Signal Processing 142 (2018) 281–291

Contents lists available at ScienceDirect

Signal Processing

journal homepage: www.elsevier.com/locate/sigpro

Analysis of a nonlinear importance sampling scheme for Bayesian parameter estimation in state-space models

Joaquín Míguez[a,*], Inés P. Mariño[b], Manuel A. Vázquez[a]

[a] Department of Signal Theory & Communications, Universidad Carlos III de Madrid. Avenida de la Universidad 30, 28911 Leganés, Madrid, Spain
[b] Department of Biology and Geology, Physics and Inorganic Chemistry, Universidad Rey Juan Carlos. C/ Tulipán s/n, 28933 Móstoles, Madrid, Spain

ARTICLE INFO

Article history:
Received 22 January 2017
Revised 15 June 2017
Accepted 24 July 2017
Available online 25 July 2017

Keywords:
Importance sampling
Population Monte Carlo
State space models
Bayesian inference
Adaptive importance sampling
Parameter estimation

ABSTRACT

The Bayesian estimation of the unknown parameters of state-space (dynamical) systems has received considerable attention over the past decade, with a handful of powerful algorithms being introduced. In this paper we tackle the theoretical analysis of the recently proposed *nonlinear* population Monte Carlo (NPMC). This is an iterative importance sampling scheme whose key features, compared to conventional importance samplers, are (i) the approximate computation of the importance weights (IWs) assigned to the Monte Carlo samples and (ii) the nonlinear transformation of these IWs in order to prevent the degeneracy problem that flaws the performance of conventional importance samplers. The contribution of the present paper is a rigorous proof of convergence of the nonlinear IS (NIS) scheme as the number of Monte Carlo samples, M, increases. Our analysis reveals that the NIS approximation errors converge to 0 almost surely and with the optimal Monte Carlo rate of $M^{-\frac{1}{2}}$. Moreover, we prove that this is achieved even when the mean estimation error of the IWs remains constant, a property that has been termed *exact approximation* in the Markov chain Monte Carlo literature. We illustrate these theoretical results by means of a computer simulation example involving the estimation of the parameters of a state-space model typically used for target tracking.

1. Introduction

The estimation of the static unknown parameters of state-space dynamic models is a classical problem in statistical signal processing [1–6] which has also received considerable attention, very recently, from the computational statistics community [7–9] (see also [10] for a recent survey) partly because of the ubiquity of the problem in science and engineering and partly because of the availability of more powerful computational resources to address it.

The particle Markov chain Monte Carlo (pMCMC) method, originally proposed in [7], has gained popularity in the signal processing community [6,11–14]. This is a Markov chain Monte Carlo (MCMC) algorithm [15] where the target probability density function (pdf) is the posterior density of the unknown parameters conditional on the available observations. This pdf is analytically intractable and, hence, it is approximated (for each element of the chain) via particle filtering [16–20]. The most popular MCMC schemes (including Metropolis and Metropolis-Hastings algorithms) admit a pMCMC implementation. A key feature of these methods is that they have the so-called *exact approximation* property. This means that, even if the acceptance test of the MCMC algorithm is only approximate (since the true target pdf is intractable), the stationary distribution of the Markov chain is still the actual posterior density of the parameters. While popular, pMCMC procedures suffer from the same limitations as regular MCMC schemes [15,21]:

- Convergence of the chain is purely asymptotic and potentially slow: we need to generate a chain that is long enough to converge to its stationary distribution; then we need to generate a sufficiently large number of additional samples in the chain to compute any desired estimators. There are no known convergence rates, neither for the convergence of the chain to its stationary distribution nor for the convergence of the resulting Monte Carlo estimators.
- The Monte Carlo samples in the chain are correlated (hence the difficulty to obtain theoretical convergence rates). Correlation reduces the accuracy of estimators compared to methods that produce independent samples.
- If the target pdf is multimodal, MCMC algorithms may get trapped in local maxima of the function.

An alternative to pMCMC methods is to employ schemes based on importance sampling (IS) [21]. This class of techniques includes

* Corresponding author.
E-mail addresses: joaquin.miguez@uc3m.es, jmiguez@ieee.org (J. Míguez), ines.perez@urjc.es (I.P. Mariño), mvazquez@tsc.uc3m.es (M.A. Vázquez).

http://dx.doi.org/10.1016/j.sigpro.2017.07.030
0165-1684/

population Monte Carlo (PMC) [22], the sequential Monte Carlo square (SMC^2) of [23] or the nested particle filter of [9]. In general, IS methods aim at approximating a complicated, or directly intractable, *target* probability distribution by generating Monte Carlo samples from a simpler *proposal* distribution (different from the target). The samples are assigned importance weights (IWs) in order to account for the mismatch between the target and the proposal. Note that, in the setup of interest in this paper, the target is the posterior distribution of the unknown parameters of the state-space model.

