p=1}^{n} \boldsymbol{M}_p^s(u_{p-1}, \mathrm{d}u_p)$$

and

$$\boldsymbol{\eta}_n^s(\varphi) = \frac{\boldsymbol{\gamma}_n^s(\varphi)}{\boldsymbol{\gamma}_n^s(1)}.$$

Throughout this appendix, we assume (D), and that Assumption 2(i) holds, i.e. there exists $\check{\eta}_0 \in \mathscr{P}(\mathsf{X} \times \mathsf{X})$ such that for any $A \in \mathcal{X}$

$$\check{\eta}_0(A \times \mathsf{X}) = \eta_0^F(A) \qquad \check{\eta}_0(\mathsf{X} \times A) = \eta_0^C(A)$$

and moreover for any $n \geq 1$ there exists Markov kernels $\{\check{M}_n\}$, $\check{M}_n : \mathsf{X} \times \mathsf{X} \to \mathscr{P}(\mathsf{X} \times \mathsf{X})$ such that for any $A \in \mathcal{X}$, $(x, x') \in \mathsf{X} \times \mathsf{X}$:

$$\check{M}_n(A \times \mathsf{X})(x, x') = M_n^F(A)(x) \qquad \check{M}_n(\mathsf{X} \times A)(x, x') = M_n^C(A)(x'). \tag{17}$$

A.2. **Delta particle filter.** Define $x_p = (x_p^F, x_p^C) \in \mathsf{X} \times \mathsf{X}$ and

$$\check{G}_p(x_p) = \frac{1}{2}(G_p(x_p^F) + G_p(x_p^C)),$$

as in Assumption 2(ii). Set, for $n \geq 0$, $x_{0:n} \in \mathsf{X}^{2(n+1)}$

$$\check{\boldsymbol{G}}_n(x_{0:n}) = \check{G}_n(x_n)$$

and for $n \geq 1$, $x_{0:n-1} \in \mathsf{X}^{2n}$

$$\check{\boldsymbol{M}}_n(x_{0:n-1}, \mathrm{d}x'_{0:n}) = \delta_{\{x_{0:n-1}\}}(\mathrm{d}x'_{0:n-1}) \check{M}_n(x'_{n-1}, \mathrm{d}x'_n), .$$

Note that coupling assumption (17) for $\check{M}_n$ can be equivalently formulated for $\check{\boldsymbol{M}}_n$.

For $n \geq 0$, $\varphi \in \mathcal{B}_b(\mathsf{E}_n \times \mathsf{E}_n)$, $u_n \in \mathsf{E}_n \times \mathsf{E}_n$, we have

$$\check{\boldsymbol{\gamma}}_n(\varphi) = \int_{\mathsf{E}_0^2 \times \cdots \times \mathsf{E}_n^2} \varphi(u_n) \Big(\prod_{p=0}^{n-1} \check{\boldsymbol{G}}_p(u_p) \Big) \check{\eta}_0(\mathrm{d}u_0) \prod_{p=1}^{n} \check{\boldsymbol{M}}_p(u_{p-1}, \mathrm{d}u_p)$$

and

$$\check{\boldsymbol{\eta}}_n(\varphi) = \frac{\check{\boldsymbol{\gamma}}_n(\varphi)}{\check{\boldsymbol{\gamma}}_n(1)}.$$

As noted in [19] it is simple to establish that for $\varphi \in \mathcal{B}_b(\mathsf{E}_n)$, if

$$\psi(x_{0:n}) = \check{\boldsymbol{G}}_n(x_{0:n}) \Big(\varphi(x^F_{0:n}) \prod_{p=0}^{n} \frac{\boldsymbol{G}_p(x^F_{0:p})}{\check{\boldsymbol{G}}_p(x_{0:p})} - \varphi(x^C_{0:n}) \prod_{p=0}^{n} \frac{\boldsymbol{G}_p(x^C_{0:p})}{\check{\boldsymbol{G}}_p(x_{0:p})} \Big) \tag{18}$$

then

$$\check{\boldsymbol{\gamma}}_n(\psi) = \check{\boldsymbol{\gamma}}_n(1) \check{\boldsymbol{\eta}}_n(\psi) = \gamma^F_n(G_n\varphi) - \gamma^C_n(G_n\varphi). \tag{19}$$

Note

$$\check{\boldsymbol{\gamma}}_n(1) = \prod_{p=0}^{n-1} \check{\boldsymbol{\eta}}_p(\check{\boldsymbol{G}}_p).$$

In order to approximate $\check{\boldsymbol{\gamma}}_n(\psi)$ one can run the following abstract version of Algorithm 2 (recall from Section 3 that we will only consider multinomial resampling). Define for $n \geq 1$, $\mu \in \mathscr{P}(\mathsf{E}_{n-1} \times \mathsf{E}_{n-1})$, $\varphi \in \mathcal{B}_b(\mathsf{E}_n \times \mathsf{E}_n)$

$$\check{\boldsymbol{\phi}}_n(\mu)(\varphi) = \frac{\mu(\check{\boldsymbol{G}}_{n-1} \check{\boldsymbol{M}}_n(\varphi))}{\mu(\check{\boldsymbol{G}}_{n-1})}.$$

The algorithm begins by generating $u_0^i \in \mathsf{E}_0 \times \mathsf{E}_0$, $i \in \{1, \ldots, N\}$ with joint law

$$\prod_{i=1}^{N} \check{\eta}_0(\mathrm{d}u_0^i) = \prod_{i=1}^{N} \check{\boldsymbol{\eta}}_0(\mathrm{d}u_0^i).$$

Defining

$$\check{\boldsymbol{\eta}}_0^N(\mathrm{d}u_0) = \frac{1}{N} \sum_{i=1}^{N} \delta_{u_0^i}(\mathrm{d}u_0)$$

we then generate $u_1^i \in \mathsf{E}_1 \times \mathsf{E}_1$, $i \in \{1, \ldots, N\}$ with joint law

$$\prod_{i=1}^{N} \check{\boldsymbol{\phi}}_1(\check{\boldsymbol{\eta}}_0^N)(\mathrm{d}u_1^i).$$

This proceeds recursively, so the joint law of the particles up to time n is

$$\Big(\prod_{i=1}^{N} \check{\boldsymbol{\eta}}_0(\mathrm{d}u_0^i) \Big) \Big(\prod_{p=1}^{n} \prod_{i=1}^{N} \check{\boldsymbol{\phi}}_p(\check{\boldsymbol{\eta}}_{p-1}^N)(\mathrm{d}u_p^i) \Big).$$

Hence we have the estimate

$$\check{\boldsymbol{\gamma}}_n^N(\psi) = \Big(\prod_{p=0}^{n-1} \check{\boldsymbol{\eta}}_p^N(\check{\boldsymbol{G}}_p) \Big) \check{\boldsymbol{\eta}}_n^N(\psi).$$

Remark 27. Note that $\check{\boldsymbol{\gamma}}_n^N(\psi)$ corresponds to the quantity $\Delta_\ell(\varphi)$ in (6) from the ΔPF output (Algorithm 2).

A.3. **General hidden Markov model case.** Define for $p \geq 1$ the semigroup

$$\check{\boldsymbol{Q}}_p(x_{0:p-1}, \mathrm{d}x'_{0:p}) = \check{\boldsymbol{G}}_{p-1}(x_{0:p-1})\check{\boldsymbol{M}}_p(x_{0:p-1}, \mathrm{d}x'_{0:p})$$

with the definition for $0 \leq p \leq n$, $\varphi \in \mathcal{B}_b(\mathsf{E}_n \times \mathsf{E}_n)$

$$\check{\boldsymbol{Q}}_{p,n}(\varphi)(u_p) = \int \varphi(u_n) \prod_{j=p+1}^{n} \check{\boldsymbol{Q}}_j(u_{j-1}, \mathrm{d}u_j)$$

if $p = n$ clearly $\check{\boldsymbol{Q}}_{n,n}$ is the identity operator. For any $0 \leq n$, $\varphi \in \mathcal{B}_b(\mathsf{E}_n \times \mathsf{E}_n)$ we set $\check{\boldsymbol{Q}}_{-1,n}(\varphi)(u_{-1}) = 0$.

Now following [7, Chapter 7] we have the following martingale (w.r.t. the natural filtration of the particle system), $\varphi \in \mathcal{B}_b(\mathsf{E}_n \times \mathsf{E}_n)$:

$$\check{\boldsymbol{\gamma}}_n^N(\varphi) - \check{\boldsymbol{\gamma}}_n(\varphi) = \sum_{p=0}^{n} \check{\boldsymbol{\gamma}}_p^N(1)[\check{\boldsymbol{\eta}}_p^N - \check{\boldsymbol{\phi}}_p(\check{\boldsymbol{\eta}}_{p-1}^N)](\check{\boldsymbol{Q}}_{p,n}(\varphi)) \tag{20}$$

with the convention that $\check{\boldsymbol{\phi}}_p(\check{\boldsymbol{\eta}}_{p-1}^N) = \check{\boldsymbol{\eta}}_0$ if $p = 0$. The representation immediately establishes that

$$\mathbb{E}[\check{\boldsymbol{\gamma}}_n^N(\varphi)] = \check{\boldsymbol{\gamma}}_n(\varphi)$$

where the expectation is w.r.t. the law associated to the particle system. We will use the following convention that C' is a finite positive constant that does not depend upon n, N or any of the G_n, M_n^s ($s \in \{F, C\}$, $\check{M}_n$. The value of C' may change from line-to-line. Define for $0 \leq p \leq n < \infty$

$$\overline{G}_{p,n} = \prod_{q=p}^{n} \|G_q\|$$

with the convention that if $p = 0$ we write $\overline{G}_n$. We have the following result.

Proposition 28. *Suppose that $\|G_n\| < \infty$ for each $n \geq 0$. Then there exist a $C' < \infty$ such that for any $n \geq 0$, $\varphi \in \mathcal{B}_b(\mathsf{E}_n \times \mathsf{E}_n)$*

$$\mathbb{E}\Big[\Big(\check{\boldsymbol{\gamma}}_n^N(\varphi) - \check{\boldsymbol{\gamma}}_n(\varphi)\Big)^2\Big] \leq \frac{C'}{N} \sum_{p=0}^{n} \overline{G}_{p-1}^2 \mathbb{E}[\check{\boldsymbol{Q}}_{p,n}(\varphi)(u_p^1)^2].$$

Proof. Set

$$\check{\boldsymbol{S}}_{p,n}^N(\varphi) = \check{\boldsymbol{\gamma}}_p^N(1)[\check{\boldsymbol{\eta}}_p^N - \check{\boldsymbol{\phi}}_p(\check{\boldsymbol{\eta}}_{p-1}^N)](\check{\boldsymbol{Q}}_{p,n}(\varphi))$$

By (20), one can apply the Burkholder-Gundy-Davis inequality to obtain

$$\mathbb{E}\Big[\Big(\check{\boldsymbol{\gamma}}_n^N(\varphi) - \check{\boldsymbol{\gamma}}_n(\varphi)\Big)^2\Big] \leq C' \sum_{p=0}^{n} \mathbb{E}[\check{\boldsymbol{S}}_{p,n}^N(\varphi)^2]. \tag{21}$$

Now, we have that

$$\mathbb{E}[\check{\boldsymbol{S}}_{p,n}^N(\varphi)^2] \leq \overline{G}_{p-1}^2 \mathbb{E}[[\check{\boldsymbol{\eta}}_p^N - \check{\boldsymbol{\phi}}_p(\check{\boldsymbol{\eta}}_{p-1}^N)](\check{\boldsymbol{Q}}_{p,n}(\varphi))^2].$$

Application of the (conditional) Marcinkiewicz-Zygmund inequality yields

$$\mathbb{E}[\check{\boldsymbol{S}}_{p,n}^N(\varphi)^2] \leq \frac{C'\overline{G}_{p-1}^2}{N} \mathbb{E}\Big[\Big(\check{\boldsymbol{Q}}_{p,n}(\varphi)(u_p^1) - \check{\boldsymbol{\phi}}_p(\check{\boldsymbol{\eta}}_{p-1}^N)(\check{\boldsymbol{Q}}_{p,n}(\varphi))\Big)^2\Big].$$

After applying C_2 and Jensen inequalities, we then conclude by (21). □

A.4. **Diffusion case.** We now consider the model of Section 3, where we recall that θ is omitted from the notation. A series of technical results are given and the proofs for Theorem 8 and Corollary 9 are given at the end of this section.

We recall that the joint probability density of the observations and the unobserved diffusion at the observation times is given by

$$\prod_{p=0}^{n} G_p(x_p) Q^{(\infty)}(x_{p-1}, x_p).$$

As the true dynamics can not be simulated, in practice we work with

$$\prod_{p=0}^{n} G_p(x_p) Q^{(\ell)}(x_{p-1}, x_p).$$

Recall an (Euler) approximation scheme with discretisation $h_\ell = 2^{-\ell}$, $\ell \geq 0$. In our context then, M_n^F corresponds $Q^{(\ell)}$ ($\ell \geq 1$) and M_n^C corresponds $Q^{(\ell-1)}$. The initial distribution η_0 is simply the (Euler) kernel started at some given x_0. As noted earlier in Remark 4(i), a natural coupling of M_n^F and M_n^C (and hence of η_0) exists. As established in [18, eq. (32)] one has (uniformly in θ as (D) holds with θ independent constants) for $C' < \infty$

$$\sup_{\mathcal{A}} \sup_{x \in \mathsf{X}} |M_n^F(\varphi)(x) - M_n^C(\varphi)(x)| \leq C' h_\ell \tag{22}$$

where $\mathcal{A} = \{\varphi \in \mathcal{B}_b(\mathsf{X}) \cap \mathrm{Lip}(\mathsf{X}) : \|\varphi\| \leq 1|\}$. We also recall that (8) holds (recall (D) is assumed).

We will use $M < \infty$ to denote a constant that may change from line-to-line. It will not depend upon θ nor N, ℓ, but may depend on the time parameter or a function. The following result will be needed later on. The proof is given after the proof of Lemma 30 below.

Proposition 29. *Assume (A1 (i)-(ii),2). Then for any $n \geq 0$ and $\varphi \in \mathcal{B}_b(\mathsf{X}^{n+1}) \cap \mathrm{Lip}(\mathsf{X}^{n+1})$ there exists a $M < \infty$ such that*

$$|\gamma_n^F(G_n\varphi) - \gamma_n^C(G_n\varphi)| \leq M h_\ell$$

We write expectations w.r.t. the time-inhomogeneous Markov chain associated to the sequence of kernels $(M_p^F)_{p\geq 1}$ (resp. $(M_p^C)_{p\geq 1}$) as $\mathbb{E}^F$, (resp. $\mathbb{E}^C$).

Lemma 30. *Assume (A1(i)-(ii),2). Let $s \in \{F, C\}$ and $\varphi \in \mathcal{B}_b(\mathsf{X}^{n+1}) \cap \mathrm{Lip}(\mathsf{X}^{n+1})$, then, define the function for $0 \leq p \leq n$*

$$\varphi_{p,n}^s(x_{0:p}) := \mathbb{E}^s[\varphi(x_{0:p}, X_{p+1:n}) \prod_{q=p+1}^{n} G_q(X_q)|x_p].$$

Then we have that $\varphi_{p,n}^s \in \mathcal{B}_b(\mathsf{X}^{p+1}) \cap \mathrm{Lip}(\mathsf{X}^{p+1})$.

Proof. The case $p = n$ follows immediately from $\varphi \in \mathcal{B}_b(\mathsf{X}^{n+1}) \cap \mathrm{Lip}(\mathsf{X}^{n+1})$. We will use a backward inductive argument on p. Suppose $p = n-1$ then we have for any $(x_{0:n-1}, x'_{0:n-1}) \in \mathsf{X}^n \times \mathsf{X}^n$

$$|\varphi_{n-1,n}^s(x_{0:n-1}) - \varphi_{n-1,n}^s(x'_{0:n-1})| =$$
$$|\mathbb{E}^s[\varphi(x_{0:n-1}, X_n)G_n(X_n)|x_{n-1}] - \mathbb{E}^s[\varphi(x'_{0:n-1}, X_n)G_n(X_n)|x'_{n-1}]| \leq$$
$$|\mathbb{E}^s[\varphi(x_{0:n-1}, X_n)G_n(X_n)|x_{n-1}] - \mathbb{E}^s[\varphi(x'_{0:n-1}, X_n)G_n(X_n)|x_{n-1}]| +$$
$$|\mathbb{E}^s[\varphi(x'_{0:n-1}, X_n)G_n(X_n)|x_{n-1}] - \mathbb{E}^s[\varphi(x'_{0:n-1}, X_n)G_n(X_n)|x'_{n-1}]|$$

By $\varphi \in \mathrm{Lip}(\mathsf{X}^{n+1})$ it easily follows via (A1(i)) that

$$|\mathbb{E}^s[\varphi(x_{0:n-1}, X_n)G_n(X_n)|x_{n-1}] - \mathbb{E}^s[\varphi(x'_{0:n-1}, X_n)G_n(X_n)|x_{n-1}]| \leq M\sum_{j=0}^{n-1}|x_j - x'_j|.$$

By (A1(ii)) and $\varphi \in \mathrm{Lip}(\mathsf{X}^{n+1})$, $\varphi(x_{0:n})G_n(x_n)$ is Lipschitz in x_n and hence by (A2)

$$|\mathbb{E}^s[\varphi(x'_{0:n-1}, X_n)G_n(X_n)|x_{n-1}] - \mathbb{E}^s[\varphi(x'_{0:n-1}, X_n)G_n(X_n)|x'_{n-1}]| \leq M|x_{n-1} - x'_{n-1}|. \tag{23}$$

Hence it follows

$$|\varphi^s_{n-1,n}(x_{0:n-1}) - \varphi^s_{n-1,n}(x'_{0:n-1})| \leq M\sum_{j=0}^{n-1}|x_j - x'_j|.$$

The induction step follows by almost the same argument as above and is hence omitted. □

Proof of Proposition 29. We have the following standard collapsing sum representation:

$$\begin{aligned}\gamma^F_n(G_n\varphi) - \gamma^C_n(G_n\varphi) \;=\; \sum_{p=0}^{n}\Bigg(&\mathbb{E}^F[\prod_{q=0}^{p}G_q(X_q)\mathbb{E}^C[\varphi(X_{0:n})\prod_{q=p+1}^{n}G_q(X_q)|X_p]] - \\ &\mathbb{E}^F[\prod_{q=0}^{p-1}G_q(X_q)\mathbb{E}^C[\varphi(X_{0:n})\prod_{q=p}^{n}G_q(X_q)|X_{p-1}]]\Bigg)\end{aligned}$$

The summand is

$$T_p := \mathbb{E}^F\Big[\Big(\prod_{q=0}^{p-1}G_q(X_q)\Big)(\mathbb{E}^F - \mathbb{E}^C)\Big(\mathbb{E}^C[\varphi(X_{0:n})\prod_{q=p+1}^{n}G_q(X_q)|X_p]G_p(X_p)\Big|X_{p-1}\Big)\Big].$$

By Lemma 30, $\mathbb{E}^C[\varphi(x_{0:p}, X_{p+1:n})\prod_{q=p+1}^{n}G_q(X_q)|x_p] \in \mathcal{B}_b(\mathsf{X}^{p+1}) \cap \mathrm{Lip}(\mathsf{X}^{p+1})$ and by (A1) (i) and (ii) $G_p \in \mathcal{B}_b(\mathsf{X}) \cap \mathrm{Lip}(\mathsf{X})$. So by (22)

$$\Big|(\mathbb{E}^F - \mathbb{E}^C)\Big(\mathbb{E}^C[\varphi(X_{0:n})\prod_{q=p+1}^{n}G_q(X_q)|X_p]G_p(X_p)\Big|X_{p-1}\Big)\Big| \leq$$

$$Mh_\ell \sup_{x_{0:p}\in\mathsf{X}^{p+1}}|\mathbb{E}^C[\varphi(x_{0:p}, X_{p+1:n})\prod_{q=p+1}^{n}G_q(X_q)|$$

and hence

$$|T_p| \leq Mh_\ell\mathbb{E}^F[\prod_{q=0}^{p-1}G_q(X_q)]\sup_{x_{0:p}\in\mathsf{X}^{p+1}}|\mathbb{E}^C[\varphi(x_{0:p}, X_{p+1:n})\prod_{q=p+1}^{n}G_q(X_q)|x_p]G_p(x_p)|.$$

Application of (A1) (i) gives $|T_p| \leq Mh_\ell$ and the proof is hence concluded. □

Lemma 31. *Assume (A1). Then for any $n \geq 0$ there exists a $M < \infty$ such that for any $x_{0:n} \in \mathsf{X}^{2(n+1)}$*

$$\Big|\prod_{p=0}^{n}\frac{G_p(x^F_p)}{\check{G}_p(x_p)} - \prod_{p=0}^{n}\frac{G_p(x^C_p)}{\check{G}_p(x_p)}\Big| \leq M\sum_{p=0}^{n}|x^F_p - x^C_p|.$$

Proof. The is proof by induction. The case $n = 0$:

$$\Big|\frac{G_0(x^F_0)}{\check{G}_0(x_0)} - \frac{G_0(x^C_0)}{\check{G}_0(x_0)}\Big| = \frac{1}{\check{G}_0(x_0)}|G_0(x^F_0) - G_0(x^C_0)|.$$

Application of (A1) (ii) and (iii) yield that

$$\Big|\frac{G_0(x_0^F)}{\check{G}_0(x_0)} - \frac{G_0(x_0^C)}{\check{G}_0(x_0)}\Big| \leq M|x_0^F - x_0^C|.$$

The result is assumed to hold at rank $n-1$, then

$$\Big|\prod_{p=0}^{n}\frac{G_p(x_p^F)}{\check{G}_p(x_p)} - \prod_{p=0}^{n}\frac{G_p(x_p^C)}{\check{G}_p(x_p)}\Big| \leq$$

$$\Big|\frac{G_n(x_n^F)}{\check{G}_n(x_n)} - \frac{G_n(x_n^C)}{\check{G}_n(x_n)}\Big| \cdot \prod_{p=0}^{n-1}\frac{G_p(x_p^F)}{\check{G}_p(x_p)} + \Big|\prod_{p=0}^{n-1}\frac{G_p(x_p^F)}{\check{G}_p(x_p)} - \prod_{p=0}^{n-1}\frac{G_p(x_p^C)}{\check{G}_p(x_p)}\Big| \cdot \frac{G_n(x_n^C)}{\check{G}_n(x_n)}.$$

For the first term of the R.H.S. one can follow the argument at the initialisation and apply (A1) (i) and (iii). For the second term of the R.H.S., the induction hypothesis and (A1) (i) and (iii) can be used. That is one can deduce that

$$\Big|\prod_{p=0}^{n}\frac{G_p(x_p^F)}{\check{G}_p(x_p)} - \prod_{p=0}^{n}\frac{G_p(x_p^C)}{\check{G}_p(x_p)}\Big| \leq M\sum_{p=0}^{n}|x_p^F - x_p^C|.$$

□

Recall (18) for the definition of ψ and that $x_p = (x_p^F, x_p^C) \in \mathsf{X} \times \mathsf{X}$.