The family of PMC methods includes adaptive IS schemes in which the proposal functions used to generate the samples are improved across a number of iterations [24–26]. The intuition behind this approach is rather straightforward: if we are able to produce an initial approximation of the target probability via IS, using some starting proposal distribution, then we should be able to use that approximation in order to design an improved proposal (e.g., closer to the target) that we can use to apply IS again and obtain an improved approximation. See [27–30] for recent applications, and new developments, of this methodology in statistical signal processing.

The SMC^2 method is a generalisation of the iterative batch importance sampling (IBIS) algorithm of [31]. It mimics the standard particle filter, but the Monte Carlo samples are drawn from the space of the (static) parameters and they are sequentially updated using a pMCMC kernel. All these methods, including SMC^2, are batch, meaning that the whole record of observations is typically processed many times. A purely recursive version of the SMC^2 algorithm has been proposed in [9]. The reduction in computational complexity, however, is obtained at the expense of a reduction in the convergence rate of the algorithm. It is worth mentioning that all these techniques (including pMCMC) can be fit within the theoretical framework of sequential Monte Carlo samplers introduced in [32].

The key feature of IS-based methods is the use of almost-arbitrary proposal functions to generate Monte Carlo samples and the computation of IWs for these samples. While this is a very flexible approach, it suffers from the well-known problem of degeneracy of the IWs [8,18,21,33]: when the target pdf is concentrated in a very small region of the space of the unknowns, the largest IW tends to be orders of magnitude greater than all other IWs. As a result the IS-based scheme practically yields a degenerate one-sample approximation.

In this paper we address the analysis of the nonlinear population Monte Carlo (NPMC) algorithm proposed in [8]. This is a PMC-type method, in which the proposal functions are adapted (intuitively, to be closer to the target) through an iterative scheme. The key feature of the NPMC algorithm is that the IWs undergo a nonlinear transformation to control their variance and, in this way, mitigate the degeneracy problem. In [8] it was proved that the approximation of the target distribution produced *at each iteration* of the NPMC method converges asymptotically, with the number of Monte Carlo samples M, and almost surely (a.s.). Therefore, the weight transformation preserves asymptotic convergence, while it has been shown through numerical examples that performance for finite M is consistently improved compared to conventional PMC procedures. The analysis in [8], however

- relies on the exact computation of the IWs, which is not feasible for general state-space models,
- and does not provide explicit convergence rates[1]

In this paper we analyse the performance of NPMC methods for the Bayesian estimation of the unknown parameters of state space models. In the vein of [8], we focus on the convergence of the IS estimators with transformed weights, for a fixed iteration, as the number of samples is increased (we do not analyse the convergence of the iterative process for a fixed number of samples). Based on an unbiasedness property of particle filters, we prove that IS with nonlinearly-transformed IWs also yields asymptotic convergence when the weights are approximate, i.e., computed via a particle filter with a fixed computational budget that introduces non-vanishing errors. In other words, we prove that the nonlinear importance sampler enjoys the same exact approximation property as pMCMC and SMC^2 algorithms. Moreover, the analysis of this paper also extends considerably the results of [8] by obtaining an explicit (and almost sure) estimation error rate of order $M^{-\frac{1}{2}+\epsilon}$, where $\epsilon > 0$ is an arbitrarily small constant. This result holds for approximate weights and under mild assumptions typical of classic IS analyses. It is worth mentioning that the analytical approach developed in this paper can be applied, in a rather natural way, to the study of recently proposed PMC-like algorithms [28,34] when the target distribution is the posterior density of the parameters of a state space model.

The rest of the paper is organised as follows. The necessary background material, including notation, state-space models and particle filters, is presented in Section 2. The nonlinear IS scheme and its iterative implementation (the NPMC algorithm) are detailed in Section 3 for the case in which the target probability distribution is the posterior distribution of the unknown parameters of a state-space model. In Section 4 we introduce the new analytical results on the convergence of nonlinear importance samplers, which is the main contribution of the paper. We illustrate the exact approximation property, and numerically compare the NPMC algorithm with a pMCMC scheme through computer simulations for a target tracking model in Section 5. Finally, some brief concluding remarks are made in Section 6.

2. Background and problem statement

2.1. State-space model

A Markov state-space model consists of two sequences of random variables (r.v.'s), $\{\mathbf{x}_n\}_{n\geq 0}$ and $\{\mathbf{y}_n\}_{n\geq 1}$. The first sequence, $\{\mathbf{x}_n\}$, is termed the system state. We assume it takes values on some space $\mathcal{X} \subseteq \mathbb{R}^{d_x}$, hence $\mathbf{x}_n$ is a random $d_x \times 1$ vector. The state dynamics are described by a prior probability measure $\mathcal{K}_0(\mathrm{d}\mathbf{x}_0)$ and a sequence of Markov kernels $\mathcal{K}_{n,\theta}(\mathrm{d}\mathbf{x}_n|\mathbf{x}_{n-1})$ that depend on a parameter vector $\theta \in \mathrm{S} \subset \mathbb{R}^{d_\theta}$. In this paper, θ is assumed unknown and modelled as a random vector, with prior pdf $p_0(\theta)$ with respect to (w.r.t.) the Lebesgue measure. The support set of the parameter vector, S, is assumed to be compact.

The state $\mathbf{x}_n$ cannot be observed directly. Instead, some noisy observations $\mathbf{y}_n \in \mathcal{Y} \subseteq \mathbb{R}^{d_y}$, $n = 1, 2, \ldots$, are collected. We note that $\mathbf{y}_n$ is a $d_y \times 1$ vector, with $d_y \neq d_x$ in general.

We assume that the observations are conditionally independent given the system states and the parameter vector θ, with a conditional pdf w.r.t. the Lebesgue measure, denoted $l_{n,\theta}(\mathbf{y}_n|\mathbf{x}_n) > 0$, which depends on the parameter vector θ as well.

2.2. The optimal filter and its Monte Carlo approximation

Let $\mathbf{y}_{1:n} = \{\mathbf{y}_1, \ldots, \mathbf{y}_n\}$ denote the sequence of observations collected up the time n. The posterior probability measure of the state $\mathbf{x}_n$ conditional on the observations $\mathbf{y}_{1:n}$ and the parameter vector θ is denoted $\pi_{n,\theta}$, i.e., for any Borel set $A \subset \mathcal{X}$,

$$\pi_{n,\theta}(A) = \int_A \pi_{n,\theta}(\mathrm{d}\mathbf{x}) \tag{1}$$

is the posterior probability of the event "$\mathbf{x}_n \in A$", given θ and $\mathbf{y}_{1:n}$.

[1] Error rates are found in [8] for convergence in probability (not for almost sure convergence) when the IWs are computed exactly.

Similarly, $\xi_{n,\theta}$ denotes the posterior probability measure of $\mathbf{x}_n$ conditional on θ and $\mathbf{y}_{1:n-1}$ (i.e., not including $\mathbf{y}_n$). This is often referred to as the one-step-ahead predictive measure. For a Borel set $A \subset \mathcal{X}$,

$$\xi_{n,\theta}(A) = \int_A \xi_{n,\theta}(\mathrm{d}\mathbf{x}) \tag{2}$$

is the posterior probability of the event "$\mathbf{x}_n \in A$", given θ and $\mathbf{y}_{1:n-1}$.

We refer to $\pi_{n,\theta}$ as the optimal filter conditional on the parameter vector θ. It is not possible, in general, to obtain either $\pi_{n,\theta}$ or $\xi_{n,\theta}$ in closed-form (with the notable exception of linear-Gaussian state space models, for which $\pi_{n,\theta}$ and $\xi_{n,\theta}$ are computed recursively and exactly using the Kalman filter [35]) and, therefore, numerical approximation algorithms are needed. One of the most popular schemes is the standard particle filter, also known as bootstrap filter (BF) [16,18,36].

The BF with N particles (i.e., Monte Carlo samples on the state space $\mathcal{X}$) conditional on a given parameter vector θ can be briefly outlined as follows.