Lemma 32. *Assume (A1-2). Then for any $0 \leq p < n$, $\varphi \in \mathcal{B}_b(\mathsf{X}^{n+1}) \cap \mathrm{Lip}(\mathsf{X}^{n+1})$ there exists a $M < \infty$ such that for any $x_{0:p} \in \mathsf{E}_p \times \mathsf{E}_p$*

$$|\check{\boldsymbol{Q}}_{p,n}(\psi)(x_{0:p})| \leq M\Big(\sum_{j=0}^{p}|x_j^F - x_j^C| + h_\ell\Big)$$

Proof. We have

$$\begin{aligned}\check{\boldsymbol{Q}}_{p,n}(\psi)(x_{0:p}) = \check{G}_p(x_p) \times \Big(&\prod_{q=0}^{p}\frac{G_q(x_q^F)}{\check{G}_q(x_q)}\mathbb{E}^F[\varphi(x_{0:p}^F, Y_{p+1:n})\prod_{s=p+1}^{n}G_s(X_s^F)|x_p^F] \\ &- \prod_{q=0}^{p}\frac{G_q(x_q^C)}{\check{G}_q(x_q)}\mathbb{E}^C[\varphi(x_{0:p}^C, Y_{p+1:n})\prod_{s=p+1}^{n}G_s(X_s^C)|x_p^C]\Big).\end{aligned}$$

It then follows that $\check{\boldsymbol{Q}}_{p,n}(\psi)(x_{0:p}) = \check{G}_p(x_p)(T_1 + T_2)$ where

$$\begin{aligned}T_1 &= \Big(\prod_{q=0}^{p}\frac{G_q(x_q^F)}{\check{G}_q(x_q)} - \prod_{q=0}^{p}\frac{G_q(x_q^C)}{\check{G}_q(x_q)}\Big)\mathbb{E}^F[\varphi(x_{0:p}^F, Y_{p+1:n})\prod_{s=p+1}^{n}G_s(X_s^F)|x_p^F] \\ T_2 &= \prod_{q=0}^{p}\frac{G_q(x_q^C)}{\check{G}_q(x_q)}\Big(\mathbb{E}^F[\varphi(x_{0:p}^F, Y_{p+1:n})\prod_{s=p+1}^{n}G_s(X_s^F)|x_p^F] - \mathbb{E}^C[\varphi(x_{0:p}^C, Y_{p+1:n})\prod_{s=p+1}^{n}G_s(X_s^C)|x_p^C]\Big).\end{aligned}$$

By Lemma 31, $\varphi \in \mathcal{B}_b(\mathsf{X}^{n+1}) \cap \mathrm{Lip}(\mathsf{X}^{n+1})$ and (A1) (i)

$$|T_1| \leq M\sum_{j=0}^{p}|x_j^F - x_j^C|.$$

Now $T_2 = T_3 + T_4$ where

$$T_3 = \prod_{q=0}^{p} \frac{G_q(x_q^C)}{\check{G}_q(x_q)} \Big(\mathbb{E}^F[\varphi(x_{0:p}^F, Y_{p+1:n}) \prod_{q=p+1}^{n} G_s(X_s^F) | x_p^F] - \mathbb{E}^F[\varphi(x_{0:p}^F, Y_{p+1:n}) \prod_{s=p+1}^{n} G_s(X_s^F) | x_p^C] \Big)$$

$$T_4 = \prod_{q=0}^{p} \frac{G_q(x_q^C)}{\check{G}_q(x_q)} \Big(\mathbb{E}^F[\varphi(x_{0:p}^F, Y_{p+1:n}) \prod_{s=p+1}^{n} G_s(X_s^F) | x_p^C] - \mathbb{E}^C[\varphi(x_{0:p}^C, Y_{p+1:n}) \prod_{s=p+1}^{n} G_s(X_s^C) | x_p^C] \Big).$$

For T_3 one can use Lemma 30 (along with (A1) (i) and (iii)) to get that

$$|T_3| \leq M \sum_{j=0}^{p} |x_j^F - x_j^C|.$$

For T_4 a similar collapsing sum argument that is used in the proof of Proposition 29 can be used to deduce that

$$|T_4| \leq M h_\ell.$$

One can then conclude the proof via the above bounds (along with (A1) (i)). □

Below $\mathbb{E}$ denotes expectation w.r.t. the particle system described in Section A.2 started at position (x, x) at time $n = 0$ with $x \in \mathsf{X}$, in the diffusion case of Section A.4. Recall the particle $U_n^i \in \mathsf{E}_n \times \mathsf{E}_n$ at time $n \geq 0$ in path space. We denote by $U_n^{i,s}(j) \in \mathsf{X}$ as the $j \in \{0, \ldots, n\}$ component of particle $i \in \{1, \ldots, N\}$ at time $n \geq 0$ of $s \in \{F, C\}$ component. Recall $(U_n^{i,F}(n), U_n^{i,C}(n))$ for $n \geq 1$ is sampled from the kernel $\check{M}_n((\bar{u}_{n-1}^{i,F}(n-1), \bar{u}_{n-1}^{i,C}(n-1)), \cdot)$ where the $\bar{u}$ denotes post-resampling and the component $(U_n^{i,F}(j), U_n^{i,C}(j)) = (\bar{u}_{n-1}^{i,F}(j), \bar{u}_{n-1}^{i,C}(j))$ for $j \in \{0, \ldots, n-1\}$ is kept the same for the earlier components of the particle.

Lemma 33. *Assume (A1 (i) (iii), 2). Then for any $n \geq 0$ there exists a $M < \infty$ such that*

$$\mathbb{E}[\sum_{j=0}^{n} |U_n^{1,F}(j) - U_n^{1,C}(j)|^2] \leq M h_\ell^\beta.$$

where β is as in (8).

Proof. Our proof is by induction, the case $n = 0$ following by (8). Assuming the result at $n - 1$ we have

$$\mathbb{E}[\sum_{j=0}^{n} |U_n^{1,F}(j) - U_n^{1,C}(j)|^2] = \mathbb{E}[\sum_{j=0}^{n-1} |\bar{U}_{n-1}^{1,F}(j) - \bar{U}_{n-1}^{1,C}(j)|^2 + |U_n^{1,F}(n) - U_n^{1,C}(n)|^2].$$

Now

$$\begin{aligned} \mathbb{E}[\sum_{j=0}^{n-1} |\bar{U}_{n-1}^{1,F}(j) - \bar{U}_{n-1}^{1,C}(j)|^2] &= N \sum_{j=0}^{n-1} \mathbb{E}\Big[\frac{\check{G}_{n-1}(U_{n-1}^{1,F}(n-1), U_{n-1}^{1,C}(n-1))}{\sum_{j=1}^{N} \check{G}_{n-1}(U_{n-1}^{j,F}(n-1), U_{n-1}^{j,C}(n-1))} \times \\ &\qquad |U_{n-1}^{1,F}(j) - U_{n-1}^{1,C}(j)|^2 \Big] \\ &\leq M \mathbb{E}[\sum_{j=0}^{n-1} |U_{n-1}^{1,F}(j) - U_{n-1}^{1,C}(j)|^2] \end{aligned}$$

where we have used (A1) (i) and (iii). Applying the induction hypothesis along with (8) yields

$$\mathbb{E}[\sum_{j=0}^{n} |U_n^{1,F}(j) - U_n^{1,C}(j)|^2] \leq M \Big(h_\ell^\beta + \mathbb{E}[|\bar{U}_{n-1}^{1,F}(n-1) - \bar{U}_{n-1}^{1,C}(n-1)|^2] \Big)$$

Now

$$\mathbb{E}[|\bar{U}^{1,F}_{n-1}(n-1)-\bar{U}^{1,C}_{n-1}(n-1)|^2]=$$

$$N\mathbb{E}\Big[\frac{\check{G}_{n-1}(U^{1,F}_{n-1}(n-1),U^{1,C}_{n-1}(n-1))}{\sum_{j=1}^N\check{G}_{n-1}(U^{j,F}_{n-1}(n-1),U^{j,C}_{n-1}(n-1))}|U^{1,F}_{n-1}(n-1)-U^{1,C}_{n-1}(n-1)|^2\Big]$$

Then by (A1) (i) and (iii)

$$\mathbb{E}\Big[\frac{\check{G}_{n-1}(U^{1,F}_{n-1}(n-1),U^{1,C}_{n-1}(n-1))}{\sum_{j=1}^N\check{G}_{n-1}(U^{j,F}_{n-1}(n-1),U^{j,C}_{n-1}(n-1))}|U^{1,F}_{n-1}(n-1)-U^{1,C}_{n-1}(n-1)|^2\Big]\leq$$

$$\frac{M}{N}\mathbb{E}[|U^{1,F}_{n-1}(n-1)-U^{1,C}_{n-1}(n-1)|^2]\leq\frac{M}{N}\mathbb{E}[\sum_{j=0}^{n-1}|U^{1,F}_{n-1}(j)-U^{1,C}_{n-1}(j)|^2].$$

Hence via the induction hypothesis, one has

$$\mathbb{E}[|\bar{U}^{1,F}_{n-1}(n-1)-\bar{U}^{1,C}_{n-1}(n-1)|^2]\leq Mh_\ell^\beta$$

and the proof is concluded. □

Recall Remark 27.

Proof of Theorem 8. This follows first by applying Proposition 28, followed by Lemma 32 and then some standard calculations followed by Lemma 33. □

Proof of Corollary 9. Easily follows by adding and subtracting $\check{\boldsymbol{\gamma}}_n(\psi)$ the C_2 inequality along with Theorem 8, and then using (19) combined with Proposition 29. □

Appendix B. Proof of consistency of the Markov chain Monte Carlo

Proof of Theorem 11. Denote

$$\xi_k(g):=\big(\sum_{i=1}^N V_k^{(i)}+\epsilon\big)^{-1}\big[\sum_{i=1}^N V_k^{(i)}g(\Theta_k,X_k^{(i)})+\tilde{\Delta}_k(g^{(\Theta_k)})\big],\tag{24}$$

where $g^{(\theta)}(x):=g(\theta,x)$ and $\tilde{\Delta}_k(g^{(\theta)}):=p_{L_k}^{-1}\sum_{i=1}^{2N}V_{k,L_k}^{(i)}g^{(\theta)}(\mathbf{X}_{k,L_k}^{(i)})$. Then $E_{m_{\text{iter}},N,\mathbf{p}}(f)=\frac{\sum_{k=1}^{m_{\text{iter}}}\xi_k(f)}{\sum_{j=1}^{m_{\text{iter}}}\xi_k(\mathbf{1})}$. Furthermore, by Assumption 5 [cf. 26, 31], we have

$$\begin{aligned}\mathbb{E}[\tilde{\Delta}_k^2(g)\mid\Theta_k=\theta]&=s_g(\theta),\\ \mathbb{E}[\tilde{\Delta}_k(g)\mid\Theta_k=\theta]&=\gamma_n^{(\theta,\infty)}(G_ng)-\gamma_n^{(\theta,0)}(G_ng)\end{aligned}$$

for $g=1$ and $g=f^{(\theta)}$. This implies for $g=f$ and $g=1$,

$$\begin{aligned}\mu_g(\theta,v^{(1:N)},\mathbf{x}^{(1:N)})&:=\mathbb{E}[\xi_k(g)\mid(\Theta_k,V_k^{(1:N)},\mathbf{X}_k^{(1:N)})=(\theta,v^{(1:N)},\mathbf{x}^{(1:N)})]\\&=\frac{1}{\sum_{j=1}^N v^{(j)}+\epsilon}\Big[\sum_{i=1}^N v^{(i)}g(\theta,x^{(i)})-\gamma_n^{(\theta,0)}(G_ng)+\gamma_n^{(\theta,\infty)}(G_ng)\Big],\\ m_g^{(1)}(\theta,v^{(1:N)},\mathbf{x}^{(1:N)})&:=\mathbb{E}[|\xi_k(g)|\mid(\Theta_k,V_k^{(1:N)},\mathbf{X}_k^{(1:N)})=(\theta,v^{(1:N)},\mathbf{x}^{(1:N)})]\\&\leq\frac{1}{\sum_{j=1}^N v^{(j)}+\epsilon}\Big[\sum_{i=1}^N v^{(i)}|g(\theta,x^{(i)})|+\sqrt{s_{g^{(\theta)}}(\theta)}\Big].\end{aligned}$$

It is direct to check that the PMMH type chain $(\Theta_k, X_k^{(1:N)}, V_k^{(1:N)})$ is reversible with respect to the probability

$$\Pi(\mathrm{d}\theta, \mathrm{d}x^{(1:N)}, \mathrm{d}v^{(1:N)}) = c_0 \mathrm{pr}(\theta)\mathrm{d}\theta R_\theta^{(0)}(\mathrm{d}x^{(1:N)}, \mathrm{d}v^{(1:N)})\Big(\sum_{i=1}^N v^{(i)} + \epsilon\Big), \tag{25}$$

where $c_0 > 0$ is a normalisation constant and $R_\theta^{(0)}(\,\cdot\,)$ stands for the law of the output of Algorithm 1 with $(M_{0:n}^{(\theta,0)}, G_{0:n}^{(\theta,0)}, N)$, and therefore is Harris recurrent as a full-dimensional Metropolis–Hastings that is ψ-irreducible [cf. 27, Theorem 8]. It is direct to check that $\Pi(m_f^{(1)}) < \infty$, $\Pi(m_1^{(1)}) < \infty$, $\Pi(\mu_f) = c\pi^{(\infty)}(f)$ and $\Pi(\mu_1) = c$, where $c > 0$ is a constant, so the result follows from [32, Theorem 3]. □

Appendix C. Proofs about asymptotic efficiency and allocations

Proof of Proposition 16. By Harris ergodicity, $m^{-1}\mathscr{C}(m) \to \mathbb{E}[\boldsymbol{\tau}]$ almost surely. Dividing the inequality

$$\mathscr{C}(\mathscr{L}(\kappa)) \le \kappa < \mathscr{C}(\mathscr{L}(\kappa) + 1)$$

by $\mathscr{L}(\kappa)$ and taking the limit $\kappa \to \infty$, which implies $\mathscr{L}(\kappa) \to \infty$, we get that $\kappa/\mathscr{L}(\kappa) \to \mathbb{E}[\boldsymbol{\tau}]$ almost surely. Also, by Proposition 13,

$$\sqrt{\mathscr{L}(\kappa)}\big[E_{\mathscr{L}(\kappa),N,\mathbf{p}}(f) - \pi^{(\infty)}(f)\big] \xrightarrow{\kappa\to\infty} \mathcal{N}(0, \sigma^2), \qquad \text{in distribution,}$$

so the result follows by Slutsky's theorem. □

Proof of Proposition 20. We have that

$$\mathbb{E}[\mathscr{C}(m)] = \sum_{k=1}^m \mathbb{E}[\tau_{\Theta_k, L_k}] = \sum_{k=1}^m \sum_{\ell=1}^\infty \mathbb{E}[\tau_{\Theta_k,\ell}] p_\ell. \le Cm \sum_{k=1}^\infty p_\ell 2^{\gamma\ell(1+\rho)},$$

by Assumption 18(i), which is finite if $r > \gamma(1+\rho)$. Also,

$$s_g(\theta) = \mathbb{E}[\tilde{\Delta}_k^2(g)|\Theta_k = \theta] = \sum_{\ell\ge1} \frac{\mathbb{E}\Delta_\ell^2}{p_\ell} \le C \sum \Big(2^{-\ell(\beta+\rho-r)} + 2^{-\ell(2\alpha-r)}\Big),$$

which is finite if $r < \min(\beta+\rho, 2\alpha)$. Therefore, $\sigma^2 < \infty$, and the CLT follows by Proposition 16. □

Lemma 34. *Let $\{X_k\}_{k\ge1}$ be a sequence of independent random variables with $\mathbb{E}[X_{k_0}] = \infty$ for at least one k_0, and let $\{a_k\}_{k\ge1}$ be a sequence of monotonically increasing real numbers with $a_k/k \longrightarrow \infty$. Suppose one of the following assumptions holds:*

(i) $\sum_{k\ge1} \mathbb{P}[X_k > a_k] < \infty$, and $\{X_k\}_{k\ge1}$ are also identically distributed, or
(ii) $\sum_{k\ge1} \sup_{m\ge1} \mathbb{P}[X_m > a_k] < \infty$.

Then

$$\mathbb{P}[\sum_{k=1}^m X_k > a_m \text{ infinitely many } m \in \mathbb{N}] = 0.$$

Proof. (i) is [11, Theorem 2] since $\mathbb{E}[X_{k_0}] = \infty$ implies $\mathbb{E}[X_k] = \infty$ for all $k \ge 1$ as $\{X_k\}_{k\ge1}$ are i.i.d. For (ii), note that if X_k has c.d.f. denoted F_k, then it is straightforward to check that

$$F^*(x) := \inf_{k\ge1} F_k(x)$$

is a c.d.f. also. With $X_k^* \sim F^*$ i.i.d. for $k \ge 1$, we have

$$\mathbb{P}[X_k^* > a_k] = 1 - F^*(a_k) = \sup_{m\ge1} 1 - F_m(a_k) = \sup_{m\ge1} P[X_m > a_k].$$

Summing over $k \geq 1$, we obtain $\sum_{k\geq 1} \mathbb{P}[X_k^* > a_k] < \infty$. In addition,

$$\mathbb{E}[X_k^*] = \int \mathbb{P}[X_k^* > x]\mathrm{d}x \geq \int \mathbb{P}[X_{k_0} > x]\mathrm{d}x = \infty,$$

for all $k \geq 1$. Hence, we can apply (i) for i.i.d. random variables, obtaining

$$0 = \mathbb{P}[\sum_{k=1}^{m} X_k^* > a_m \text{ infinitely many } m] \geq \mathbb{P}[\sum_{k=1}^{m} X_k > a_m \text{ infinitely many } m],$$

where the first equality comes from (i), and so we conclude. □

Proof of Proposition 22. Conditional on output $\{\Theta_k\}_{k\geq 1}$ of Algorithm 4, $\{\tau_{\Theta_k,L_k}\}_{k\geq 1}$ are independent random variables. Our assumptions imply Lemma 34(ii) holds, so

$$\mathbb{P}[\mathscr{C}(m) > a_m \text{ infinitely many } m] = 0,$$

which means that $\mathscr{C}(m)$ is asymptotically bounded by a_m. Setting $m = O(\epsilon^{-2})$ allows us to conclude. □

The proofs below of Proposition 24 and 26 are similar to that of [26, Proposition 4 and 5].

Proof of Proposition 24. With the prescribed choice of p_ℓ we have finite variance, as

$$s_g(\theta) = \sum_{\ell\geq 1} \frac{\mathbb{E}\Delta_\ell^2}{p_\ell} \leq C \sum_{\ell\geq 1} \frac{1}{\ell[\log_2(\ell+1)]^\eta} < \infty,$$

uniformly in $\theta \in \mathsf{T}$. To determine the order of complexity, we would like to apply Lemma 34(i) to the i.i.d sequence $\{\tau_{L_k}^*\}_{k\geq 1}$, where $\tau_\ell^* := C2^{\gamma\ell(1+\rho)}$. For any $k \geq 1$, where $a_k > 0$ is some positive real number, we have,

$$\mathbb{P}[\tau_{L_k}^* > a_k] = \sum_{\ell\geq 1} \mathbb{P}[\tau_\ell^* > a_k] p_\ell = \sum_{\ell\geq 1} \mathbf{1}\left\{\ell > \frac{1}{\gamma(1+\rho)} \log_2 \frac{a_k}{C}\right\} p_\ell. \tag{26}$$

Because $\sum_{\ell\geq 1} p_\ell = 1$ and p_ℓ is monotonically decreasing, we have $\sum_{\ell\geq \ell_*} p_\ell$ is $O(p_{\ell_*})$. Setting $\ell_* = \lfloor \frac{1}{\gamma(1+\rho)} \log_2 \frac{a_k}{C} \rfloor$, we therefore obtain that (26) is of order

$$a_k^{-\frac{2b}{\gamma(1+\rho)}} \big(\log_2 a_k\big)\big(\log_2 \log_2 a_k\big)^\eta.$$

Setting

$$a_k := [k(\log_2 k)^q]^{\frac{\gamma(1+\rho)}{2b}} \tag{27}$$

then ensures that $\sum_{k\geq 1} \mathbb{P}[\tau_{L_k}^* > a_k] < \infty$. As $\beta \leq 1$, it is easy to check that $\mathbb{E}[\tau_{L_k}^*] = \infty$. We then apply Lemma 34(i), obtaining

$$0 = \mathbb{P}[\sum_{k=1}^{m} \tau_{L_k}^* > a_m \text{ infinitely many } m] \geq \mathbb{P}[\sum_{k=1}^{m} \tau_{\Theta_k,L_k} > a_m \text{ infinitely many } m].$$

and conclude as before, by using that $\mathscr{C}(m)$ is asymptotically bounded by a_m and setting $m = O(\epsilon^{-2})$. □

Proof of Proposition 26. We are in the basic setting of Proposition 24 as before, but additionally may choose $\rho \geq 0$ as we please. The growth of a_k given in (27) is essentially determined by $\gamma(1+\rho)/2b$, which can be made small when $\rho = 2\alpha - \beta$, implying $b = \alpha$. □

References

[1] C. Andrieu and J. Thoms. A tutorial on adaptive MCMC. *Statist. Comput.*, 18(4): 343–373, Dec. 2008.