1. ***Initialisation.*** Draw N samples $\mathbf{x}_0^1, \ldots, \mathbf{x}_0^N$ from the prior distribution $\mathcal{K}(\mathrm{d}\mathbf{x}_0)$. The particle approximation of $\pi_{0,\theta}(\mathrm{d}\mathbf{x}_0) \equiv \mathcal{K}_0(\mathrm{d}\mathbf{x}_0)$ is
$$\pi_{0,\theta}^N(\mathrm{d}\mathbf{x}_0) = \frac{1}{N}\sum_{i=1}^N \delta_{\mathbf{x}_0^i}(\mathrm{d}\mathbf{x}_0), \tag{3}$$
where $\delta_{\mathbf{x}_0^i}$ denotes the Dirac delta measure centred at $\mathbf{x}_0^i \in \mathcal{X}$.
2. ***Recursive step.*** Given the approximation $\pi_{n-1,\theta}^N(\mathrm{d}\mathbf{x}_{n-1}) = \frac{1}{N}\sum_{i=1}^N \delta_{\mathbf{x}_{n-1}^i}(\mathrm{d}\mathbf{x}_{n-1})$, take the following steps:
   (a) Randomly propagate each particle using the Markov kernel in the model, i.e., draw $\tilde{\mathbf{x}}_n^i$ from $\mathcal{K}_{n,\theta}(\mathrm{d}\mathbf{x}_n|\mathbf{x}_{n-1}^i)$, $i = 1, \ldots, N$.
   (b) Compute IWs, $\tilde{u}_n^i = l_{n,\theta}(\mathbf{y}_n|\tilde{\mathbf{x}}_n^i)$, for $i = 1, \ldots, N$, and
   (c) normalise them as
   $$u_n^i = \frac{\tilde{u}_n^i}{\sum_{j=1}^N \tilde{u}_n^j}, \quad i = 1, \ldots, N. \tag{4}$$
   (d) Resample: draw N times independently from the discrete distribution $\tilde{\pi}_{n,\theta}^N(\mathrm{d}\mathbf{x}_n) = \sum_{i=1}^N u_n^i \delta_{\tilde{\mathbf{x}}_n^i}(\mathrm{d}\mathbf{x}_n)$ and denote the resulting samples as $\{\mathbf{x}_n^i\}_{i=1}^N$. Construct the *unweighted* approximation $\pi_{n,\theta}^N(\mathrm{d}\mathbf{x}_n) = \frac{1}{N}\sum_{i=1}^N \delta_{\mathbf{x}_n^i}(\mathrm{d}\mathbf{x}_n)$.

The resampling step (d) above can be implemented in a number of different ways (see, e.g., [37,38] or [20] for a brief survey of methods). Here, for simplicity, we have adopted a scheme which is often referred to as multinomial resampling [18,37] but most asymptotic convergence results hold true for several other schemes as well [38,39]. The measure-valued r.v. $\pi_{n,\theta}^N$ is an approximation of the optimal filter $\pi_{n,\theta}$ (conditional on θ). Let us use the shorthand

$$(f, \pi) = \int f(\mathbf{x})\pi(\mathrm{d}\mathbf{x}) \tag{5}$$

for the integral of a real function $f : \mathbb{R}^d \to \mathbb{R}$ w.r.t. a measure π. We note that when the measure is discrete, as it is the case with $\pi_{n,\theta}^N$, the integral reduces to a sum, hence

$$(f, \pi_{n,\theta}^N) = \int f(\mathbf{x}_n)\pi_{n,\theta}^N(\mathrm{d}\mathbf{x}_n) = \frac{1}{N}\sum_{i=1}^N f(\mathbf{x}_n^i). \tag{6}$$

Under very mild assumptions it can be shown that

$$\lim_{N\to\infty}(f, \pi_{n,\theta}^N) = (f, \pi_{n,\theta}) \tag{7}$$

almost a.s. for any bounded function $f : \mathcal{X} \to \mathbb{R}$ [38,39]. Moreover, if we denote $||f||_\infty = \sup|f(\mathbf{x})|$, $E[Z]$ indicates the expected value of a r.v. Z and $||Z||_p = (E[|Z|^p])^{\frac{1}{p}}$ is its L_p norm ($p \geq 1$), then it can be proved [40] that

$$||(f, \pi_{n,\theta}^N) - (f, \pi_{n,\theta})||_p \leq \frac{C_n||f||_\infty}{\sqrt{N}} \tag{8}$$

where C_n is a constant independent of N.

The algorithm also produces a Monte Carlo approximation of the predictive measure $\xi_{n,\theta}$, namely

$$\xi_{n,\theta}^N(\mathrm{d}\mathbf{x}_n) = \frac{1}{N}\sum_{i=1}^N \delta_{\tilde{\mathbf{x}}_n^i}(\mathrm{d}\mathbf{x}_n). \tag{9}$$

If we write $\mathbf{y} = \mathbf{y}_{1:n}$ for the complete sequence of observations up to time n, it turns out that the conditional pdf of $\mathbf{y}$ given the parameter vector θ, denoted $\ell(\mathbf{y}|\theta)$, can be written in terms of integrals w.r.t. to the predictive measures $\xi_{k,\theta}$, $k = 1, \ldots, n$. To be specific,

$$\ell(\mathbf{y}|\theta) = \prod_{k=1}^n (l_{k,\theta}(\mathbf{y}_k|\cdot), \xi_{k,\theta}), \tag{10}$$

where

$$(l_{k,\theta}(\mathbf{y}_k|\cdot), \xi_{k,\theta}) = \int_{\mathcal{X}} l_{k,\theta}(\mathbf{y}_k|\mathbf{x}_k)\xi_{k,\theta}(\mathrm{d}\mathbf{x}_k). \tag{11}$$

The conditional pdf $\ell(\mathbf{y}|\theta)$ is the likelihood of the parameter vector θ given the available data $\mathbf{y}$ and the BF yields the straightforward estimator

$$\ell^N(\mathbf{y}|\theta) = \prod_{k=1}^n (l_{k,\theta}(\mathbf{y}_k|\cdot), \xi_{k,\theta}^N) \tag{12}$$

which can be shown to be unbiased (i.e., $E[\ell^N(\mathbf{y}|\theta)] = \ell(\mathbf{y}|\theta)$) under very mild assumptions ([39], Theorem 7.4.2).

2.3. Problem statement

Let $\mathbf{y}_{=}\{\mathbf{y}_1, \ldots, \mathbf{y}_R\}$ be the available data set, with $R < \infty$. Our goal is to approximate the probability measure associated to the posterior pdf of the parameter vector, θ, given the data, $\mathbf{y}$. We denote this pdf as $p(\theta|\mathbf{y})$ and it is straightforward to show, using Bayes' theorem, that

$$p(\theta|\mathbf{y}) \propto \ell(\mathbf{y}|\theta)p_0(\theta) \tag{13}$$

where, we recall, $p_0(\theta)$ is the prior pdf of θ.

In the next section, we describe an iterative importance sampling algorithm, originally introduced in [8], for the approximation of $p(\theta|\mathbf{y})\mathrm{d}\theta$.

3. Algorithm

3.1. The procedure

The NPMC algorithm of [8] is an iterative importance sampling (IS) scheme that seeks to approximate a target probability distribution, in our case given by the posterior pdf $p(\theta|\mathbf{y})$, using weighted Monte Carlo samples. It generates a sequence of proposal pdf's $q_k(\theta)$, $k = 1, \ldots, K$, from which samples can be drawn and importance weights (IWs) can be computed. This sequence of proposals is expected to yield increasingly better approximations of the target as the algorithm converges. The key feature of the NPMC method, which departs from the classical PMC technique of [22], is to compute a set of *transformed* importance weights (TIWs) by applying a nonlinear function to the standard IWs. The aim of this transformation is to mitigate the well-known problem of the degeneracy of the IWs (common to many IS methods, see [8,18]) by controlling the weight variability.

For the case of general state space models, an additional difficulty encountered when trying to estimate the unknown model parameters (denoted θ in our setup) is that the likelihood $\ell(\mathbf{y}|\theta)$ is intractable[2]. In the last few years, though, it has become a common approach to approximate this likelihood via particle filtering (PF) (see, e.g., [7,8,23,42]). To be specific, we let $\ell^N(\mathbf{y}|\theta)$ stand for the approximation of $\ell(\mathbf{y}|\theta)$ computed using a standard bootstrap filter (BF) [16,43] with N particles (see Eq. (12) in Section 2.2). One key feature of this approach, that we exploit for our analysis in Section 4, is that $\ell^N(\mathbf{y}|\theta)$ can be proved to be an unbiased estimator of $\ell(\mathbf{y}|\theta)$ [39,44]. Note that, since this property holds for a more general class of particle filters [39], the BF can be replaced by more efficient algorithms that reduce the running times or improve the accuracy of the likelihood estimates. For simplicity, however, we assume that $\ell^N(\mathbf{y}|\theta)$ is computed via a simple BF in the rest of this paper.

The NPMC algorithm applied to a state space model, with K iterations, M Monte Carlo samples per iteration, plain Gaussian proposals $\{q_k\}_{k\geq 1}$, and approximate likelihoods is outlined below.