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

[3] A. Beskos and G. Roberts. Exact simulation of diffusions. *Ann. Appl. Probab.*, 15(4): 2422–2444, 11 2005. doi: 10.1214/105051605000000485.

[4] A. Beskos, O. Papaspiliopoulos, G. O. Roberts, and P. Fearnhead. Exact and computationally efficient likelihood-based estimation for discretely observed diffusion processes. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 68(3):333–382, 2006. (with discussion).

[5] O. Cappé, E. Moulines, and T. Ryden. *Inference in Hidden Markov Models*. Springer, New York, 2005.

[6] A. Cliffe, M. Giles, R. Scheichl, and A. Teckentrup. Multilevel Monte Carlo methods and applications to elliptic PDEs with random coefficients. *Comput. Vis. Sci.*, 14(1): 3, 2011.

[7] P. Del Moral. *Feynman-Kac Formulae*. Springer, New York, 2004.

[8] R. Douc, O. Cappé, and E. Moulines. Comparison of resampling schemes for particle filtering. In *Proc. Image and Signal Processing and Analysis, 2005*, pages 64–69, 2005.

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

[10] P. Fearnhead, K. Latuszynski, G. Roberts, and G. Sermaidis. Continious-time importance sampling: Monte Carlo methods which avoid time-discretisation error. Preprint arXiv:1712.06201, 2017.

[11] W. Feller. A limit theorem for random variables with infinite moments. *Amer. J. Math.*, 68(2):257–262, 1946.

[12] J. Franks and M. Vihola. Importance sampling correction versus standard averages of reversible MCMCs in terms of the asymptotic variance. Preprint arXiv:1706.09873v3, 2017.

[13] M. Giles and L. Szpruch. Antithetic multilevel Monte Carlo estimation for multi-dimensional SDEs without Lévy area simulation. *Ann. Appl. Probab.*, 24(4):1585–1620, 2014.

[14] M. B. Giles. Multilevel Monte Carlo path simulation. *Oper. Res.*, 56(3):607–617, 2008.

[15] P. Glynn and W. Whitt. The asymptotic efficiency of simulation estimators. *Oper. Res.*, 40(3):505–520, 1992.

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

[17] S. Heinrich. Multilevel Monte Carlo methods. In *Large-scale scientific computing*, pages 58–67. Springer, 2001.

[18] A. Jasra, K. Kamatani, K. J. H. Law, and Y. Zhou. Multilevel particle filters. *SIAM J. Numer. Anal.*, 55:3068–3096, 2017.

[19] A. Jasra, K. Kamatani, K. J. H. Law, and Y. Zhou. Bayesian static parameter estimation for partially observed diffusions via multilevel Monte Carlo. *SIAM J. Sci. Comp.*, 40:A887–A902, 2018.

[20] A. Jasra, K. Kamatani, K. J. H. Law, and Y. Zhou. A multi-index Markov chain Monte Carlo method. *Intern. J. Uncertainty Quantif.*, 8(1), 2018.

[21] A. Jasra, K. J. Law, and P. P. Osei. Multilevel particle filters for Lévy-driven stochastic differential equations. Preprint arXiv:1804.04444, 2018.
[22] P. Kloeden and E. Platen. *Numerical solution of stochastic differential equations*. Springer, Berlin Heidelberg, 3rd edition, 1999.
[23] D. McLeish. A general method for debiasing a Monte Carlo estimator. *Monte Carlo Methods Appl.*, 17(4):301–315, 2011.
[24] J. K. Møller and H. Madsen. From state dependent diffusion to constant diffusion in stochastic differential equations by the Lamperti transform. *IMM-Technical Report-2010-16*, 2010.
[25] O. Papaspiliopoulos and G. Roberts. Importance sampling techniques for estimation of diffusion models. *Statistical methods for stochastic differential equations*, 124:311–340, 2012.
[26] C.-H. Rhee and P. W. Glynn. Unbiased estimation with square root convergence for SDE models. *Oper. Res.*, 63(5):1026–1043, 2015.
[27] G. Roberts and J. Rosenthal. Harris recurrence of Metropolis-within-Gibbs and trans-dimensional Markov chains. *Ann. Appl. Probab.*, 16(4):2123–2139, 2006.
[28] G. Sermaidis, O. Papaspiliopoulos, G. Roberts, A. Beskos, and P. Fearnhead. Markov chain Monte Carlo for exact inference for diffusions. *Scand. J. Statist.*, 40(2):294–321, 2013.
[29] C. Sherlock, A. H. Thiery, G. O. Roberts, and J. S. Rosenthal. On the efficiency of pseudo-marginal random walk Metropolis algorithms. *Ann. Statist.*, 43(1):238–275, 2015.
[30] H. Sørensen. Parametric inference for diffusion processes observed at discrete points in time: a survey. *Intern. Statist. Review*, 72(3):337–354, 2004.
[31] M. Vihola. Unbiased estimators and multilevel Monte Carlo. *Oper. Res.*, 66(2):448–462, 2018.
[32] M. Vihola, J. Helske, and J. Franks. Importance sampling type estimators based on approximate marginal MCMC. Preprint arXiv:1609.02541v4, 2016.
[33] Q. Wang, V. Rao, and Y. W. Teh. An exact auxiliary variable Gibbs sampler for a class of diffusions. Preprint arXiv:1903.10659, 2019.

J. Franks & M. Vihola, Department of Mathematics and Statistics P.O.Box 35, FI-40014 University of Jyväskylä, FI. Email: jordan.j.franks@jyu.fi & matti.s.vihola@jyu.fi.

A. Jasra, Department of Statistics & Applied Probability, National University of Singapore, Singapore, 117546, SG. Email: staja@nus.edu.sg.

K. J. H. Law, School of Mathematics, University of Manchester, Manchester, M13 9PL, UK. Email: kodylaw@gmail.com

Article [D]

On the use of ABC-MCMC with inflated tolerance and post-correction

Matti Vihola and Jordan Franks

Preprint arXiv:1902.00412, 2019.

ON THE USE OF ABC-MCMC WITH INFLATED TOLERANCE AND POST-CORRECTION

MATTI VIHOLA AND JORDAN FRANKS

Abstract. Approximate Bayesian computation (ABC) allows for inference of complicated probabilistic models with intractable likelihoods using model simulations. The ABC Markov chain Monte Carlo (MCMC) inference is often sensitive to the tolerance parameter: low tolerance leads to poor mixing and large tolerance entails excess bias. We consider an approach using a relatively large tolerance for the ABC-MCMC to ensure sufficient mixing, and post-processing the output of ABC-MCMC leading to estimators for a range of finer tolerances. We introduce an approximate confidence interval for the related post-corrected estimators, which can be calculated from any run of ABC-MCMC, with little extra cost. We propose an adaptive ABC-MCMC, which finds a 'balanced' tolerance level automatically, based on acceptance rate optimisation. Tolerance adaptation, combined with proposal covariance adaptation, leads to an easy-to-use adaptive ABC-MCMC, with subsequent post-correction over a range of tolerances. Our experiments show that post-processing based estimators can perform better than direct ABC-MCMC, that our confidence intervals are reliable, and that our adaptive ABC-MCMC leads to reliable inference with little user specification.

1. Introduction

Approximate Bayesian computation (ABC) is a form of likelihood-free inference (see, e.g., the reviews Marin et al., 2012; Sunnåker et al., 2013) which is used when exact Bayesian inference of a parameter $\theta \in \mathsf{T}$ with posterior density $\pi(\theta) \propto \mathrm{pr}(\theta)L(\theta)$ is impossible, where $\mathrm{pr}(\theta)$ is the prior density and $L(\theta) := g(y^* \mid \theta)$ is an intractable likelihood with data $y^* \in \mathsf{Y}$. More specifically, when the generative model of observations $g(\cdot \mid \theta)$ cannot be evaluated, but allows for simulations, ABC can be used for relatively straightforward approximate inference, based on a pseudo-posterior

$$\pi_\epsilon(\theta) \propto \mathrm{pr}(\theta)L_\epsilon(\theta), \qquad \text{where} \qquad L_\epsilon(\theta) := \mathbb{E}[K_\epsilon(Y_\theta, y^*)], \quad Y_\theta \sim g(\cdot \mid \theta), \tag{1}$$

where $\epsilon > 0$ is a 'tolerance' parameter, and $K_\epsilon : \mathsf{Y}^2 \to [0, \infty)$ is a 'kernel' function, which is often taken as a simple cut-off $K_\epsilon(y, y^*) = \mathbf{1}\left(\|s(y) - s(y^*)\| \le \epsilon\right)$, where $s : \mathsf{Y} \to \mathbb{R}^d$ extracts a vector of summary statistics from the (pseudo) observations

The summary statistics are often chosen based on the application at hand, and reflect what is relevant for the inference task; see also (Fearnhead and Prangle, 2012). Because $L_\epsilon(\theta)$ may be regarded as a smoothed version of the true likelihood $g(y^* \mid \theta)$ using the kernel K_ϵ, it is intuitive that using a too large ϵ may blur the likelihood and bias the inference. Therefore, it is generally desirable to use as small a tolerance $\epsilon > 0$ as possible, but because the computational ABC methods suffer from inefficiency with small ϵ, the choice of tolerance level is difficult (cf. Bortot et al., 2007; Sisson and Fan, 2018; Tanaka et al., 2006).

We discuss a simple post-processing procedure which allows for consideration of a range of values for the tolerance $\epsilon \le \epsilon_0$, based on a single run of ABC Markov chain Monte Carlo (ABC-MCMC) (Marjoram et al., 2003) with tolerance ϵ_0. Post-processing has been

Key words and phrases. Adaptive, approximate Bayesian computation, confidence interval, importance sampling, Markov chain Monte Carlo, tolerance choice.

suggested earlier at least in (Wegmann et al., 2009) (in the special case of simple cut-off), and it can be regarded as an importance sampling correction of pseudo-marginal type MCMC (cf. Vihola et al., 2016). The method, discussed further in Section 2, can be useful for two reasons:

- A range of tolerances $\epsilon \leq \epsilon_0$ may be routinely inspected, which can reveal excess bias of ABC-MCMC with tolerance ϵ_0.
- The ABC-MCMC may be implemented with sufficiently large ϵ_0 to allow for good mixing, and post-correction $\epsilon_0 \to \epsilon$ may be used for inference.

Our contribution is two-fold. We suggest straightforward-to-calculate approximate confidence intervals for the post-processing output, with some theoretical properties discussed in Section 3. We also introduce an adaptive ABC-MCMC in Section 4 which finds a balanced ϵ_0 during burn-in, using acceptance rate as a proxy. We provide some experimental results regarding the suggested confidence interval and the tolerance adaptation in Section 5, and conclude with discussion in Section 6.

2. ABC-MCMC with post-processing over a range of tolerances

For the rest of the paper, we assume that the kernel function has the following form:

$$K_\epsilon(y, y^*) := \phi\Big(\frac{d(y, y^*)}{\epsilon}\Big),$$

where $d : \mathsf{Y}^2 \to [0, \infty)$ is any 'dissimilarity' function and $\phi : [0, \infty) \to [0, 1]$ is a non-increasing 'cut-off' function. Typically $d(y, y^*) = \|s(y) - s(y^*)\|$, where $s : \mathsf{Y}^2 \to \mathbb{R}^d$ are the chosen summaries, and in case of the simple cut-off discussed in Section 1, $\phi(t) = \phi_{\text{simple}}(t) := \mathbf{1}\,(t \leq 1)$. We also assume that the ABC posterior π_ϵ given in (1) is well-defined for all $\epsilon > 0$ of interest (that is, $\int \mathrm{pr}(\theta) L_\epsilon(\theta) \mathrm{d}\theta > 0$).

The following summarises the ABC-MCMC algorithm suggested by Marjoram et al. (2003), using a proposal density q and a tolerance $\epsilon > 0$:

Algorithm 1 (ABC-MCMC(ϵ))**.** Suppose $\Theta_0 \in \mathsf{T}$ and $Y_0 \in \mathsf{Y}$ are any starting values, such that $\mathrm{pr}(\Theta_0) > 0$ and $\phi(d(Y_0, y^*)/\epsilon) > 0$. For $k \geq 1$, iterate:

(i) Draw $\tilde{\Theta}_k \sim q(\Theta_{k-1}, \cdot\,)$ and $\tilde{Y}_k \sim g(\cdot \mid \tilde{\Theta}_k)$.
(ii) With probability $\alpha_\epsilon(\Theta_{k-1}, Y_{k-1}; \tilde{\Theta}_k, \tilde{Y}_k)$ accept and set $(\Theta_k, Y_k) \leftarrow (\tilde{\Theta}_k, \tilde{Y}_k)$; otherwise reject and set $(\Theta_k, Y_k) \leftarrow (\Theta_{k-1}, Y_{k-1})$, where

$$\alpha_\epsilon(\theta, y; \tilde{\theta}, \tilde{y}) := \min\left\{1, \frac{\mathrm{pr}(\tilde{\theta}) q(\tilde{\theta}, \theta) \phi\big(d(\tilde{y}, y^*)/\epsilon\big)}{\mathrm{pr}(\theta) q(\theta, \tilde{\theta}) \phi\big(d(y, y^*)/\epsilon\big)}\right\}.$$

Note that Algorithm 1 may be implemented by storing only Θ_k and the related distances $T_k := d(Y_k, y^*)$, and in what follows, we regard either $(\Theta_k, Y_k)_{k\geq 1}$ or $(\Theta_k, T_k)_{k\geq 1}$ as the output of Algorithm 1. Note also that in practice, the initial values (Θ_0, Y_0) should be taken as the state of the Algorithm 1 run for a number of initial 'burn-in' iterations, during which time an adaptive algorithm for parameter tuning may be employed (Section 4).

Definition 2. Suppose $(\Theta_k, T_k)_{k=1,\dots,n}$ is the output of ABC-MCMC(ϵ_0) for some $\epsilon_0 > 0$. For any $\epsilon \in (0, \epsilon_0]$ such that $\phi(T_k/\epsilon) > 0$ for some $k = 1, \dots, n$, and for any function

$f : \mathsf{T} \to \mathbb{R}$, define

$$W_k^{(\epsilon_0,\epsilon)} := \frac{U_k^{(\epsilon_0,\epsilon)}}{\sum_{j=1}^n U_j^{(\epsilon_0,\epsilon)}}, \qquad U_k^{(\epsilon_0,\epsilon)} := \frac{\phi(T_k/\epsilon)}{\phi(T_k/\epsilon_0)},$$

$$E_{\epsilon_0,\epsilon}(f) := \sum_{k=1}^n W_k^{(\epsilon_0,\epsilon)} f(\Theta_k), \qquad S_{\epsilon_0,\epsilon}(f) := \sum_{k=1}^n (W_k^{(\epsilon_0,\epsilon)})^2 \big[f(\Theta_k) - E_{\epsilon_0,\epsilon}(f)\big]^2.$$

The estimator $E_{\epsilon_0,\epsilon}(f)$ approximates $\mathbb{E}_{\pi_\epsilon}[f(\Theta)]$ and $S_{\epsilon_0,\epsilon}(f)$ may be used to construct a confidence interval; see Algorithm 6 below. The following algorithm shows that in case of simple cut-off, $E_{\epsilon_0,\epsilon}(f)$ and $S_{\epsilon_0,\epsilon}(f)$ may be calculated simultaneously for all tolerances efficiently:

Algorithm 3. Suppose $\phi = \phi_{\text{simple}}$ and $(\Theta_k, T_k)_{k=1,\dots,n}$ is the output of ABC-MCMC(ϵ_0).

(i) Sort $(\Theta_k, T_k)_{k=1,\dots,n}$ with respect to T_k:
- Find indices $I_1, \dots, I_n$ such that $T_{I_k} \le T_{I_{k+1}}$ for all $k = 1, \dots, n-1$.
- Denote $(\hat\Theta_k, \hat T_k) \leftarrow (\Theta_{I_k}, T_{I_k})$.

(ii) For all unique values $\epsilon \in \{\hat T_1, \dots, \hat T_n\}$, let $m_\epsilon := \max\{k \ge 1 : \hat T_k \le \epsilon\}$, and define

$$E_{\epsilon_0,\epsilon}(f) := \frac{1}{m_\epsilon} \sum_{k=1}^{m_\epsilon} f(\hat\Theta_k), \qquad \text{and} \qquad S_{\epsilon_0,\epsilon}(f) := \frac{1}{m_\epsilon^2} \sum_{k=1}^{m_\epsilon} \big[f(\hat\Theta_k) - E_{\epsilon_0,\epsilon}(f)\big]^2.$$

(and for $\hat T_k < \epsilon < \hat T_{k+1}$, let $E_{\epsilon_0,\epsilon}(f) := E_{\epsilon_0,\hat T_k}(f)$ and $S_{\epsilon_0,\epsilon}(f) := S_{\epsilon_0,\hat T_k}(f)$.)

The sorting in Algorithm 3(i) may be performed in $O(n \log n)$ time, and $E_{\epsilon_0,\epsilon}(f)$ and $S_{\epsilon_0,\epsilon}(f)$ may all be calculated in $O(n)$ time by forming appropriate cumulative sums.

Theorem 5 below details consistency of $E_{\epsilon_0,\epsilon}(f)$, and relates $S_{\epsilon_0,\epsilon}(f)$ to the limiting variance, in case the following (well-known) condition ensuring a central limit theorem holds:

Assumption 4 (Finite integrated autocorrelation)**.** Suppose that $\mathbb{E}_{\pi_\epsilon}[f^2(\Theta)] < \infty$ and $\sum_{k\ge1} \rho_k^{(\epsilon_0,\epsilon)}$ is finite, with $\rho_k^{(\epsilon_0,\epsilon)} := \operatorname{Corr}\big(h_{\epsilon_0,\epsilon}(\Theta_0^{(s)}, Y_0^{(s)}), h_{\epsilon_0,\epsilon}(\Theta_k^{(s)}, Y_k^{(s)})\big)$, where $(\Theta_k^{(s)}, Y_k^{(s)})_{k\ge1}$ is a stationary version of the ABC-MCMC(ϵ_0) chain, and

$$h_{\epsilon_0,\epsilon}(\theta, y) := w_{\epsilon_0,\epsilon}(y) f(\theta) \qquad \text{where} \qquad w_{\epsilon_0,\epsilon}(y) := \phi\big(d(y,y^*)/\epsilon\big)/\phi\big(d(y,y^*)/\epsilon_0\big).$$

Theorem 5. *Suppose $(\Theta_k, T_k)_{k\ge1}$ is the output of ABC-MCMC(ϵ_0), and denote by $E_{\epsilon_0,\epsilon}^{(n)}(f)$ and $S_{\epsilon_0,\epsilon}^{(n)}(f)$ the estimators in Definition 2. If $(\Theta_k, T_k)_{k\ge1}$ is ψ-irreducible, then, for any $\epsilon \in (0, \epsilon_0)$, we have as $n \to \infty$:*

(i) $E_{\epsilon_0,\epsilon}^{(n)}(f) \to \mathbb{E}_{\pi_\epsilon}[f(\Theta)]$ almost surely, whenever the expectation is finite.

(ii) Under Assumption 4, $\sqrt{n}\big(E_{\epsilon_0,\epsilon}^{(n)}(f) - \mathbb{E}_{\pi_\epsilon}[f(\Theta)]\big) \to N\big(0, v_{\epsilon_0,\epsilon}(f)\tau_{\epsilon_0,\epsilon}(f)\big)$ in distribution, where

$$\tau_{\epsilon_0,\epsilon}(f) := \Big(1 + 2\sum_{k\ge1} \rho_k^{(\epsilon_0,\epsilon)}\Big) \in [0,\infty), \qquad \text{and} \qquad nS_{\epsilon_0,\epsilon}^{(n)}(f) \xrightarrow{a.s.} v_{\epsilon_0,\epsilon}(f) \in [0,\infty).$$

Proof of Theorem 5 is given in Appendix A. Based on Theorem 5, we suggest to report the following approximate confidence intervals for the suggested estimators:

Algorithm 6. Suppose $(\Theta_k, T_k)_{k=1,\dots,n}$ is the output of ABC-MCMC(ϵ_0) and $f : \Theta \to \mathbb{R}$ is a function, then for any $\epsilon \le \epsilon_0$:

(i) Calculate $E_{\epsilon_0,\epsilon}(f)$ and $S_{\epsilon_0,\epsilon}(f)$ as in Definition 2 (or in Algorithm 3).

(ii) Calculate $\hat\tau_{\epsilon_0}(f)$, an estimate of the integrated autocorrelation of $\big(f(\Theta_k)\big)_{k=1,\dots,n}$.

(iii) Report the confidence interval

$$\Big[E_{\epsilon_0,\epsilon}(f) \pm \beta\sqrt{S_{\epsilon_0,\epsilon}(f)\hat{\tau}_{\epsilon_0}(f)}\Big],$$

where $\beta > 0$ corresponds to the desired normal quantile.

The classical choice for $\hat{\tau}_{\epsilon_0}(f)$ in Algorithm 6(ii) is windowed autocorrelation, $\hat{\tau}_{\epsilon_0}(f) = \sum_{k=-\infty}^{\infty} \omega(k)\hat{\rho}_k$, with some $0 \le \omega(k) \le 1$, where $\hat{\rho}_k$ is the sample autocorrelation of $\big(f(\Theta_k)\big)$ (cf. Geyer, 1992), but also more sophisticated techniques for the calculation of the asymptotic variance have been suggested (e.g. Flegal and Jones, 2010).

Because computing an estimate of $\tau_{\epsilon_0,\epsilon}(f)$ is computationally demanding, and because such an estimate is likely to be unstable for small ϵ, Algorithm 6 is based on the use of $\hat{\tau}_{\epsilon_0}(f)$ as a common autocorrelation for all $\epsilon \le \epsilon_0$. This relies on the approximation $\tau_{\epsilon_0,\epsilon}(f) \lessapprox \tau_{\epsilon_0,\epsilon_0}(f)$, which may not always be entirely accurate, but likely to be reasonable, as illustrated by Theorem 7 in Section 3 below.

We remark that, although we focus on the case of using a common cut-off for both the ABC-MCMC and post-correction, one could also consider using two different cut-offs. The extension to Definition 2 is straightforward, and Algorithm 3 holds with simple post-correction cut-off, under a support condition.

3. Confidence interval and efficiency

The following result, whose proof is given in Appendix A, gives an expression for the integrated autocorrelation in case of simple cut-off.

Theorem 7. *Suppose Assumption 4 holds and $\phi = \phi_{\mathrm{simple}}$, then*

$$\tau_{\epsilon_0,\epsilon}(f) - 1 = \frac{\big(\check{\tau}_{\epsilon_0,\epsilon}(f) - 1\big)\mathrm{var}_{\pi_{\epsilon_0}}(f_{\epsilon_0,\epsilon}) + 2\int \pi_{\epsilon_0}(\theta)\bar{w}_{\epsilon_0,\epsilon}(\theta)\big(1-\bar{w}_{\epsilon_0,\epsilon}(\theta)\big)\frac{r_{\epsilon_0}(\theta)}{1-r_{\epsilon_0}(\theta)}f^2(\theta)\mathrm{d}\theta}{\mathrm{var}_{\pi_{\epsilon_0}}(f_{\epsilon_0,\epsilon}) + \int \pi_{\epsilon_0}(\theta)\bar{w}_{\epsilon_0,\epsilon}(\theta)\big(1-\bar{w}_{\epsilon_0,\epsilon}(\theta)\big)f^2(\theta)\mathrm{d}\theta},$$

where $\bar{w}_{\epsilon_0,\epsilon}(\theta) := L_\epsilon(\theta)/L_{\epsilon_0}(\theta)$, $f_{\epsilon_0,\epsilon}(\theta) := f(\theta)\bar{w}_{\epsilon_0,\epsilon}(\theta)$, $\check{\tau}_{\epsilon_0,\epsilon}(f)$ is the integrated autocorrelation of $\{f_{\epsilon_0,\epsilon}(\Theta_k^{(s)})\}_{k\ge1}$ and $r_{\epsilon_0}(\theta)$ the rejection probability of the ABC-MCMC(ϵ_0) chain at θ.

We next discuss how this loosely suggests that $\tau_{\epsilon_0,\epsilon}(f) \lessapprox \tau_{\epsilon_0,\epsilon_0}(f)$. Note that $\bar{w}_{\epsilon_0,\epsilon_0} \equiv 1$, and under suitable regularity conditions both $\bar{w}_{\epsilon_0,\epsilon}(\theta)$ and $\check{\tau}_{\epsilon_0,\epsilon}(f)$ are continuous with respect to ϵ, and $\bar{w}_{\epsilon_0,\epsilon}(\theta) \to 0$ as $\epsilon \to 0$. Then, for $\epsilon \approx \epsilon_0$, we have $\bar{w}_{\epsilon_0,\epsilon} \approx 1$ and therefore $\tau_{\epsilon_0,\epsilon_0}(f) \approx \tau_{\epsilon_0,\epsilon}(f)$. For small ϵ, the terms with $\mathrm{var}_{\pi_{\epsilon_0}}(f_{\epsilon_0,\epsilon})$ are of order $O(\bar{w}^2_{\epsilon_0,\epsilon})$, and are dominated by the other terms of order $O(\bar{w}_{\epsilon_0,\epsilon})$. The remaining ratio may be written as

$$\frac{2\int \pi_{\epsilon_0}(\theta)\bar{w}_{\epsilon_0,\epsilon}(\theta)\big(1-\bar{w}_{\epsilon_0,\epsilon}(\theta)\big)\frac{r_{\epsilon_0}(\theta)}{1-r_{\epsilon_0}(\theta)}f^2(\theta)\mathrm{d}\theta}{\int \pi_{\epsilon_0}(\theta)\bar{w}_{\epsilon_0,\epsilon}(\theta)\big(1-\bar{w}_{\epsilon_0,\epsilon}(\theta)\big)f^2(\theta)\mathrm{d}\theta} = 2\mathbb{E}_{\pi_{\epsilon_0}}\Big[\bar{g}^2_{\epsilon_0,\epsilon}(\Theta)\frac{r_{\epsilon_0}(\Theta)}{1-r_{\epsilon_0}(\Theta)}\Big],$$

where $\bar{g}_{\epsilon_0,\epsilon} \propto (\bar{w}_{\epsilon_0,\epsilon}(1-\bar{w}_{\epsilon_0,\epsilon}))^{1/2} f$ with $\pi_{\epsilon_0}(\bar{g}^2_{\epsilon_0,\epsilon}) = 1$. If $r_{\epsilon_0}(\theta) \le r_* < 1$, then the term is upper bounded by $2r_*(1-r_*)^{-1}$, and we believe it to be often less than $\tau_{\epsilon_0,\epsilon_0}(f)$, because the latter expression is similar to the contribution of rejections to the integrated autocorrelation; see the proof of Theorem 7.

For general ϕ, it appears to be hard to obtain similar theoretical result, but we expect the approximation to be still sensible. Theorem 7 relies on $Y_k^{(s)}$ being independent of $(\Theta_k^{(0)}, Y_k^{(0)})$ conditional on $\Theta_k^{(s)}$, assuming at least single acceptance. This is not true with other cut-offs, but we believe that the dependence of $Y_k^{(s)}$ from $(\Theta_0^{(s)}, Y_0^{(s)})$ given $\Theta_k^{(s)}$ is generally weaker than dependence of $\Theta_k^{(s)}$ and $\Theta_0^{(s)}$, suggesting similar behaviour.

Let us state next a general upper bound for the IS-corrected ABC-MCMC as we suggest, with respect to a direct ABC-MCMC with a smaller tolerance.

Theorem 8. *For any $\epsilon \leq \epsilon_0$, denote by $\sigma^2_{\epsilon_0,\epsilon}(f) := v_{\epsilon_0,\epsilon}(f)\tau_{\epsilon_0,\epsilon}(f)$ the asymptotic variance of the estimator of Definition 2 (see Theorem 5(ii)). Define*

$$w_{\epsilon_0,\epsilon}(y) := \frac{\phi\big(d(y,y^*)/\epsilon\big)}{\phi\big(d(y,y^*)/\epsilon_0\big)}, \quad c_\epsilon := \int \mathrm{pr}(\theta)L_\epsilon(\theta)\mathrm{d}\theta, \quad \tilde{\pi}_\epsilon(\theta,y) := \frac{\mathrm{pr}(\theta)g(y \mid \theta)K_\epsilon(y,y^*)}{c_\epsilon},$$

and denote $\bar{f}(\theta) = f(\theta) - E_{\pi_\epsilon}[f(\Theta)]$. Then for any $\epsilon \leq \epsilon_0$,

$$\sigma^2_{\epsilon_0,\epsilon}(f) \leq \frac{c_{\epsilon_0}}{c_\epsilon}\Big[\sigma^2_\epsilon(f) + \tilde{\pi}_\epsilon\big(\bar{f}^2(1-w_{\epsilon_0,\epsilon})\big)\Big],$$

where $\sigma^2_\epsilon(f) := \sigma^2_{\epsilon,\epsilon}(f)$ is the asymptotic variance of the direct ABC-MCMC(ϵ).

Theorem 8 follows directly from (Franks and Vihola, 2017, Corollary 4). The upper bound guarantees that a moderate correction, that is, ϵ close to ϵ_0 and c_{ϵ_0} close to c_ϵ, is nearly as efficient as direct ABC-MCMC. Indeed, typically $w_{\epsilon_0,\epsilon} \to 1$ and $c_\epsilon \to c_{\epsilon_0}$ as $\epsilon \to \epsilon_0$, in which case Theorem 8 implies $\limsup_{\epsilon\to\epsilon_0} \sigma^2_{\epsilon_0,\epsilon}(f) \leq \sigma^2_\epsilon(f)$. However, as $\epsilon \to 0$, the bound becomes less informative.

4. A tolerance adaptive ABC-MCMC algorithm

We propose Algorithm 9 below to adapt the tolerance ϵ_0 in the ABC-MCMC during burn-in, in order to obtain a user-specified overall acceptance rate α^*. Together with the approach based on post-correction of Section 2, we thus obtain an automated ABC inference solution that does not require prior choice of an ϵ_0.

In Algorithm 9, we assume that a desired acceptance rate $\alpha^* \in (0,1)$ is specified. We used $\alpha^* = 0.1$ in our experiments, and discuss this choice later. We also assume a choice of decreasing positive step sizes $(\gamma_k)_{k\geq 1}$. We used $\gamma_k = k^{-2/3}$ in our experiments. For convenience, we denote the distance distribution here as $T \sim Q_\theta(\,\cdot\,)$, where $T := d(Y,y^*)$ for $Y \sim g(\,\cdot\,|\theta)$.

Algorithm 9 (TA(n_b, α^*))**.** Suppose $\Theta_0 \in \mathsf{T}$ is a starting value with $\mathrm{pr}(\Theta_0) > 0$.

1. Initialise $\epsilon_0 := T_0$ where $T_0 \sim Q_{\Theta_0}(\,\cdot\,)$ and $T_0 > 0$.
2. For $k = 0, \ldots, n_b - 1$, iterate:
   (i) Draw $\Theta'_k \sim q(\Theta_k, \,\cdot\,)$.
   (ii) Draw $T'_k \sim Q_{\Theta'_k}(\,\cdot\,)$.
   (iii) Accept, by setting $(\Theta_{k+1}, T_{k+1}) \leftarrow (\Theta'_k, T'_k)$, with probability

$$\alpha_{\epsilon_k}(\Theta_k, T_k; \Theta'_k, T'_k) := \min\left\{1, \frac{\mathrm{pr}(\Theta'_k)q(\Theta'_k,\Theta_k)\phi(T'_k/\epsilon_k)}{\mathrm{pr}(\Theta_k)q(\Theta_k,\Theta'_k)\phi(T_k/\epsilon_k)}\right\} \tag{2}$$

   and otherwise reject, by setting $(\Theta_{k+1}, T_{k+1}) \leftarrow (\Theta_k, T_k)$.
   (iv) $\log \epsilon_{k+1} \leftarrow \log \epsilon_k + \gamma_{k+1}\big(\alpha^* - \alpha'_{\epsilon_k}(\Theta_k, \Theta'_k, T'_k)\big)$.
3. Output $(\Theta_{n_b}, \epsilon_{n_b})$.

The following simple conditions suffice for convergence of the adaptation

Assumption 10. Suppose $\phi = \phi_{\mathrm{simple}}$ and the following hold:

(i) $\gamma_k := Ck^{-r}$ with $r \in (\frac{1}{2}, 1]$ and $C > 0$ a constant.
(ii) The domain $\mathsf{T} \subset \mathbb{R}^{n_\theta}$, $n_\theta \geq 1$, is a nonempty open set.
(iii) $\mathrm{pr}(\theta)$ is uniformly bounded on T, and $\mathrm{pr}(\theta) = 0$ for $\theta \notin \mathsf{T}$.
(iv) q is a uniformly bounded density, and $q \geq \delta_q$ on T^2, $\delta_q > 0$.
(v) $Q_\theta(\mathrm{d}t)$ admits a uniformly bounded density $Q_\theta(t)$.

(vi) ϵ_k stays in a set $[a, b]$ almost surely, where $0 < a \leq b < +\infty$.
(vii) $\int \mathrm{pr}(\mathrm{d}\theta) L_\epsilon(\theta) > 0$ for all $\epsilon \in [a, b]$.

Theorem 11. *Under Assumption 10, the expected value of the acceptance probability (2), taken with respect to the stationary measure of the chain, converges to α^*.*

Proof of Theorem 11 will follow from the more general Theorem 13 of Appendix B. Theorem 13 is phrased for geometrically ergodic chains on possibly unbounded domains without the lower bound in Assumption 10(iv). See Appendix C for the proofs of both theorems.

In practice, the tolerance adaptation is the most straightforward to apply with a symmetric random walk proposal q adapted simultaneously with proposal covariance adaptation (Andrieu and Moulines, 2006; Haario et al., 2001) (see Algorithm 22 of Appendix D for a detailed description of the resulting algorithm). Such simultaneous use of different optimisation criteria within adaptive MCMC has been discussed, for example, in the review (Andrieu and Thoms, 2008). While we do not consider Algorithm 22 explicitly in our theoretical analysis, our results could be elaborated, along the lines of Andrieu and Moulines (2006), to accommodate Algorithm 22 in detail.

In the standard Adaptive Metropolis algorithm (Haario et al., 2001), the limiting acceptance rate is often around 0.234 (Roberts et al., 1997). In the ABC-MCMC context, this acceptance rate would be reached if the tolerance would be made infinite, and if the prior distribution would be regular enough (e.g. Gaussian). Because the mean acceptance rate of ABC-MCMC typically decreases when tolerance is decreased (see Lemma 15 of Appendix C in case of ϕ_{simple}), and because the likelihood approximation must be reasonable, the desired acceptance rate should be substantially lower than 0.234.

As ABC-MCMC may be interpreted as an instance of pseudo-marginal MCMC, for which there are certain conditions under which the optimal acceptance rate of about 0.07 is reached (Sherlock et al., 2015), one could take this as a guideline as well for the ABC-MCMC. However, the context of Sherlock et al. (2015) is quite dissimilar to that of ABC-MCMC, and so we decided to push the acceptance rate a little higher to ensure sufficient mixing. As well, we do subsequent post-correction, which is further justification for a slightly inflated tolerance and therefore acceptance rate.

5. Experiments

We experiment with our methods on two models, a lightweight Gaussian toy example, and a Lotka-Volterra model. Our experiments aim at providing information regarding the following questions:

- Can ABC-MCMC with larger tolerance and post-correction to a desired tolerance deliver more accurate results than direct application of ABC-MCMC?
- Does our approximate confidence interval appear reliable?
- How well does the adaptive ABC-MCMC work in practice?

In all our experiments, we apply the Adaptive Metropolis (Andrieu and Moulines, 2006; Haario et al., 2001) covariance adaptation, which is run during the whole simulation, using an identity covariance initially.

Regarding our first question, we investigate running the ABC-MCMC starting near the posterior mode with different pre-selected tolerances, both selected in a preliminary pilot experiment. We first attempted to perform the experiments by initialising the chains from independent samples of the prior distribution, but in this case, most of the chains did not accept a single move during the whole run. In contrast, our experiments with tolerance adaptation do start from initial points drawn from the prior distribution, and both the

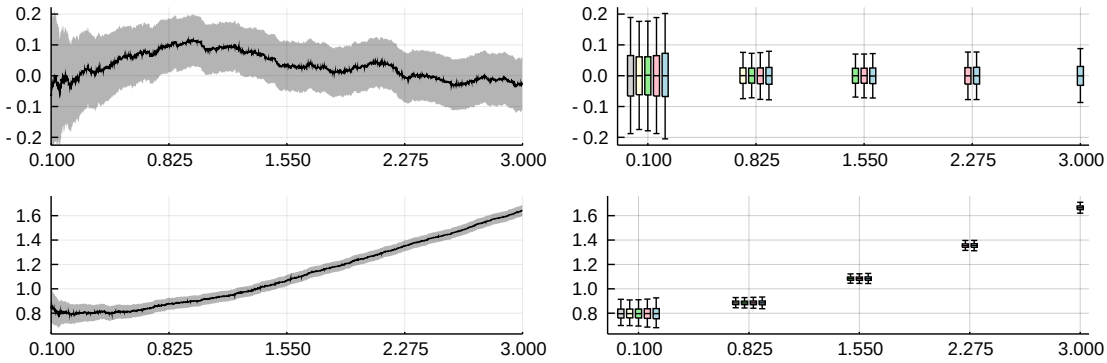


FIGURE 1. Gaussian model with simple cut-off.

tolerances and the covariances are adjusted fully automatically by our algorithm. The latter assumes no prior information of the model at all, which we aim at.

In our tests about confidence intervals, we employ a simple 'automatic window' estimator of integrated autocorrelation of the form $\hat{\tau}_{\epsilon_0,M} = 1+2\sum_{i=1}^{M}\hat{\rho}_k(f)$, where $\hat{\rho}_k$ are lag-k sample autocorrelations, and where M is the smallest positive integer such that $M \geq 5\hat{\tau}_{\epsilon_0,M}$ (Sokal, 1996).

When running the covariance adaptation alone, we employ the covariance adaptation of Andrieu and Moulines (2006) with step size n^{-1}, which behaves similar to the original Adaptive Metropolis algorithm of Haario et al. (2001). In case we apply tolerance adaptation, we use step size $n^{-2/3}$ for both the tolerance adaptation and for the covariance adaptation. Slower decaying step sizes such as this often behave better with acceptance rate adaptation (cf. Vihola, 2012, Remark 3).

All the experiments are implemented in Julia (Bezanson et al., 2017), and the codes are available in `https://bitbucket.org/mvihola/abc-mcmc`.

5.1. **One-dimensional Gaussian model.** Our first model is a toy model with $\mathrm{pr}(\theta) = N(\theta; 0, 30^2)$ and $Y = \Theta + Z$, where Z is standard Gaussian random variable. The true posterior without ABC approximation is Gaussian. While this scenario is clearly academic, the prior is far from the posterior, which we believe to be common in practice. It is clear that π_ϵ has zero mean for all ϵ, and also that the distribution π_ϵ is spread wider for bigger ϵ. We experiment with both simple cut-off ϕ_{simple} and Gaussian cut-off $\phi_{\mathrm{Gauss}}(t) := e^{-t^2/2}$.

We run the experiments with 10,000 independent chains, each for 11,000 iterations including 1,000 burn-in. The chains were always started from $\theta_0 = 0$. Figures 1 and 2 show results of the same experiments with simple and Gaussian cut-off. On the left, a single realisation of the estimates and confidence intervals calculated for all $\epsilon \leq \epsilon_0 = 3$ are shown for functions $f_1(\theta) = \theta$ (above) and $f_2(\theta) = |\theta|$ (below). The figures on the right show box plots of the final estimators calculated for each chain, for five equispaced tolerance values between 0.1 and 3.0 (x axis labels indicate these tolerances). The leftmost box plot in each group corresponds to the direct ABC-MCMC targeting that tolerance, and the rightmost box plot corresponds to the post-corrected estimators from the ABC-MCMC with $\epsilon_0 = 3.0$, and the second from the right with $\epsilon_0 = 2.275$ etc. The colour indicates the ϵ_0. Some post-corrected estimates appear to be slightly more accurate than ABC-MCMC, and in the results suggest that $\epsilon_0 = 0.82$ might be a good choice, if the desired tolerance is $\epsilon = 0.1$.

Table 1 indicates the frequencies of the calculated 95% confidence intervals containing the 'ground truth', over the 10,000 independent experiments, as well as acceptance rates. The ground truth for $\mathbb{E}_{\pi_\epsilon}[f_1(\Theta)]$ is known to be zero for all ϵ, and the overall mean of all the

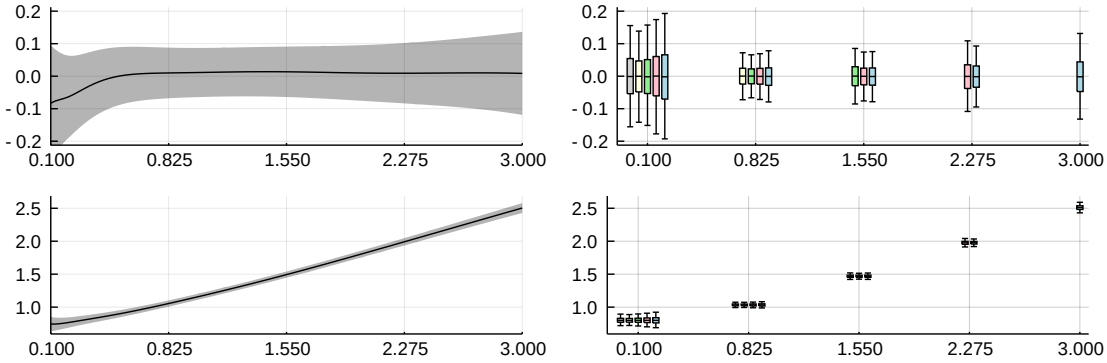


FIGURE 2. Gaussian model with Gaussian cut-off.

TABLE 1. Frequencies of the 95% confidence intervals containing the ground truth in the Gaussian model.

Cutoff		$f(x) = x$					$f(x) = \lvert x \rvert$					Acc.
	$\epsilon_0 \setminus \epsilon$	0.10	0.82	1.55	2.28	3.00	0.10	0.82	1.55	2.28	3.00	rate
ϕ_{simple}	0.1	0.93					0.93					0.03
	0.82	0.97	0.95				0.95	0.94				0.22
	1.55	0.97	0.97	0.95			0.96	0.95	0.95			0.33
	2.28	0.98	0.97	0.96	0.95		0.96	0.96	0.96	0.95		0.4
	3.0	0.98	0.98	0.97	0.97	0.95	0.96	0.96	0.96	0.95	0.95	0.43
ϕ_{Gauss}	0.1	0.93					0.93					0.05
	0.82	0.94	0.95				0.92	0.95				0.29
	1.55	0.94	0.94	0.95			0.94	0.94	0.95			0.38
	2.28	0.95	0.95	0.95	0.95		0.95	0.95	0.96	0.95		0.41
	3.0	0.95	0.95	0.95	0.95	0.95	0.95	0.96	0.95	0.95	0.95	0.42

calculated estimates is used as the ground truth for $\mathbb{E}_{\pi_\epsilon}[f_2(\Theta)]$. The frequencies appear close to ideal with the post-correction approach, being slightly pessimistic in case of simple cut-off as anticipated by the theoretical considerations (see Theorem 7 and discussion below).

Figure 3 shows progress of tolerance adaptations during the burn-in, and histogram of the mean acceptance rates of the chain after burn-in. The lines on the left show the median, and the shaded regions indicate the 50%, 75%, 95% and 99% quantiles. The figures indicate concentration, but suggest that the adaptation has not fully converged yet. This is also indicated by the mean acceptance rate over all realisations, which are 0.17 and 0.12 with simple and Gaussian cutoff, respectively. Table 2 shows root mean square errors from the ground truth, for both the fixed tolerance estimators, and the adaptive algorithms, for tolerance $\epsilon = 0.1$. Here, only the adaptive chains with final tolerance ≥ 0.1 were included (9,997 and 9,996 out of 10,000 chains for the simple and Gaussian cut-offs, respectively).

5.2. **Lotka-Volterra model.** Our second experiment is a Lotka-Volterra model suggested in (Boys et al., 2008), and also analysed in the ABC context in (Fearnhead and Prangle, 2012). The model is a Markov process $(X_t, Y_t)_{t\geq 0}$ of counts, corresponding to a reaction network $X \to 2X$ with rate θ_1, $X + Y \to 2Y$ with rate θ_2 and $Y \to \emptyset$ with rate θ_3. The reaction rates $\theta = (\theta_1, \theta_2, \theta_3)^\top$ are parameters, which we equip here with a uniform prior, $(\log\theta_1, \log\theta_2, \log\theta_3)^\top \sim U([-6, 0]^3)$. The data is a simulated trajectory from the model

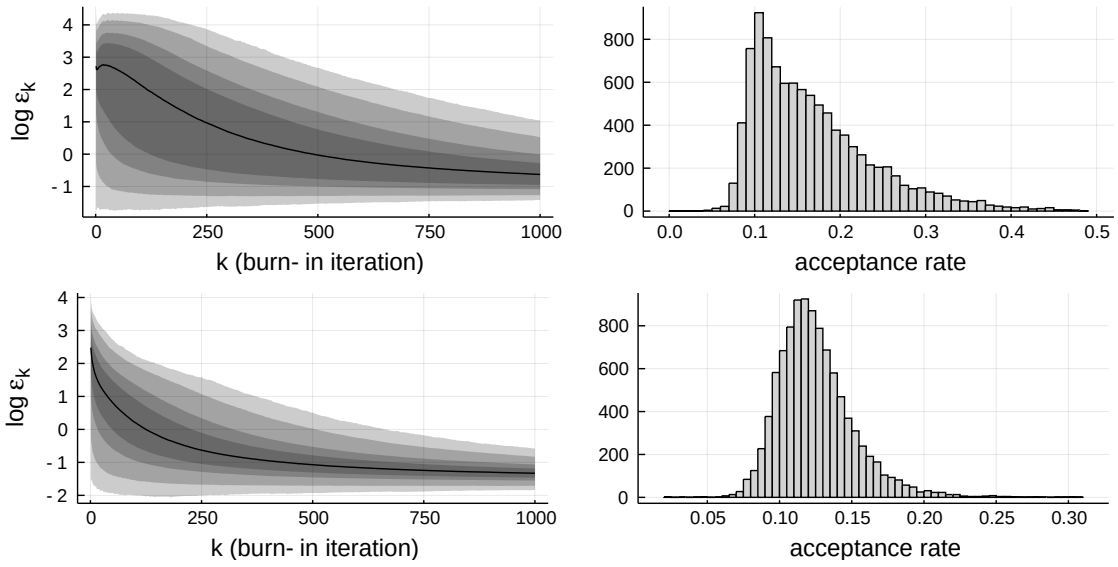


FIGURE 3. Progress of tolerance adaptation (left) and histogram of acceptance rates (right) in the Gaussian model experiment with simple cutoff (top) and Gaussian cutoff (bottom).

TABLE 2. RMSEs ($\times 10^{-2}$) with fixed tolerance and with the adaptive algorithms in the Gaussian model, for tolerance $\epsilon = 0.1$.

	ϕ_{simple}						ϕ_{Gauss}					
	Fixed tolerance					Adapt	Fixed tolerance					Adapt
ϵ_0	0.1	0.82	1.55	2.28	3.0	0.64	0.1	0.82	1.55	2.28	3.0	0.28
x	9.68	8.99	9.21	9.67	10.36	9.16	7.97	7.12	7.82	8.94	9.93	9.26
$\lvert x\rvert$	5.54	5.38	5.5	5.85	6.21	5.44	4.47	4.22	4.68	5.26	5.95	5.46

with $\theta = (0.5, 0.0025, 0.3)^\top$ until time 40. The ABC is based on Euclidean distance of a six-dimensional summary statistic, which consists of:

- Sample autocorrelation of $X_t^{(1)}$ at lag 10, multiplied by 100.
- 10% and 90% (time-averaged) quantiles of both $X_t^{(1)}$ and $X_t^{(2)}$.
- Number of jumps (or events), divided by 10.

The summary statistics are then $(-51.0711, 29.0, 304.0, 65.0, 404.0, 749.4)^\top$ for the observed series.

We first run comparisons similar to Section 5.1, but now only with 1,000 independent chains and simple cut-off. We investigate the effect of post-correction, with 20,000 samples, including 10,000 burn-in, for each chain. The MCMC is run on log-transformed θ, and all chains were started from near the mode, from $\log \theta = (-0.55, -5.77, -1.09)^\top$. Figure 4 and Table 3 show similar comparisons as in Section 5.1. The results suggest that post-corrected ABC does provide slightly more accurate estimators, particularly with smaller tolerances.

In addition, we experiment with the tolerance adaptation, using also 20,000 samples out of which 10,000 are burn-in. Figure 5 shows the progress of the log-tolerance during the burn-in, and histogram of the realised mean acceptance rates during the estimation phase. The realised acceptance rates are concentrated around the desired 10%, with mean 0.10. Table 4 shows RMSEs of both the fixed tolerance ABC-MCMC outputs and with tolerance

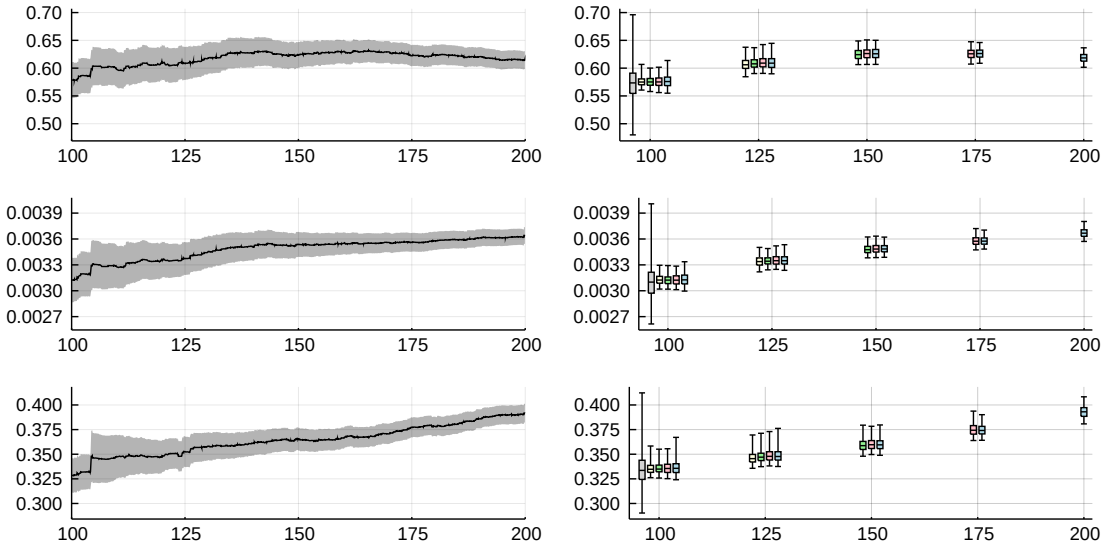


FIGURE 4. Lotka-Volterra model with simple cut-off.

TABLE 3. Frequencies of the 95% confidence intervals in the Lotka-Volterra experiment and mean acceptance rates.

	$f(\theta)=\theta_1$					$f(\theta)=\theta_2$					$f(\theta)=\theta_3$					
ϵ_0 \ ϵ	100.0	125.0	150.0	175.0	200.0	100.0	125.0	150.0	175.0	200.0	100.0	125.0	150.0	175.0	200.0	Acc. rate
100.0	0.59					0.55					0.53					0.04
125.0	0.97	0.88				0.97	0.88				0.96	0.81				0.11
150.0	0.99	0.97	0.92			0.99	0.97	0.92			0.99	0.95	0.88			0.13
175.0	0.99	0.97	0.96	0.92		0.99	0.98	0.96	0.92		0.99	0.98	0.97	0.92		0.16
200.0	0.98	0.98	0.98	0.96	0.94	0.99	0.99	0.98	0.97	0.92	0.98	0.97	0.96	0.96	0.92	0.18

adaptation. Again, only the adaptive chains with final tolerance ≥ 100.0 were included (999 out of 1,000 chains).

In this case, the chains run with the tolerance adaptation led to better results than those run only with the covariance adaptation (and fixed tolerance). This perhaps surprising result may be due to the initial behaviour of the covariance adaptation, which may be unstable when there are many rejections. Different initialisation strategies, for instance following (Haario et al., 2001, Remark 2), might lead to more stable behaviour compared to using the adaptation of Andrieu and Moulines (2006) from the start, as we do. The different step size sequences (n^{-1} and $n^{-2/3}$) could also play a rôle. We repeated the experiment for the chains with fixed tolerances, but now with covariance adaptation step size $n^{-2/3}$. This led to more stable behaviour of the ABC-MCMC with tolerance $\epsilon_0 = 100.0$. In any case, also here, the adaptive ABC-MCMC using the tolerance adaptation delivered slightly better results (see supplementary results in Appendix E).

6. Discussion

We believe our approach consisting of ABC-MCMC with post-processing is a useful addition, and complements some earlier and related work. As previously mentioned, trimming

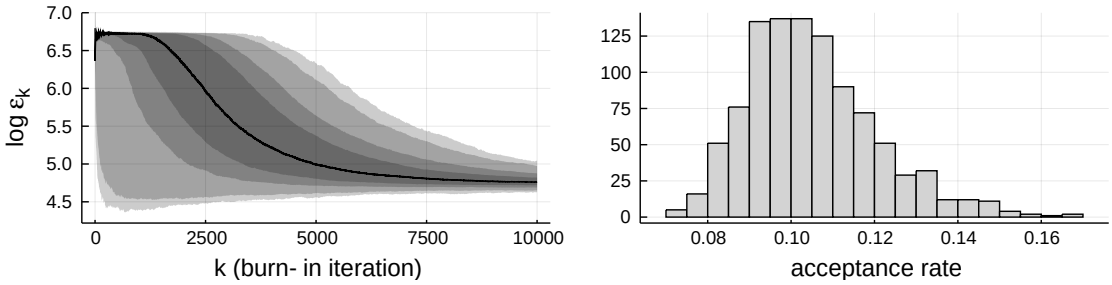


FIGURE 5. Progress of tolerance adaptation (left) and histogram of acceptance rates (right) in the Lotka-Volterra experiment.

TABLE 4. RMSEs with fixed tolerance and with the adaptive algorithms in the Lotka-Volterra model, for tolerance $\epsilon = 100$.

	Fixed tolerance					Adapt
ϵ_0	100.0	125.0	150.0	175.0	200.0	119.1
θ_1 $(\times 10^{-2})$	5.07	1.39	1.13	1.31	1.74	0.79
θ_2 $(\times 10^{-4})$	3.15	0.85	0.69	0.74	1.02	0.54
θ_3 $(\times 10^{-2})$	2.94	1.09	0.87	0.85	1.39	0.51

of ABC-MCMC output to finer tolerances has been considered earlier (e.g. Wegmann et al., 2009). Our experimental results suggest that this can indeed be beneficial, and our confidence interval may make the approach more appealing in practice.

Another related approach by Bortot et al. (2007) makes tolerance an auxiliary variable with a user-specified prior, and ABC-MCMC is run targeting the joint posterior of parameter and tolerance. While this approach avoids tolerance selection, we believe that our approach, where the effect of tolerance can be investigated explicitly, can be helpful in interpretation of the ABC posterior. In fact, Bortot et al. (2007) also provide tolerance-dependent analysis, but we believe that our estimators, with associated confidence intervals, have a more immediate interpretation.

Automatic selection of tolerance in ABC-MCMC has been considered earlier in Ratmann et al. (2007), who propose an algorithm based on tempering and a cooling schedule. It has been remarked by Sisson and Fan (2018) that acceptance rate based adaptation could be used to deal with the choice of a suitable tolerance. Based on our experiments, the adaptive ABC-MCMC we present in this paper appears to perform well in practice, and provides reliable results with post-correction. The tolerance adaptation also seems to benefit the covariance adaptation in the early phases. For the adaptive ABC-MCMC to work efficiently, the MCMC chains must be taken relatively long, rendering the approach difficult for computationally demanding models. However, we believe that our approach using adaptive ABC-MCMC provides a straightforward way to do inference with ABC models.

Our estimators, and their uncertainty estimators, could also turn out to be useful in the regression adjustment context (Beaumont et al., 2002; Blum, 2010; Wegmann et al., 2009). We did not consider such adjustments, but note that approximate normality and the confidence bounds may be used to derive an appropriately weighted estimator that reflects the uncertainty of the estimators.

We conclude with a brief discussion of certain extensions of the suggested post-correction method. The first extension is based on 'recycling' the rejected samples in the estimator (Ceperley et al., 1977). This may improve the accuracy (but can also reduce accuracy in

certain pathological cases; see Delmas and Jourdain (2009)). The 'waste recycling' estimator is

$$E^{\mathrm{WR}}_{\epsilon_0,\epsilon}(f) := \sum_{k=1}^{n} W_k^{(\epsilon_0,\epsilon)} \big[\alpha_{\epsilon_0}(\Theta_k, Y_k; \tilde{\Theta}_{k+1}, \tilde{Y}_{k+1}) f(\tilde{\Theta}_{k+1}) + [1 - \alpha_{\epsilon_0}(\Theta_k, Y_k; \tilde{\Theta}_{k+1}, \tilde{Y}_{k+1})] f(\Theta_k)\big].$$

When $E_{\epsilon_0,\epsilon}(f)$ is consistent under Theorem 5(i), this is also a consistent estimator. Namely, as in the proof (in Appendix A) of Theorem 5, we find that $(\Theta_k, Y_k, \tilde{\Theta}_{k+1}, Y_{k+1})_{k\geq 1}$ is a Harris recurrent Markov chain with invariant distribution

$$\hat{\pi}_{\epsilon_0}(\theta, y, \tilde{\theta}, \tilde{y}) = \tilde{\pi}_{\epsilon_0}(\theta, y)\tilde{q}(\theta, y; \tilde{\theta}, \tilde{y}),$$

and $\hat{\pi}_\epsilon(\theta, y, \tilde{\theta}, \tilde{y})/\hat{\pi}_{\epsilon_0}(\theta, y, \tilde{\theta}, \tilde{y}) = c_\epsilon w_{\epsilon_0,\epsilon}(y)$, where $\tilde{q}(\theta, y; \theta', y') = q(\theta, \theta')g(y' \mid \theta')$. Therefore, $E^{\mathrm{WR}}_{\epsilon_0,\epsilon}(f)$ is a strongly consistent estimator of

$$\mathbb{E}_{\hat{\pi}_\epsilon}\big[\alpha_{\epsilon_0}(\Theta, Y; \tilde{\Theta}, \tilde{Y}) f(\tilde{\Theta}) + [1 - \alpha_{\epsilon_0}(\Theta, Y; \tilde{\Theta}, \tilde{Y})] f(\Theta)\big] = \mathbb{E}_{\pi_\epsilon}[f(\Theta)].$$

See (Rudolf and Sprungk, 2018; Schuster and Klebanov, 2018) for alternative waste recycling estimators based on importance sampling analogues.

Another extension, which could be considered, is about enhancing the accuracy of the estimator with smaller values of ϵ, by performing further simulations from the model (which may be calculated in parallel for different Θ_k). Namely, a new estimator may be formed as follows:

$$\hat{E}_{\epsilon_0,\epsilon}(f) = \frac{\sum_{k=1}^{n}\sum_{j=0}^{m} \hat{U}_{k,j}^{(\epsilon_0,\epsilon)} f(\Theta_k)}{\sum_{\ell=1}^{n}\sum_{i=0}^{m} \hat{U}_{\ell,i}^{(\epsilon_0,\epsilon)}}, \qquad \hat{U}_{k,0}^{(\epsilon_0,\epsilon)} := U_k^{(\epsilon_0,\epsilon)} \qquad \text{and} \qquad \hat{U}_{k,j}^{(\epsilon_0,\epsilon)} := \frac{\hat{N}_k \phi(\hat{T}_{k,j}/\epsilon)}{\phi(T_k/\epsilon_0)},$$

for $j \geq 1$, where $\hat{N}_k$ is the number of independent random variables $\hat{Z}_1, \hat{Z}_2, \ldots \sim g(\cdot \mid \Theta_k)$ generated before observing $\phi(\hat{T}_{k,\hat{N}_k}/\epsilon_0) > 0$ where $\hat{T}_{k,j} := d(\hat{Z}_j, y^*)$, and $\hat{T}_k := d(\hat{Y}_k, y^*)$ with independent $\hat{Y}_k \sim g(\cdot \mid \Theta_k)$. This ensures that

$$\mathbb{E}[\hat{N}_k \phi(\hat{T}_{k,j}/\epsilon) \mid \Theta_k = \theta, Y_k = y] = \frac{L_\epsilon(\theta)}{\mathbb{P}_{g(\cdot\mid\theta)}\big(\phi\big(d(Y, y^*)/\epsilon_0\big) > 0\big)},$$

which is sufficient to ensure that $\xi_{k,j}(f) := \hat{U}_{k,j}^{(\epsilon_0,\epsilon)} f(\Theta_k)$ is a proper weighting scheme from $\tilde{\pi}_{\epsilon_0}$ to π_ϵ; see (Vihola et al., 2016, Proposition 17(ii)), and consequently the average $\xi_k(f) := (m+1)^{-1}\sum_{j=0}^{m} \xi_{k,j}(f)$ is a proper weighting.

The latter extension, which involves additional simulations as post-processing, is similar to 'lazy ABC', which incorporates a randomised stopping rule for simulation (Prangle, 2015, 2016), and to unbiased 'exact' ABC (Tran and Kohn, 2015), which may lead to estimators which get rid of ϵ-bias entirely, using the debiasing approach lately investigated in (McLeish, 2011; Rhee and Glynn, 2015).

7. Acknowledgements

This work was supported by Academy of Finland (grants 274740, 284513 and 312605). The authors wish to acknowledge CSC, IT Center for Science, Finland, for computational resources. The authors wish to thank Christophe Andrieu for useful discussions.

References

C. Andrieu and É. Moulines. On the ergodicity properties of some adaptive MCMC algorithms. *Ann. Appl. Probab.*, 16(3):1462–1505, 2006.

C. Andrieu and J. Thoms. A tutorial on adaptive MCMC. *Statist. Comput.*, 18(4):343–373, Dec. 2008.

C. Andrieu, É. Moulines, and P. Priouret. Stability of stochastic approximation under verifiable conditions. *SIAM J. Control Optim.*, 44(1):283–312, 2005.

M. Beaumont, W. Zhang, and D. Balding. Approximate Bayesian computation in population genetics. *Genetics*, 162(4):2025–2035, 2002.

J. Bezanson, A. Edelman, S. Karpinski, and V. B. Shah. Julia: A fresh approach to numerical computing. *SIAM review*, 59(1):65–98, 2017.

M. Blum. Approximate Bayesian computation: a nonparametric perspective. *J. Amer. Statist. Assoc.*, 105(491):1178–1187, 2010.

P. Bortot, S. Coles, and S. Sisson. Inference for stereological extremes. *J. Amer. Statist. Assoc.*, 102(477):84–92, 2007.

R. J. Boys, D. J. Wilkinson, and T. B. Kirkwood. Bayesian inference for a discretely observed stochastic kinetic model. *Stat. Comput.*, 18(2):125–135, 2008.

D. Burkholder, B. Davis, and R. Gundy. Integral inequalities for convex functions of operators on martingales. In *Proc. Sixth Berkeley Symp. Math. Statist. Prob*, volume 2, pages 223–240, 1972.

D. Ceperley, G. Chester, and M. Kalos. Monte Carlo simulation of a many-fermion study. *Phys. Rev. D*, 16(7):3081, 1977.

J.-F. Delmas and B. Jourdain. Does waste recycling really improve the multi-proposal Metropolis–Hastings algorithm? an analysis based on control variates. *J. Appl. Probab.*, 46(4):938–959, 2009.

P. Fearnhead and D. Prangle. Constructing summary statistics for approximate Bayesian computation: semi-automatic approximate Bayesian computation. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 74(3):419–474, 2012.

J. M. Flegal and G. L. Jones. Batch means and spectral variance estimators in Markov chain Monte Carlo. *Ann. Statist.*, 38(2):1034–1070, 2010.

J. Franks and M. Vihola. Importance sampling correction versus standard averages of reversible MCMCs in terms of the asymptotic variance. Preprint arXiv:1706.09873v3, 2017.

C. J. Geyer. Practical Markov chain Monte Carlo. *Statist. Sci.*, pages 473–483, 1992.

H. Haario, E. Saksman, and J. Tamminen. An adaptive Metropolis algorithm. *Bernoulli*, 7 (2):223–242, 2001.

J.-M. Marin, P. Pudlo, C. P. Robert, and R. J. Ryder. Approximate Bayesian computational methods. *Statist. Comput.*, 22(6):1167–1180, 2012.

P. Marjoram, J. Molitor, V. Plagnol, and S. Tavaré. Markov chain Monte Carlo without likelihoods. *Proc. Natl. Acad. Sci. USA*, 100(26):15324–15328, 2003.

D. McLeish. A general method for debiasing a Monte Carlo estimator. *Monte Carlo Methods Appl.*, 17(4):301–315, 2011.

S. Meyn and R. L. Tweedie. *Markov Chains and Stochastic Stability*. Cambridge University Press, 2nd edition, 2009. ISBN 978-0-521-73182-9.

D. Prangle. Lazier ABC. Preprint arXiv:1501.05144, 2015.

D. Prangle. Lazy ABC. *Statist. Comput.*, 26(1-2):171–185, 2016.

O. Ratmann, O. Jørgensen, T. Hinkley, M. Stumpf, S. Richardson, and C. Wiuf. Using likelihood-free inference to compare evolutionary dynamics of the protein networks of H. pylori and P. falciparum. *PLoS Comput. Biol.*, 3(11):e230, 2007.

C.-H. Rhee and P. W. Glynn. Unbiased estimation with square root convergence for SDE models. *Oper. Res.*, 63(5):1026–1043, 2015.

G. Roberts, A. Gelman, and W. Gilks. Weak convergence and optimal scaling of random walk Metropolis algorithms. *Ann. Appl. Probab.*, 7(1):110–120, 1997.

G. O. Roberts and J. S. Rosenthal. Harris recurrence of Metropolis-within-Gibbs and trans-dimensional Markov chains. *Ann. Appl. Probab.*, 16(4):2123–2139, 2006.
D. Rudolf and B. Sprungk. On a Metropolis-Hastings importance sampling estimator. Preprint arXiv:1805.07174, 2018.
I. Schuster and I. Klebanov. Markov chain importance sampling - a highly efficient estimator for MCMC. Preprint arXiv:1805.07179, 2018.
C. Sherlock, A. H. Thiery, G. O. Roberts, and J. S. Rosenthal. On the efficiency of pseudo-marginal random walk Metropolis algorithms. *Ann. Statist.*, 43(1):238–275, 2015.
S. Sisson and Y. Fan. ABC samplers. In S. Sisson, Y. Fan, and M. Beaumont, editors, *Handbook of Markov chain Monte Carlo*. Chapman & Hall/CRC Press, 2018.
A. D. Sokal. Monte Carlo methods in statistical mechanics: Foundations and new algorithms. Lecture notes, 1996.
M. Sunnåker, A. G. Busetto, E. Numminen, J. Corander, M. Foll, and C. Dessimoz. Approximate Bayesian computation. *PLoS computational biology*, 9(1):e1002803, 2013.
M. Tanaka, A. Francis, F. Luciani, and S. Sisson. Using approximate Bayesian computation to estimate tuberculosis transmission parameters from genotype data. *Genetics*, 173(3): 1511–1520, 2006.
M. N. Tran and R. Kohn. Exact ABC using importance sampling. Preprint arXiv:1509.08076, 2015.
M. Vihola. Robust adaptive Metropolis algorithm with coerced acceptance rate. *Statist. Comput.*, 22(5):997–1008, 2012.
M. Vihola, J. Helske, and J. Franks. Importance sampling type estimators based on approximate marginal MCMC. Preprint arXiv:1609.02541v5, 2016.
D. Wegmann, C. Leuenberger, and L. Excoffier. Efficient approximate Bayesian computation coupled with Markov chain Monte Carlo without likelihoods. *Genetics*, 182(4):1207–1218, 2009.

Appendix A. Proofs for the post-correction estimators

Proof of Theorem 5. Algorithm 1 is a Metropolis-Hastings algorithm with proposal $\tilde{q}(\theta, y; \theta', y') = q(\theta, \theta')g(y' \mid \theta')$ and with target

$$\tilde{\pi}_\epsilon(\theta, y) \propto \mathrm{pr}(\theta)g(y \mid \theta)\phi\big(d(y, y^*)/\epsilon\big).$$

The chain $(\Theta_k, Y_k)_{k\geq 1}$ is Harris-recurrent, as a full-dimensional Metropolis-Hastings which is ψ-irreducibile (Roberts and Rosenthal, 2006).

Because ϕ is monotone and $\epsilon \leq \epsilon_0$, we have $\phi\big(d(y, y^*)/\epsilon_0\big) \geq \phi\big(d(y, y^*)/\epsilon\big)$, and therefore $\tilde{\pi}_\epsilon$ is absolutely continuous with respect to $\tilde{\pi}_{\epsilon_0}$, so

$$w_{\epsilon_0,\epsilon}(y) = c_{\epsilon_0,\epsilon}\frac{\tilde{\pi}_\epsilon(\theta, y)}{\tilde{\pi}_{\epsilon_0}(\theta, y)},$$

where $c_{\epsilon_0,\epsilon} > 0$ is a constant. If we denote $\xi_k(f) := U_k^{(\epsilon_0,\epsilon)} f(\Theta_k)$ and $\xi_k(\mathbf{1}) := U_k^{(\epsilon_0,\epsilon)} = w_{\epsilon_0,\epsilon}(Y_k)$, then we may write

$$E^{(n)}_{\epsilon_0,\epsilon}(f) = \frac{\sum_{k=1}^n \xi_k(f)}{\sum_{j=1}^n \xi_j(\mathbf{1})} \xrightarrow{n\to\infty} \mathbb{E}_{\tilde{\pi}_\epsilon}[f(\Theta)],$$

by Harris recurrence and $\tilde{\pi}_\epsilon$ invariance (e.g. Vihola et al., 2016). The claim (i) follows because π_ϵ is the marginal density of $\tilde{\pi}_\epsilon$.

The chain $(\Theta_k, Y_k)_{k\geq 1}$ is reversible, so (ii) follows by (Vihola et al., 2016, Theorem 7(i)), because $m_f^{(2)}(\theta, y) := w_{\epsilon_0,\epsilon}^2(y) f^2(\theta)$ satisfies

$$\mathbb{E}_{\tilde{\pi}_{\epsilon_0}}[m_f^{(2)}(\Theta, Y)] = c_{\epsilon_0,\epsilon}\mathbb{E}_{\tilde{\pi}_\epsilon}[w_{\epsilon_0,\epsilon}(Y) f^2(\Theta)] \leq c_{\epsilon_0,\epsilon}\mathbb{E}_{\pi_\epsilon}[f^2(\Theta)] < \infty,$$

and because the asymptotic variance of the function $h_{\epsilon_0,\epsilon}$ with respect to $(\Theta_k, Y_k)_{k\geq 1}$ may be expressed as $\mathrm{var}_{\tilde{\pi}_{\epsilon_0}}\big(h_{\epsilon_0,\epsilon}(\Theta, Y)\big)\tau_{\epsilon_0,\epsilon}(f)$, so we may conclude that $v_{\epsilon_0,\epsilon}(f) = \mathrm{var}_{\tilde{\pi}_{\epsilon_0}}\big(h_{\epsilon_0,\epsilon}(\Theta, Y)\big)/c_{\epsilon_0,\epsilon}^2$.

The convergence $nS_{\epsilon_0,\epsilon}^{(n)}(f) \to v_{\epsilon_0,\epsilon}(f)$ follows from (Vihola et al., 2016, Theorem 9). □

Proof of Theorem 7. Note that $\tilde{\pi}_{\epsilon_0}(\theta, y) = \pi_{\epsilon_0}(\theta)\bar{g}_{\epsilon_0}(y \mid \theta)$, where

$$\bar{g}_{\epsilon_0}(y \mid \theta) := g(y \mid \theta)\mathbf{1}\left(d(y, y^*) \leq \epsilon_0\right)/L_{\epsilon_0}(\theta).$$

Notice that $\int \bar{g}_{\epsilon_0}(y \mid \theta) w_{\epsilon_0,\epsilon}^p(y)\mathrm{d}y = \bar{w}_{\epsilon_0,\epsilon}(\theta)$ for $p \in \{1, 2\}$, and consequently $\tilde{\pi}_{\epsilon_0}(h_{\epsilon_0,\epsilon}) = \pi_{\epsilon_0}(f_{\epsilon_0,\epsilon})$ and $\tilde{\pi}_{\epsilon_0}(h_{\epsilon_0,\epsilon}^2) = \pi_{\epsilon_0}(f^2\bar{w}_{\epsilon_0,\epsilon})$. Therefore,

$$\mathrm{var}_{\tilde{\pi}_{\epsilon_0}}(h_{\epsilon_0,\epsilon}) = \big[\mathrm{var}_{\pi_{\epsilon_0}}(f_{\epsilon_0,\epsilon}) + \pi_{\epsilon_0}\big(\bar{w}_{\epsilon_0,\epsilon}(1 - \bar{w}_{\epsilon_0,\epsilon})f^2\big)\big].$$

Hereafter, let $a_{\epsilon_0,\epsilon} := \big(\mathrm{var}_{\tilde{\pi}_{\epsilon_0}}(h_{\epsilon_0,\epsilon})\big)^{-1/2}$ and denote $\tilde{h}_{\epsilon_0,\epsilon} := a_{\epsilon_0,\epsilon}h_{\epsilon_0,\epsilon}$ and $\tilde{f}_{\epsilon_0,\epsilon} := a_{\epsilon_0,\epsilon}f_{\epsilon_0,\epsilon}$. Clearly, $\mathrm{var}_{\tilde{\pi}_{\epsilon_0}}(\tilde{h}_{\epsilon_0,\epsilon}) = 1$ and

$$\rho_k^{(\epsilon_0,\epsilon)} = e_k^{(\epsilon_0,\epsilon)} - \big(\pi_{\epsilon_0}(\tilde{f}_{\epsilon_0,\epsilon})\big)^2, \qquad \text{where} \qquad e_k^{(\epsilon_0,\epsilon)} := \mathbb{E}\big[\tilde{h}_{\epsilon_0,\epsilon}(\Theta_0^{(s)}, Y_0^{(s)})\tilde{h}_{\epsilon_0,\epsilon}(\Theta_k^{(s)}, Y_k^{(s)})\big].$$

Note that with $\phi = \phi_{\mathrm{simple}}$, the acceptance ratio satisfies

$$\alpha_{\epsilon_0}(\theta, y; \hat{\theta}, \hat{y}) = \dot{\alpha}(\theta, \hat{\theta})\mathbf{1}\left(d(\hat{y}, y^*) \leq \epsilon_0\right), \qquad \text{where} \qquad \dot{\alpha}(\theta, \hat{\theta}) = \min\left\{1, \frac{\mathrm{pr}(\hat{\theta})q(\hat{\theta}, \theta)}{\mathrm{pr}(\theta)q(\theta, \hat{\theta})}\right\},$$

which is independent of y, so $(\Theta_k^{(s)})$ is marginally a Metropolis-Hastings type chain, with proposal q and acceptance probability $\alpha(\theta, \hat{\theta})L_{\epsilon_0}(\hat{\theta})$. We have

$$\begin{aligned}
&\mathbb{E}\big[\tilde{h}_{\epsilon_0,\epsilon}(\Theta_1^{(s)}, Y_1^{(s)}) \mid (\Theta_0^{(s)}, Y_0^{(s)}) = (\theta, y)\big] \\
&\qquad = a_\epsilon \int q(\theta, \hat{\theta})\dot{\alpha}(\theta, \hat{\theta})g(\hat{y} \mid \hat{\theta})w_\epsilon(\hat{y})f(\hat{\theta})\mathrm{d}\hat{\theta}\mathrm{d}\hat{y} + r_{\epsilon_0}(\theta)\tilde{h}_{\epsilon_0,\epsilon}(\theta, y) \\
&\qquad = \int q(\theta, \hat{\theta})\dot{\alpha}(\theta, \hat{\theta})L_{\epsilon_0}(\hat{\theta})\tilde{f}_{\epsilon_0,\epsilon}(\hat{\theta})\mathrm{d}\hat{\theta} + r_{\epsilon_0}(\theta)\tilde{h}_\epsilon(\theta, y).
\end{aligned}$$

Using this iteratively, we obtain that

$$e_k^{(\epsilon_0,\epsilon)} = \mathbb{E}\big[\tilde{f}_{\epsilon_0,\epsilon}(\Theta_0^{(s)})\tilde{f}_{\epsilon_0,\epsilon}(\Theta_k^{(s)})\big] + \int \tilde{\pi}_{\epsilon_0}(\theta, y)\big[\tilde{h}_{\epsilon_0,\epsilon}^2(\theta, y) - \tilde{f}_{\epsilon_0,\epsilon}^2(\theta)\big]r_{\epsilon_0}^k(\theta)\mathrm{d}\theta\mathrm{d}y,$$

and therefore with $\gamma_k^{(\epsilon_0,\epsilon)} := a_{\epsilon_0,\epsilon}^2\mathrm{cov}\big(f_{\epsilon_0,\epsilon}(\Theta_0^{(s)}), f_{\epsilon_0,\epsilon}(\Theta_k^{(s)})\big)$,

$$\sum_{k\geq 1}\rho_k^{(\epsilon_0,\epsilon)} = \sum_{k\geq 1}\gamma_k^{(\epsilon_0,\epsilon)} + a_{\epsilon_0,\epsilon}^2 \int \pi_{\epsilon_0}(\theta)\bar{w}_{\epsilon_0,\epsilon}(\theta)\big(1 - \bar{w}_{\epsilon_0,\epsilon}(\theta)\big)\frac{r_{\epsilon_0}(\theta)}{1 - r_{\epsilon_0}(\theta)}f^2(\theta)\mathrm{d}\theta.$$

We conclude by noticing that $2\sum_{k\geq 1}\gamma_k^{(\epsilon_0,\epsilon)} = a_{\epsilon_0,\epsilon}^2\mathrm{var}_{\pi_{\epsilon_0}}(f_{\epsilon_0,\epsilon})(\check{\tau}_{\epsilon_0,\epsilon}(f) - 1)$. □

Appendix B. Convergence of the tolerance adaptive ABC-MCMC under generalised conditions

This appendix details a convergence theorem, under weaker assumptions than that of Theorem 11, for the tolerance adaptation (Algorithm 9) of Section 4.

Let us set $\beta := \log \epsilon$, and consider the proposal-rejection Markov kernel

$$\dot{P}_\beta(\theta, \mathrm{d}\vartheta) := q(\theta, \mathrm{d}\vartheta)\alpha_\beta(\theta, \vartheta) + \left(1 - \int q(\theta, \mathrm{d}\vartheta)\alpha_\beta(\theta, \vartheta)\right)\mathbf{1}\left(\theta \in \mathrm{d}\vartheta\right), \tag{3}$$

where $\alpha_\beta(\theta, \vartheta) := \dot{\alpha}(\theta, \vartheta)L_\beta(\vartheta)$,

$$\dot{\alpha}(\theta, \vartheta) := \min\left\{1, \frac{\mathrm{pr}(\vartheta)q(\vartheta, \theta)}{\mathrm{pr}(\theta)q(\theta, \vartheta)}\right\}, \qquad \text{and} \qquad L_\beta(\vartheta) := \int Q_\vartheta(\mathrm{d}t)\mathbf{1}\left(t \le e^\beta\right).$$

Then $\dot{P}_{\beta_k}$ is the transition of the θ-coordinate chain of Algorithm 9 with simple cut-off at iteration k, obtained by disregarding the t-coordinate. It is easily seen to be reversible with respect to the posterior probability $\pi_\beta(\theta) \propto \mathrm{pr}(\theta)L_\beta(\theta)$ given in (1), written here in terms of $\beta = \log \epsilon$ instead of ϵ.

Assumption 12. Suppose $\phi = \phi_{\text{simple}}$ and the following hold:

(i) Step sizes $(\gamma_k)_{k\ge1}$ satisfy $\gamma_k \ge 0$, $\gamma_{k+1} \le \gamma_k$,

$$\sum_{k\ge1} \gamma_k = \infty, \qquad \text{and} \qquad \sum_{k\ge1} \gamma_k^2\Big(1 + |\log \gamma_k| + |\log \gamma_k|^2\Big) < \infty.$$

(ii) The domain $\mathsf{T} \subset \mathbb{R}^{n_\theta}$, $n_\theta \ge 1$, is a nonempty open set.

(iii) $\mathrm{pr}(\,\cdot\,)$ and $q(\theta, \,\cdot\,)$ are uniformly bounded densities on $\mathbb{R}^{n_\theta}$ (i.e. $\exists C > 0$ s.t. $q(\theta, \vartheta) < C$ and $\mathrm{pr}(\theta) < C$ for all θ, $\vartheta \in \mathbb{R}^{n_\theta}$), and $\mathrm{pr}(\theta) = 0$ for $\theta \notin \mathsf{T}$.

(iv) $Q_\theta(\mathrm{d}t)$ admits a uniformly bounded density $Q_\theta(t)$.

(v) The values $\{\beta_k\}$ remain in some compact subset $\mathsf{B} \subset \mathbb{R}$ almost surely.

(vi) $c_\beta > 0$ for all $\beta \in \mathsf{B}$, where $c_\beta := \int \mathrm{pr}(\mathrm{d}\theta)L_\beta(\theta)$.

(vii) There exists $\dot{V} : \mathsf{T} \to [1, \infty)$ such that the Markov transitions $\dot{P}_\beta$ are simultaneously $\dot{V}$-geometrically ergodic: there exist $C > 0$ and $\rho \in (0, 1)$ s.t. for all $k \ge 1$ and $f : \mathsf{T} \to \mathbb{R}$ with $|f| \le \dot{V}$, it holds that

$$|\dot{P}_\beta^k f(\theta) - \pi_\beta(f)| \le C\dot{V}(\theta)\rho^k.$$

(viii) With $\mathbb{E}[\,\cdot\,] = \mathbb{E}_{\theta,\beta}[\,\cdot\,]$ denoting expectation with respect to the law of the marginal chain (Θ_k) of Algorithm 9 started at $\theta \in \mathsf{T}$, $\beta \in \mathsf{B}$, and with $\dot{V}$ as in Assumption 12(vii), we have,

$$\sup_{\theta,\beta,k} \mathbb{E}\big[\dot{V}(\Theta_k)^2\big] < \infty.$$

Theorem 13. *Under Assumption 12, the expected value of the acceptance probability (2), taken with respect to the stationary measure of the chain, converges to α^*.*

Proof of Theorem 13 can be found in Appendix C. It relies heavily on the simple conditions of (Andrieu et al., 2005, Theorem 2.3), which says that one must essentially show that the noise in the stochastic approximation update is asymptotically controlled.

We remark that there are likely extensions of Assumption 12(v) to the general non-compact adaptation parameter case based on projections (cf. Andrieu et al., 2005).

Appendix C. Analysis of the tolerance adaptive ABC-MCMC

In this appendix we aim to prove generalised convergence (Theorem 13 of Appendix B) of the tolerance adaptation, from which Theorem 11 of Section 4 will follow as a corollary.

In this appendix, $C > 0$ denotes some constant which may change from line to line.

C.1. **Proposal augmentation.** Suppose $\dot{L}$ is a Markov kernel which can be written as

$$\dot{L}(x, \mathrm{d}y) = q(x, \mathrm{d}y)\alpha(x, y) + \left(1 - \int q(x, \mathrm{d}y')\alpha(x, y')\right)\mathbf{1}\left(x \in \mathrm{d}y\right), \tag{4}$$

where $\alpha(x, y) \in [0, 1]$ is a jointly measurable function and $q(x, \mathrm{d}y)$ is a Markov proposal kernel. With $\breve{x} := (x, x')$, we define the *proposal augmentation* to be the Markov kernel

$$L(\breve{x}, \mathrm{d}\breve{y}) = \alpha(\breve{x})\mathbf{1}\left(x' \in \mathrm{d}y\right) q(x', \mathrm{d}y') + \big(1 - \alpha(\breve{x})\big)\mathbf{1}\left(x \in \mathrm{d}y\right) q(x, \mathrm{d}y'). \tag{5}$$

It is easy to see that L need not be reversible even if $\dot{L}$ is reversible. In this case, however, L does leave a probability measure invariant.

Lemma 14. *Suppose a Markov kernel $\dot{L}$ of the form given in (4) is $\dot{\mu}$-reversible. Let L be its proposal augmentation. Then the following statements hold:*

(i) $\mu L = \mu$, where $\mu(\mathrm{d}x, \mathrm{d}x') := \dot{\mu}(\mathrm{d}x)q(x, \mathrm{d}x')$.

(ii) If $\dot{L}$ is $\dot{V}$-geometrically ergodic with constants $(\dot{C}, \dot{\rho})$, then L is V-geometrically ergodic with constants (C, ρ), where $C := 2\dot{C}/\dot{\rho}$, $\rho := \dot{\rho}$, and $V(\breve{x}) := \frac{1}{2}\big(V(x) + V(x')\big)$.

Lemma 14 extends (Schuster and Klebanov, 2018, Theorem 4), who consider the case where $\dot{P}$ is a Metropolis-Hastings chain (see also Delmas and Jourdain, 2009; Rudolf and Sprungk, 2018). The extension to the more general class of reversible proposal-rejection chains allows one to consider, for example, jump and delayed acceptance chains, as well as the marginal chain (3) of Appendix B, which will be important for our analysis of the tolerance adaptation.

Proof of Lemma 14. Part (i) follows by a direct calculation. We now consider part (ii). For $f : \mathsf{X}^2 \to \mathbb{R}$, we shall use the notation $\dot{f}(x) := \int f(\breve{x})q(x, \mathrm{d}x')$. For $f : \mathsf{X}^2 \to \mathbb{R}$, we have

$$\int q(x, \mathrm{d}x')L\big((x, x'); \mathrm{d}\breve{y}\big)f(\breve{y}) = \int q(x, \mathrm{d}x')\alpha(\breve{x})\dot{f}(x') + \int q(x, \mathrm{d}x')\big(1 - \alpha(\breve{x})\big)\dot{f}(x) = \dot{L}\dot{f}(x),$$

and then inductively, for $k \geq 1$,

$$\begin{aligned}
\int q(x, \mathrm{d}x')L^k\big((x, x'); \mathrm{d}\breve{y}\big)f(\breve{y}) &= \int q(x, \mathrm{d}x')\alpha(\breve{x})q(x', \mathrm{d}y')L^{k-1}\big((x', y'); \mathrm{d}\breve{z}\big)f(\breve{z}) \\
&\quad + \int q(x, \mathrm{d}x')\big(1 - \alpha(\breve{x})\big)q(x, \mathrm{d}y')L^{k-1}\big((x, y'); \mathrm{d}\breve{z}\big)f(\breve{z}) \\
&= \int q(x, \mathrm{d}x')\alpha(\breve{x})\dot{L}^{k-1}\dot{f}(x') + \int q(x, \mathrm{d}x')\big(1 - \alpha(\breve{x})\big)\dot{L}^{k-1}\dot{f}(x) \\
&= \dot{L}^k\dot{f}(x).
\end{aligned}$$

We then have the equality,

$$\begin{aligned}
L^k f(\breve{x}) &= \alpha(\breve{x}) \int q(x', \mathrm{d}y')L^{k-1}\big((x', y'); \mathrm{d}\breve{z}\big)f(\breve{z}) + \big(1 - \alpha(\breve{x})\big) \int q(x, \mathrm{d}y')L^{k-1}\big((x, y'); \mathrm{d}\breve{z}\big)f(\breve{z}) \\
&= \alpha(\breve{x})\dot{L}^{k-1}\dot{f}(x') + \big(1 - \alpha(\breve{x})\big)\dot{L}^{k-1}\dot{f}(x).
\end{aligned}$$

For $\|f\| \le V$, note that $\|\dot{f}\| \le \dot{V}$ since $\|q\|_\infty \le 1$, and we conclude (ii) from

$$\begin{aligned}|L^k f(\breve{x}) - \mu(f)| &\le \alpha(\breve{x})|\dot{L}^{k-1}\dot{f}(x') - \dot{\mu}(\dot{f})| + \big(1-\alpha(\breve{x})\big)|\dot{L}^{k-1}\dot{f}(x) - \dot{\mu}(\dot{f})| \\ &\le \dot{C}\dot{\rho}^{k-1}\big(\dot{V}(x') + \dot{V}(x)\big).\end{aligned}$$ □

Consider now the θ-coordinate chain $\dot{P}_\beta$ presented in (3) of Appendix B. This transition $\dot{P}_\beta$ is clearly a reversible proposal-rejection chain of the form (4). We now consider P_β, its proposal augmentation. This is the chain $\breve{\Theta}_k := (\Theta_k, \Theta'_k) \in \mathsf{T}^2$, formed by disregarding the t-parameter as with $\dot{P}_\beta$ before, but now augmenting by the proposal $\theta' \sim q(\theta, \cdot)$. Its transitions are of the form $\breve{\theta} = \breve{\Theta}_k$ goes to $\breve{\vartheta} = \breve{\Theta}_{k+1}$ in the ABC-MCMC, with $\breve{\vartheta} = (\vartheta, \vartheta')$ and kernel

$$P_\beta(\breve{\theta}, \mathrm{d}\breve{\vartheta}) := \alpha_\beta(\breve{\theta})\mathbf{1}\left(\theta' \in \mathrm{d}\vartheta\right) q(\theta', \mathrm{d}\vartheta') + \big(1-\alpha_\beta(\breve{\theta})\big)\mathbf{1}\left(\theta \in \mathrm{d}\vartheta\right) q(\theta, \mathrm{d}\vartheta')$$

By Lemma 14(i), P_β leaves $\pi'_\beta := \pi'_{\beta,u}/c_\beta$ invariant, where $\pi'_{\beta,u}(\mathrm{d}\breve{\theta}) := \mathrm{pr}(\mathrm{d}\theta)L_\beta(\theta)q(\theta, \mathrm{d}\theta')$ and $c_\beta := \int \mathrm{pr}(\mathrm{d}\theta)L_\beta(\theta)$.

C.2. **Monotonicity properties.** The following result establishes monotonicity of the mean field acceptance rate with increasing tolerance.

Lemma 15. *Assume Assumption 12(iii) and 12(iv) hold. The mapping $\beta \mapsto \pi'_\beta(\alpha_\beta)$ is monotone non-decreasing.*

Proof. Since $\mathrm{pr}(\theta)$ and $q(\theta, \theta')$ are uniformly bounded (Assumption 12(iii)), and $L_\beta(\theta) \le 1$, differentiation under the integral sign is possible in the following by the dominated convergence theorem. By the quotient rule,

$$\frac{\mathrm{d}}{\mathrm{d}\beta}\Big(\pi'_\beta(\alpha_\beta)\Big) = \frac{1}{c_\beta^2}\left(c_\beta \frac{\mathrm{d}}{\mathrm{d}\beta}\Big(\pi'_{\beta,u}(\alpha_\beta)\Big) - \pi'_{\beta,u}(\alpha_\beta)\frac{\mathrm{d}c_\beta}{\mathrm{d}\beta}\right). \tag{6}$$

By reversibility of Metropolis-Hastings targeting $\mathrm{pr}(\theta)$ with proposal q,

$$\frac{\mathrm{d}}{\mathrm{d}\beta}\Big(\pi'_{\beta,u}(\alpha_\beta)\Big) = 2e^\beta \int \mathrm{pr}(\mathrm{d}\theta)L_\beta(\theta)q(\theta, \mathrm{d}\theta')\dot{\alpha}(\theta, \theta')Q_{\theta'}(e^\beta).$$

With

$$f(\theta') := 2Q_{\theta'}(e^\beta)\int \mathrm{pr}(\mathrm{d}\tilde{\theta})L_\beta(\tilde{\theta}) - L_\beta(\theta')\int \mathrm{pr}(\mathrm{d}\tilde{\theta})Q_{\tilde{\theta}}(e^\beta),$$

we can then write (6) as

$$\frac{\mathrm{d}}{\mathrm{d}\beta}\Big(\pi'_\beta(\alpha_\beta)\Big) = \frac{e^\beta}{c_\beta^2}\int \mathrm{pr}(\mathrm{d}\theta)L_\beta(\theta)q(\theta, \mathrm{d}\theta')\dot{\alpha}(\theta, \theta')f(\theta').$$

By the same reversibility property as before, we can write this again as

$$\frac{\mathrm{d}}{\mathrm{d}\beta}\Big(\pi'_\beta(\alpha_\beta)\Big) = \frac{e^\beta}{c_\beta^2}\int f(\theta)\mathrm{pr}(\mathrm{d}\theta)\int q(\theta, \mathrm{d}\theta')L_\beta(\theta')\dot{\alpha}(\theta, \theta'),$$

We then conclude, since

$$\int f(\theta)\mathrm{pr}(\mathrm{d}\theta) = \int Q_\theta(e^\beta)\mathrm{pr}(\mathrm{d}\theta)\int L_\beta(\tilde{\theta})\mathrm{pr}(\mathrm{d}\tilde{\theta}) \ge 0.$$ □

Lemma 16. *The following statements hold:*

(i) The function $\beta \mapsto c_\beta$ is monotone non-decreasing on $\mathbb{R}$.

(ii) If Assumption 12(v) and 12(vi) hold, then there exist $C_{\min} > 0$, $C_{\max} > 0$ such that $C_{\min} \le c_\beta \le C_{\max}$ for all $\beta \in \mathsf{B}$.

Proof. Part (i) follows, for $\beta \le \beta'$, from

$$c_\beta = \int \mathrm{pr}(\mathrm{d}\theta) Q_\theta([0, e^\beta]) \le \int \mathrm{pr}(\mathrm{d}\theta) Q_\theta([0, e^{\beta'}]) = c_{\beta'}.$$

Consider part (ii). By part (i) and compactness of B (Assumption 12(v)), we can set $C_{\min} := c_{\min(\mathsf{B})}$ and $C_{\max} := c_{\max(\mathsf{B})}$, both of which are positive by Assumption 12(vi). □

C.3. **Stochastic approximation framework.** To obtain a form common in the stochastic approximation literature (cf. Andrieu et al., 2005), we write the update in Algorithm 9 as

$$\begin{aligned}\beta_{k+1} &= \beta_k + \gamma_{k+1} H_{\beta_k}(\breve{\Theta}_k, T'_k) \\ &= \beta_k + \gamma_{k+1} h(\beta_k) + \gamma_{k+1} \zeta_{k+1}\end{aligned}$$

where $H_\beta(\breve{\theta}, t') := \alpha^* - \alpha'_\beta(\breve{\theta}, t')$,

$$\alpha'_\beta(\breve{\theta}, t') := \min\left\{1, \frac{\mathrm{pr}(\theta') q(\theta', \theta)}{\mathrm{pr}(\theta) q(\theta, \theta')}\right\} \mathbf{1}\left(t' \le e^\beta\right),$$

$$h(\beta) := \pi'_\beta(\widehat{H}_\beta) = \int \pi_\beta(\mathrm{d}\theta) q(\theta, \mathrm{d}\theta') Q_{\theta'}(\mathrm{d}t') H_\beta(\theta, \theta', t'),$$

noise sequence $\zeta_{k+1} := H_{\beta_k}(\breve{\Theta}_k, T'_k) - h(\beta_k)$, and conditional expectation

$$\widehat{H}_\beta(\breve{\theta}) := \mathbb{E}[H_\beta(\breve{\Theta}, T') | \breve{\Theta} = \breve{\theta}],$$

where $T' \sim Q_{\theta'}(\,\cdot\,)$. We also set for convenience $\bar{H}_\beta(\breve{\theta}) := \widehat{H}_\beta(\breve{\theta}) - \pi'_\beta(\widehat{H}_\beta)$.

Lemma 17. *Suppose Assumption 12(vii) holds. Then the following statements hold:*

(i) The proposal augmented kernels $(P_\beta)_{\beta \in \mathsf{B}}$ are simultaneously V-geometrically ergodic, where $V(\theta, \theta') := \frac{1}{2}\big(\dot{V}(\theta) + \dot{V}(\theta')\big)$, with $\dot{V}$ as in Assumption 12(vii).

(ii) There exists $C > 0$, such that for all $\beta \in \mathsf{B}$, the formal solution $g_\beta = \sum_{k \ge 0} P_\beta^k \bar{H}_\beta$ to the Poisson equation $g_\beta - P_\beta g_\beta = \bar{H}_\beta$ satisfies $|g_\beta(\breve{\theta})| \le C V(\breve{\theta})$.

Proof. (i) follows directly from the explicit parametrisation for (C, ρ) given in Lemma 14(ii).

Part (ii) follows from part (i) and the bound, since $|\bar{H}_\beta| \le 1 \le V$,

$$|g_\beta(\breve{\theta})| \le 1 + C_\beta \sum_{k \ge 1} \rho_\beta^k V(\breve{\theta}) \le \left(1 + \frac{C_\beta}{1 - \rho_\beta}\right) V(\breve{\theta}). \quad \square$$

C.4. **Contractions.** We define for $V : \mathsf{T} \to [1, \infty)$ and $g : \mathsf{T} \to \mathbb{R}$ the V-norm $\|g\|_V := \sup_{\theta \in \mathsf{T}} \frac{|g(\theta)|}{V(\theta)}$. We define for a bounded operator A on a Banach space of bounded functions f, the operator norm $\|A\|_\infty = \sup_f \frac{\|Af\|_\infty}{\|f\|_\infty}$.

Lemma 18. *Suppose Assumption 12(iv), 12(v) and 12(vi) hold. The following hold:*

(i) $\exists C > 0$, $\exists C^+_{\mathsf{B}} > 0$ s.t. $\forall \beta_1 \in \mathsf{B}$, $\forall \beta_2 \in \mathsf{B}$, $\forall g : \mathsf{T}^2 \to \mathbb{R}$ bounded, we have

$$\|(P_{\beta_1} - P_{\beta_2}) g\|_\infty \le C \|g\|_\infty |e^{\beta_1} - e^{\beta_2}| \le C^+_{\mathsf{B}} \|g\|_\infty |\beta_1 - \beta_2|.$$

(ii) $\exists C^-_{\mathsf{B}} > 0$, $\exists C_{\mathsf{B}} > 0$, s.t. $\forall \beta_1 \in \mathsf{B}$, $\forall \beta_2 \in \mathsf{B}$, we have

$$\|\bar{H}_{\beta_1} - \bar{H}_{\beta_2}\|_\infty \le C^-_{\mathsf{B}} |e^{\beta_1} - e^{\beta_2}| \le C_{\mathsf{B}} |\beta_1 - \beta_2|.$$

(iii) $\exists C^-_{\mathsf{B}} > 0$, $\exists C_{\mathsf{B}} > 0$, s.t. $\forall \beta_1 \in \mathsf{B}$, $\forall \beta_2 \in \mathsf{B}$, $\forall g : \mathsf{T}^2 \to \mathbb{R}$ bounded, we have

$$|\pi'_{\beta_1}(g) - \pi'_{\beta_2}(g)| \le C^-_{\mathsf{B}} \|g\|_\infty |e^{\beta_1} - e^{\beta_2}| \le C_{\mathsf{B}} \|g\|_\infty |\beta_1 - \beta_2|.$$

Proof. By Assumption 12(iv), we have for all β_1, $\beta_2 \in \mathsf{B}$,

$$|L_{\beta_1}(\theta) - L_{\beta_2}(\theta)| = \int_{e^{\beta_1 \wedge \beta_2}}^{e^{\beta_1 \vee \beta_2}} Q_\theta(\mathrm{d}t) \le C|e^{\beta_1} - e^{\beta_2}|.$$

We obtain the first inequality for part (i), then, from the bound,

$$\begin{aligned}|(P_{\beta_1} - P_{\beta_2})g(\breve{\theta})| &= |\big(\alpha_{\beta_1}(\breve{\theta}) - \alpha_{\beta_2}(\breve{\theta})\big)\dot{g}(\theta') + \big(\alpha_{\beta_2}(\breve{\theta}) - \alpha_{\beta_1}(\breve{\theta})\big)\dot{g}(\theta)| \\ &\le \dot{\alpha}(\breve{\theta})|L_{\beta_1}(\theta') - L_{\beta_2}(\theta')| \int \Big(q(\theta', \mathrm{d}\vartheta')|g(\theta', \vartheta')| + q(\theta, \mathrm{d}\vartheta')|g(\theta, \vartheta')|\Big),\end{aligned}$$

The second, Lipschitz bound follows by a mean value theorem argument for the function $\beta \mapsto e^\beta$, namely

$$|e^{\beta_1} - e^{\beta_2}| \le \sup_{\beta \in \mathsf{B}} e^\beta \, |\beta_1 - \beta_2| \le C_{\mathsf{B}}^+ |\beta_1 - \beta_2|,$$

where the last inequality follows by compactness of B (Assumption 12(v)).

We now consider part (ii). We have,

$$\|\bar{H}_{\beta_1} - \bar{H}_{\beta_2}\|_\infty \le \|\widehat{H}_{\beta_1} - \widehat{H}_{\beta_2}\|_\infty + |h(\beta_1) - h(\beta_2)|.$$

For the first term, by Assumption 12(iv), as in (i), we have

$$\|\widehat{H}_{\beta_1} - \widehat{H}_{\beta_2}\|_\infty \le \sup_{\breve{\theta}} \dot{\alpha}(\breve{\theta}) \int_{e^{\beta_1 \wedge \beta_2}}^{e^{\beta_1 \vee \beta_2}} Q_{\theta'}(\mathrm{d}t) \le C|\beta_1 - \beta_2|.$$

For the other term, we have

$$|h(\beta_1) - h(\beta_2)| \le \frac{1}{c_{\beta_1}} |\pi'_{\beta_1,u}(\alpha_{\beta_1}) - \pi'_{\beta_2,u}(\alpha_{\beta_2})| + \pi'_{\beta_2,u}(\alpha_{\beta_2}) \frac{|c_{\beta_1} - c_{\beta_2}|}{c_{\beta_1} c_{\beta_2}}.$$

By the triangle inequality, we have

$$|\pi'_{\beta_1,u}(\alpha_{\beta_1}) - \pi'_{\beta_2,u}(\alpha_{\beta_2})| \le |\pi'_{\beta_1,u}(\alpha_{\beta_1}) - \pi'_{\beta_1,u}(\alpha_{\beta_2})| + |\pi'_{\beta_1,u}(\alpha_{\beta_2}) - \pi'_{\beta_2,u}(\alpha_{\beta_2})|$$

Each term above is bounded by $C|e^{\beta_1} - e^{\beta_2}|$, as is $|c_{\beta_1} - c_{\beta_2}|$. Moreover, by Lemma 16(ii), we have $c_\beta \ge c_{\min} > 0$ for all $\beta \in \mathsf{B}$, and the first inequality in part (ii) follows. The second inequality follows by a mean value theorem argument as before. Proof of (iii) is simpler. □

C.5. **Control of noise.** We state a simple standard fact used repeatedly in the proof of Lemma 20 below, our key lemma.

Lemma 19. *Suppose $(X_j)_{j \ge 1}$ are random variables with $X_j \ge 0$, $X_{j+1} \le X_j$, and $\lim_{j\to\infty} \mathbb{E}[X_j] = 0$. Then, almost surely, $\lim_{j\to\infty} X_j = 0$.*

Lemma 20. *Suppose Assumption 12 holds. Then, with $\mathcal{T}_{j,n} := \sum_{k=j}^n \gamma_k \zeta_k$, we have*

$$\lim_{j\to\infty} \sup_{n \ge j} \big|\mathcal{T}_{j,n}\big| = 0, \qquad \textit{almost surely.}$$

Proof. Similar to (Andrieu et al., 2005, Proof of Prop. 5.2), we write $\mathcal{T}_{j,n} := \sum_{i=1}^8 \mathcal{T}_{j,n}^{(j)}$, where

$$\widehat{H}_{\beta_{k-1}}(\breve{\Theta}_{k-1}) = \mathbb{E}[H_{\beta_{k-1}}(\breve{\Theta}_{k-1}, T')|\mathcal{F}'_{k-1}],$$

with $\mathcal{F}'_{k-1} = \sigma(\beta_{k-1}, \Theta_{k-1}, \Theta'_{k-1})$ representing the information obtained through running Algorithm 9 up to and including iteration $k-2$ and then also generating Θ'_{k-1}, and

$$\mathcal{T}_{j,n}^{(1)} := \sum_{k=j}^n \gamma_k \Big(H_{\beta_{k-1}}(\breve{\Theta}_{k-1}, T'_{k-1}) - \widehat{H}_{\beta_{k-1}}(\breve{\Theta}_{k-1})\Big),$$

$$\mathcal{T}^{(2)}_{j,n} := \sum_{k=j}^{n} \gamma_k \Big(g_{\beta_{k-1}}(\breve{\Theta}_{k-1}) - P_{\beta_{k-1}} g_{\beta_{k-1}}(\breve{\Theta}_{k-2}) \Big),$$

$$\mathcal{T}^{(3)}_{j,n} := \gamma_{j-1} P_{j-1} g_{\beta_{j-1}}(\breve{\Theta}_{j-2}) - \gamma_n P_{\beta_n} g_{\beta_n}(\breve{\Theta}_{n-1}),$$

$$\mathcal{T}^{(4)}_{j,n} := \sum_{k=j}^{n} \Big(\gamma_k - \gamma_{k-1} \Big) P_{\beta_{k-1}} g_{\beta_{k-1}}(\breve{\Theta}_{k-2}),$$

$$\mathcal{T}^{(5)}_{j,n} := \sum_{k=j}^{n} \gamma_k \sum_{i \geq m_k + 1} P^i_{\beta_k} \bar{H}_{\beta_k}(\breve{\Theta}_{k-1}),$$

$$\mathcal{T}^{(6)}_{j,n} := -\sum_{k=j}^{n} \gamma_k \sum_{i \geq m_k + 1} P^i_{\beta_{k-1}} \bar{H}_{\beta_{k-1}}(\breve{\Theta}_{k-1}),$$

$$\mathcal{T}^{(7)}_{j,n} := \sum_{k=j}^{n} \gamma_k \sum_{i=1}^{m_k} \Big(P^i_{\beta_k} - P^i_{\beta_{k-1}} \Big) \bar{H}_{\beta_k}(\breve{\Theta}_{k-1}),$$

$$\mathcal{T}^{(8)}_{j,n} := \sum_{k=j}^{n} \gamma_k \sum_{i=1}^{m_k} P^i_{\beta_{k-1}} \Big(\bar{H}_{\beta_k} - \bar{H}_{\beta_{k-1}} \Big)(\breve{\Theta}_{k-1}).$$

Here, g_β is the Poisson solution defined in Lemma 17(ii), and $m_k := \lceil |\log \gamma_k| \rceil$. We remind that $\bar{H}_\beta := \widehat{H}_\beta - h(\beta)$ from Section C.3.

We now show $\lim_{j\to\infty} \sup_{n \geq j} \left| \mathcal{T}^{(i)}_{j,n} \right| = 0$ for each of the terms $i \in \{1{:}8\}$ individually, which implies the result of the lemma.

(1) Since for all $n > j$,

$$\mathbb{E}[\mathcal{T}^{(1)}_{j,n} - \mathcal{T}^{(1)}_{j,n-1} | \mathcal{F}'_{n-1}] = 0,$$

we have that $(\mathcal{T}^{(1)}_{j,n})_{n \geq j}$ is a $\mathcal{F}'_n$-martingale for each $j \geq 1$. By the Burkholder-Davis-Gundy inequality for martingales (cf. Burkholder et al., 1972), we have

$$\mathbb{E}[\sup_{n \geq j} |\mathcal{T}^{(1)}_{j,n}|^2] \leq C \mathbb{E}\Big[\sum_{k=j}^{\infty} \gamma_k^2 \big(H_{\beta_{k-1}}(\breve{\Theta}_{k-1}, T'_{k-1}) - \widehat{H}_{\beta_{k-1}}(\breve{\Theta}_{k-1}) \big)^2 \Big] \leq C \sum_{k=j}^{\infty} \gamma_k^2,$$

where in the last inequality we have noted that $|H_\beta - \widehat{H}_\beta| \leq 1$. Since $\sum_{k \geq 1} \gamma_k^2 < \infty$, we get that

$$\lim_{j \to \infty} \mathbb{E}[\sup_{n \geq j} |\mathcal{T}^{(1)}_{j,n}|^2] = 0.$$

Hence, the result follows by Lemma 19.

(2) For $j \geq 2$, we have for $n > j$,

$$\mathbb{E}[\mathcal{T}^{(2)}_{j,n} - \mathcal{T}^{(2)}_{j,n-1} | \mathcal{F}'_{n-2}] = 0,$$

so that $(\mathcal{T}^{(2)}_{j,n})_{n \geq j}$ is a $\mathcal{F}'_{n-1}$-martingale, for $j \geq 2$. By the Burkholder-Davis-Gundy inequality again,

$$\mathbb{E}[\sup_{n \geq j} |\mathcal{T}^{(2)}_{j,n}|^2] \leq C \mathbb{E}\Big[\sum_{k=j}^{\infty} \gamma_k^2 \big(g_{\beta_{k-1}}(\breve{\Theta}_{k-1}) - P_{\beta_{k-1}} g_{\beta_{k-1}}(\breve{\Theta}_{k-2}) \big)^2 \Big].$$

We then use Lemma 17(ii) and $\|P_\beta\|_\infty \leq 1$, to get, after combining terms,

$$\mathbb{E}[\sup_{n \geq j} |\mathcal{T}^{(2)}_{j,n}|^2] \leq C \sum_{k=j-1}^{\infty} \gamma_k^2 \mathbb{E}\Big[V(\breve{\Theta}_{k-1})^2 \Big] \leq C \sum_{k=j-1}^{\infty} \gamma_k^2,$$

where we have used Assumption 12(viii) in the last inequality. We then conclude by Lemma 19 as before.
(3) By Lemma 17(ii), the triangle inequality, $\|P_\beta\|_\infty \le 1$, and the dominated convergence theorem, we obtain

$$\mathbb{E}[\sup_{n\ge j}|\mathcal{T}^{(3)}_{j,n}| \le C\gamma_{j-1}\mathbb{E}[V(\breve{\Theta}_{j-2})] + C\sup_{n\ge j}\gamma_n\mathbb{E}[V(\breve{\Theta}_{n-1})].$$

We then apply Assumption 12(viii) and Jensen's inequality, and use that γ_k go to zero, since $\sum\gamma_k^2 < \infty$, to get that

$$\lim_{j\to\infty}\mathbb{E}[\sup_{n\ge j}|\mathcal{T}^{(3)}_{j,n}|] \le C\Big(\lim_{j\to\infty}\gamma_{j-1} + \sup_{n\ge j}\gamma_n\Big) = 0.$$

We now may conclude by Lemma 19.
(4) By Lemma 17(ii) and $\gamma_k \le \gamma_{k-1}$, we have for $j \ge 2$,

$$\mathbb{E}[\sup_{n\ge j}|\mathcal{T}^{(4)}_{j,n}|] \le C\sup_{n\ge j}\sum_{k=j}^{n}(\gamma_{k-1}-\gamma_k)\mathbb{E}[V(\breve{\Theta}_{k-2})] \le C\sup_{n\ge j}\sum_{k=j}^{n}(\gamma_{k-1}-\gamma_k)$$

where we have used lastly Assumption 12(viii) and Jensen's inequality. Since this is a telescoping sum, we get

$$\mathbb{E}[\sup_{n\ge j}|\mathcal{T}^{(4)}_{j,n}|] \le C\sup_{n\ge j}(\gamma_{j-1}-\gamma_n) \le C\gamma_{j-1}$$

We then conclude by Lemma 19, since $\gamma_j \to 0$.
(5) By Lemma 17(i), $|P^i_\beta \bar{H}_\beta(\breve{\theta})| \le C\rho^i V(\breve{\theta})$, where C, ρ do not depend on $\beta \in \mathsf{B}$. Hence,

$$\mathbb{E}[|\mathcal{T}^{(5)}_{j,n}|] \le C\sum_{k=j}^{n}\gamma_k\sum_{i\ge m_k+1}\rho^i\mathbb{E}[V(\breve{\Theta}_{k-1})] \le C\sum_{k=j}^{n}\gamma_k\rho^{m_k},$$

where we have used lastly Assumption 12(viii) and Jensen's inequality. Since m_k was defined to be of order $|\log\gamma_k|$, we have

$$\mathbb{E}[|\mathcal{T}^{(5)}_{j,n}|] \le C\sum_{k=j}^{\infty}\gamma_k^2 < \infty$$

By the dominated convergence theorem, we then have

$$\mathbb{E}[\sup_{n\ge j}|\mathcal{T}^{(5)}_{j,n}|] \le C\sum_{k=j}^{\infty}\gamma_k^2.$$

Taking the limit $j \to \infty$, we can then conclude by using Lemma 19.
(6) The proof is essentially the same as for (5).
(7) We write for $i \ge 1$,

$$P^i_{\beta_k} - P^i_{\beta_{k-1}} = \sum_{l=0}^{i-1}P^{i-l-1}_{\beta_k}\big(P_{\beta_k} - P_{\beta_{k-1}}\big)P^l_{\beta_{k-1}}.$$

Since $\|P^i_\beta\|_\infty \le 1$ for all $i \ge 0$, and $|\bar{H}_\beta| \le 1$, by Lemma 18(i), we have

$$\|(P^i_{\beta_k} - P^i_{\beta_{k-1}})\bar{H}_{\beta_k}\| \le C\sum_{l=0}^{i-1}\|P^{i-l-1}_{\beta_k}\|_\infty|\beta_k-\beta_{k-1}|\|P^l_{\beta_{k-1}}\bar{H}_{\beta_k}\|_\infty \le C|\beta_k-\beta_{k-1}|i.$$

Since $|\beta_k - \beta_{k-1}| \le \gamma_k$ from the adaptation step in Algorithm 9, we have

$$|\mathcal{T}_{j,n}^{(7)}| \le C \sum_{k=j}^{n} \gamma_k \sum_{i=1}^{m_k} i\gamma_k \le C \sum_{k=j}^{\infty} \gamma_k^2 m_k(1+m_k) < \infty.$$

We then take $\sup_{n \ge j}$ on the left, take the expectation, and conclude by Lemma 19.
(8) Since $\|P_\beta^i\|_\infty \le 1$ and by Lemma 18(ii), we have that

$$\|P_{\beta_{k-1}}^i(\bar{H}_{\beta_k} - \bar{H}_{\beta_{k-1}})\|_\infty \le \|P_{\beta_{k-1}}^i\|_\infty \|\bar{H}_{\beta_k} - \bar{H}_{\beta_{k-1}}\|_\infty \le C|\beta_k - \beta_{k-1}|$$

Since $|\beta_k - \beta_{k-1}| \le \gamma_k$, we have

$$\mathbb{E}[\sup_{n \ge j} \mathcal{T}_{j,n}^{(8)}] \le C \sum_{k=j}^{\infty} \gamma_k^2 m_k < \infty.$$

We then conclude by Lemma 19. □

C.6. **Proofs of convergence theorems.**

Proof of Theorem 13. We define our Lyapunov function $w : \mathbb{R} \to [0, \infty)$ to be the continuously differentiable function $w(\beta) := \frac{1}{2}|e^\beta - e^{\beta^*}|^2$. We also have that $h(\beta) := \pi'_\beta(\widehat{H}_\beta)$ is continuous, which follows from Lemma 18(iii). One can then check that Assumption 12 and Lemma 20 imply that the assumptions of (Andrieu et al., 2005, Theorem 2.3) hold. The latter result implies $\lim|\beta_k - \beta^*| \to 0$, for some $\beta^* \in \mathsf{B}$ satisfying $\pi'_{\beta^*}(\alpha_{\beta^*}) = \alpha^*$, as desired. □

Lemma 21. *Suppose Assumption 10 holds. Then both $(\dot{P}_\beta)_{\beta \in \mathsf{B}}$ and $(P_\beta)_{\beta \in \mathsf{B}}$ are simultaneously 1-geometrically ergodic (i.e. uniformly ergodic).*

Proof. We have $\mathrm{pr}(\theta) \le C_{\mathrm{pr}}$ some $C_{\mathrm{pr}} > 0$, and also $0 < \delta_q \le q(\theta, \vartheta)$, for all $\theta, \vartheta \in \mathsf{T}$. Hence, for $A \subset \mathsf{T}$,

$$\dot{P}_\beta(\theta, A) \ge \int \delta_q \min\left\{1, \frac{\mathrm{pr}(\vartheta)}{\mathrm{pr}(\theta)}\right\} L_\beta(\vartheta) \mathbf{1}(\vartheta \in A) \ge \int \delta_q \frac{\mathrm{pr}(\vartheta)}{C_{\mathrm{pr}}} L_\beta(\vartheta) \mathbf{1}(\vartheta \in A)$$

By Lemma 16(ii), it holds $c_\beta \ge C_{\min}$ for some $C_{\min} > 0$ for all $\beta \in \mathsf{B}$. Therefore,

$$\dot{P}_\beta(\theta, A) \ge \delta \pi_\beta(A),$$

where $\delta := \delta_q C_{\min}/C_{\mathrm{pr}} > 0$ is independent of β. As in Nummelin's split chain construction (cf. Meyn and Tweedie, 2009), we can then define the Markov kernel $R_\beta(\theta, A) := (1-\delta)^{-1}\big(\dot{P}_\beta(\theta, A) - \delta\pi_\beta(A)\big)$ with $\pi_\beta R_\beta = \pi_\beta$. Set $\Pi_\beta(\theta, A) := \pi_\beta(A)$. For any $f \le 1$, $\beta \in \mathsf{B}$, and $k \ge 1$, we have

$$\begin{aligned}
\|\dot{P}_\beta^k f - \pi_\beta(f)\|_\infty &= (1-\delta)\|(R_\beta - \Pi_\beta)\dot{P}_\beta^{k-1} f\|_\infty = (1-\delta)\|R_\beta \dot{P}_\beta^{k-1}\big(f - \pi_\beta(f)\big)\|_\infty \\
&\le (1-\delta)\|\dot{P}_\beta^{k-1}\big(f - \pi_\beta(f)\big)\|_\infty = (1-\delta)\|\dot{P}_\beta^{k-1} f - \pi_\beta(f)\|_\infty \\
&\le \ldots \le (1-\delta)^k \|f - \pi_\beta(f)\|_\infty \le 2(1-\delta)^k \|f\|_\infty,
\end{aligned}$$

where we have used $\|R_\beta\|_\infty \le 1$ in the first inequality. Hence, $(\dot{P}_\beta)_{\beta \in \mathsf{B}}$ are simultaneously 1-geometrically ergodic, and thus so are $(P_\beta)_{\beta \in \mathsf{B}}$ by Lemma 17(i). □

Proof of Theorem 11. Since $(\dot{P}_\beta)_{\beta \in \mathsf{B}}$ are simultaneously 1-geometric ergodic by Lemma 21, it is direct to see that Assumption 10 implies Assumption 12. We conclude by Theorem 13. □

Appendix D. Simultaneous tolerance and covariance adaptation

Algorithm 22 (TA-AM(n_b, α^*))**.** Suppose $\Theta_0 \in \mathsf{T} \subset \mathbb{R}^{n_\theta}$ is a starting value with $\mathrm{pr}(\Theta_0) > 0$ and $\Gamma_0 = \mathbf{1}_{n_\theta \times n_\theta}$ is the identity matrix.

1. Initialise $\epsilon_0 := T_0$ where $T_0 \sim Q_{\Theta_0}(\,\cdot\,)$ and $T_0 > 0$. Set $\mu_0 := \Theta_0$.
2. For $k = 0, \ldots, n_b - 1$, iterate:
   (i) Draw $\Theta'_k \sim N(\Theta_k, (2.38^2/n_\theta)\Gamma_k)$
   (ii) Draw $T'_k \sim Q_{\Theta'_k}(\,\cdot\,)$.
   (iii) Accept, by setting $(\Theta_{k+1}, T_{k+1}) \leftarrow (\Theta'_k, T'_k)$, with probability

$$\alpha_{\epsilon_k}(\Theta_k, T_k; \Theta'_k, T'_k) := \min\left\{1, \frac{\mathrm{pr}(\Theta'_k)\phi(T'_k/\epsilon_k)}{\mathrm{pr}(\Theta_k)\phi(T_k/\epsilon_k)}\right\}.$$

   Otherwise reject, by setting $(\Theta_{k+1}, T_{k+1}) \leftarrow (\Theta_k, T_k)$.
   (iv) $\log \epsilon_{k+1} \leftarrow \log \epsilon_k + \gamma_{k+1}\big(\alpha^* - \alpha'_{\epsilon_k}(\Theta_k, \Theta'_k, T'_k)\big)$.
   (v) $\mu_{k+1} \leftarrow \mu_k + \gamma_{k+1}\big(\Theta_{k+1} - \mu_k\big)$.
   (vi) $\Gamma_{k+1} \leftarrow \Gamma_k + \gamma_{k+1}\big((\Theta_{k+1} - \mu_k)(\Theta_{k+1} - \mu_k)^\top - \Gamma_k\big)$.
3. Output $(\Theta_{n_b}, \epsilon_{n_b})$.

Appendix E. Supplementary results

Table 5. RMSEs ($\times 10^{-2}$) and acceptance rates in the Gaussian model.

Cutoff		$f(x) = x$					$f(x) = \lvert x\rvert$					Acc.
	$\epsilon_0 \setminus \epsilon$	0.10	0.82	1.55	2.28	3.00	0.10	0.82	1.55	2.28	3.00	rate
ϕ_{simple}	0.1	9.68					5.54					0.03
	0.82	8.99	3.81				5.38	2.14				0.22
	1.55	9.21	3.66	3.59			5.5	2.17	1.96			0.33
	2.28	9.67	3.86	3.6	3.97		5.85	2.28	2.02	2.08		0.4
	3.0	10.36	4.03	3.71	3.98	4.51	6.21	2.42	2.12	2.16	2.26	0.43
ϕ_{Gauss}	0.1	7.97					4.47					0.05
	0.82	7.12	3.67				4.22	2.08				0.29
	1.55	7.82	3.39	4.35			4.68	1.99	2.52			0.38
	2.28	8.94	3.59	3.81	5.52		5.26	2.2	2.29	3.29		0.41
	3.0	9.93	4.01	3.97	4.81	6.76	5.95	2.44	2.44	2.92	4.1	0.42

Table 6. Coverages for the adaptive algorithm in the Gaussian model, for tolerance $\epsilon = 0.1$.

	ϕ_{simple}						ϕ_{Gauss}					
	Fixed tolerance					Adapt	Fixed tolerance					Adapt
ϵ_0	0.1	0.82	1.55	2.28	3.0	0.64	0.1	0.82	1.55	2.28	3.0	0.28
x	0.93	0.97	0.97	0.98	0.98	0.96	0.93	0.94	0.94	0.95	0.95	0.94
$\lvert x\rvert$	0.93	0.95	0.96	0.96	0.96	0.95	0.93	0.92	0.94	0.95	0.95	0.92

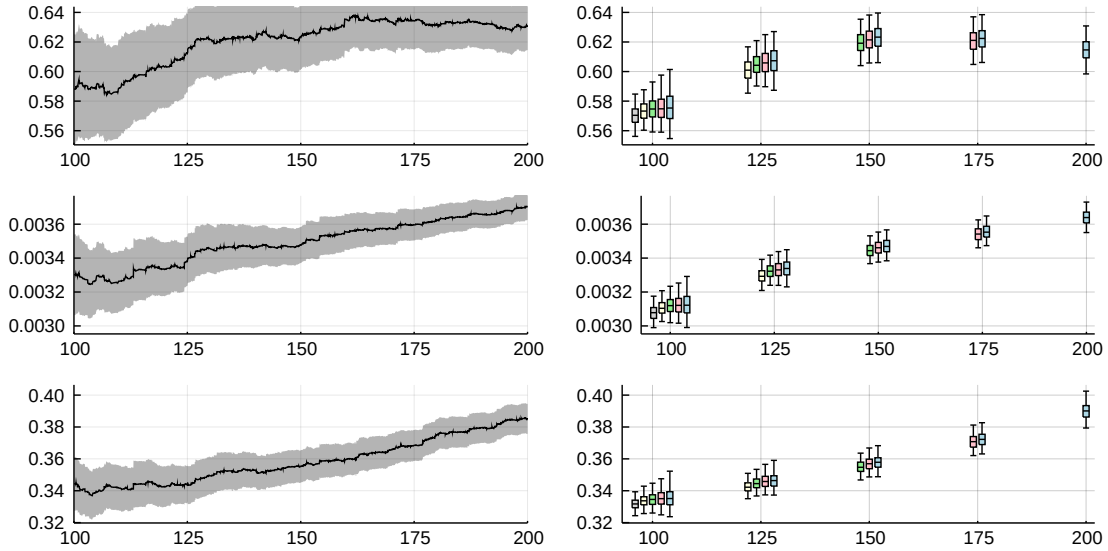


FIGURE 6. Lotka-Volterra model with simple cut-off and step size $n^{-2/3}$.

TABLE 7. RMSEs and acceptance rates in the Lotka-Volterra experiment.

	$f(\theta)=\theta_1$, RMSE $\times 10^{-2}$					$f(\theta)=\theta_2$, RMSE $\times 10^{-4}$					$f(\theta)=\theta_3$, RMSE $\times 10^{-2}$					
ϵ_0 \ ϵ	100.0	125.0	150.0	175.0	200.0	100.0	125.0	150.0	175.0	200.0	100.0	125.0	150.0	175.0	200.0	Acc. rate
100.0	5.07					3.15					2.94					0.04
125.0	1.39	1.55				0.85	0.85				1.09	0.97				0.11
150.0	1.13	1.19	1.1			0.69	0.68	0.64			0.87	0.99	0.86			0.13
175.0	1.31	1.47	1.17	1.06		0.74	0.9	0.69	0.69		0.85	1.45	1.01	0.89		0.16
200.0	1.74	1.48	1.12	0.96	0.91	1.02	0.8	0.6	0.57	0.59	1.39	1.29	0.84	0.7	0.7	0.18

TABLE 8. Coverages for the adaptive algorithm in the Lotka-Volterra model, for tolerance $\epsilon = 100$.

	Fixed tolerance					Adapt
ϵ_0	100.0	125.0	150.0	175.0	200.0	119.1
θ_1	0.59	0.97	0.99	0.99	0.98	0.95
θ_2	0.55	0.97	0.99	0.99	0.99	0.91
θ_3	0.53	0.96	0.99	0.99	0.98	0.88

TABLE 9. Frequencies of the 95% confidence intervals and mean acceptance rates in the Lotka-Volterra experiment with step size $n^{-2/3}$.

	$f(\theta)=\theta_1$					$f(\theta)=\theta_2$					$f(\theta)=\theta_3$					
ϵ_0 \ ϵ	100.0	125.0	150.0	175.0	200.0	100.0	125.0	150.0	175.0	200.0	100.0	125.0	150.0	175.0	200.0	Acc. rate
100.0	0.85					0.76					0.81					0.08
125.0	0.98	0.87				0.98	0.85				0.96	0.84				0.11
150.0	0.99	0.98	0.91			0.99	0.98	0.91			0.99	0.97	0.9			0.15
175.0	0.99	0.99	0.97	0.92		1.0	0.98	0.97	0.91		1.0	0.98	0.97	0.93		0.17
200.0	0.99	0.99	0.98	0.96	0.92	1.0	0.99	0.99	0.98	0.93	0.99	0.99	0.98	0.96	0.91	0.2

TABLE 10. RMSEs and acceptance rates in the Lotka-Volterra experiment with step size $n^{-2/3}$.

	$f(\theta)=\theta_1$, RMSE $\times 10^{-2}$					$f(\theta)=\theta_2$, RMSE $\times 10^{-4}$					$f(\theta)=\theta_3$, RMSE $\times 10^{-2}$					
ϵ_0 \ ϵ	100.0	125.0	150.0	175.0	200.0	100.0	125.0	150.0	175.0	200.0	100.0	125.0	150.0	175.0	200.0	Acc. rate
100.0	1.34					1.03					1.11					0.08
125.0	0.79	0.98				0.92	0.89				0.59	0.61				0.11
150.0	1.24	1.12	1.12			0.62	0.51	0.49			0.61	0.57	0.64			0.15
175.0	0.99	0.91	0.86	0.85		0.62	0.53	0.47	0.47		0.64	0.53	0.49	0.52		0.17
200.0	1.2	1.04	0.89	0.82	0.8	0.78	0.59	0.49	0.46	0.46	0.78	0.58	0.5	0.51	0.57	0.2

TABLE 11. RMSEs with fixed tolerance and step size $n^{-2/3}$ and with the adaptive algorithms in the Lotka-Volterra model, for tolerance $\epsilon = 100$.

	Fixed tolerance					Adapt
ϵ_0	100.0	125.0	150.0	175.0	200.0	119.1
θ_1 ($\times 10^{-2}$)	1.34	0.79	1.24	0.99	1.2	0.71
θ_2 ($\times 10^{-4}$)	1.03	0.92	0.62	0.62	0.78	0.46
θ_3 ($\times 10^{-2}$)	1.11	0.59	0.61	0.64	0.78	0.39

DEPARTMENT OF MATHEMATICS AND STATISTICS, UNIVERSITY OF JYVÄSKYLÄ P.O.BOX 35, FI-40014 UNIVERSITY OF JYVÄSKYLÄ, FINLAND

Email address: matti.s.vihola@jyu.fi, jordan.j.franks@jyu.fi