Initialisation . Draw M i.i.d. samples $\theta_0^1, \theta_0^2, \ldots, \theta_0^M$ from the prior pdf $p_0(\theta)$. Then,

1. compute non-normalised IWs $\tilde{\mathrm{w}}_0^i \propto \ell^N(\mathbf{y}|\theta_0^i)$, $i = 1, \ldots, M$,
2. compute TIWs as $\hat{\mathrm{w}}_0^i = \mathcal{T}_M\left(i, \{\tilde{\mathrm{w}}_0^j\}_{j=1}^M\right)$, where $\mathcal{T}_M : \{1, \ldots, M\} \times \{\tilde{\mathrm{w}}_0^j\}_{j=1}^M \to [0, +\infty)$ is a nonlinear transformation, and
3. normalise the TIWs, $\mathrm{w}_0^i = \frac{\hat{\mathrm{w}}_0^i}{\sum_{j=1}^M \hat{\mathrm{w}}_0^j}$, $i = 1, \ldots, M$.

Iteration . For $k = 1, \ldots, K$, take the following steps:

1. Let $q_k(\theta) = \mathcal{N}(\theta|\mu_k, \Sigma_k)$ be a multivariate Gaussian pdf with mean vector and covariance matrix obtained, respectively, as

$$\mu_k = \sum_{i=1}^M \mathrm{w}_{k-1}^i \theta_{k-1}^i \quad \text{and} \quad \Sigma_k = \sum_{i=1}^M \mathrm{w}_{k-1}^i \left(\theta_{k-1}^i - \mu_k\right)\left(\theta_{k-1}^i - \mu_k\right)^\top. \tag{14}$$

   Note that the random variates θ_{k-1}^i, $i = 1, \ldots, M$, are $d_\theta \times 1$ vectors. The superscript $^\top$ denotes transposition.
2. Draw i.i.d. samples θ_k^i, $i = 1, \ldots, M$, from $q_k(\theta)$.
3. Compute IWs, $\tilde{\mathrm{w}}_k^i = \frac{\ell^N(\mathbf{y}|\theta_k^i)p_0(\theta_k^i)}{q_k(\theta_k^i)}$, $i = 1, \ldots, M$.
4. Compute TIWs, $\hat{\mathrm{w}}_k^i = \mathcal{T}_M\left(i, \{\tilde{\mathrm{w}}_k^j\}_{j=1}^M\right)$, $i = 1, \ldots, M$, using the same nonlinear map as for $k = 0$.
5. Normalise the TIWs, $\mathrm{w}_k^i = \frac{\hat{\mathrm{w}}_k^i}{\sum_{j=1}^M \hat{\mathrm{w}}_k^j}$, $i = 1, \ldots, M$.

3.2. Remarks

The NPMC algorithm described above is an adaptive IS algorithm where the proposal pdf $q_k(\theta) = \mathcal{N}(\theta|\mu_k, \Sigma_k)$ is updated at each iteration k (see Eq. (14) above). The intuition is that using the latest Monte Carlo approximation of the posterior distribution, given by the weighted sample $\{\theta_{k-1}^i, \mathrm{w}_{k-1}^i\}_{i=1}^M$, we should obtain a proposal q_k that is closer to the target than its predecessor q_{k-1}. Hence, as we obtain a better Monte Carlo approximations we can construct a better proposal which, in turn, should yield a better Monte Carlo approximation at the next iteration. Note, however, that the sequence of proposals does not necessarily converge to the pdf $p(\theta|\mathbf{y})$, since $q_k(\theta)$ and $p(\theta|\mathbf{y})$ may belong to different families (for example, the q_k's in Section 3.1 are Gaussian while $p(\theta|\mathbf{y})$ may be multimodal). In the sequel we focus on the issue of the convergence of the Monte Carlo approximation of the posterior distribution as the sample size M increases and k is kept fixed.

The role of the nonlinear transformation $\mathcal{T}_M$ is to reduce the variability of the weights which, in turn, affects the variance of the estimators computed from the weighted sample $\{\theta_{k-1}^i, \mathrm{w}_{k-1}^i\}_{i=1}^M$. It has been found numerically, for several different examples [45], that the use of TIWs instead of standard IWs increases the effective sample size (ESS) [33]. Here, we follow [8] and choose the nonlinear map $\mathcal{T}_M$ as a "clipping" transformation. In particular, let $i_1, i_2, \ldots, i_M$ be a permutation of the indices $1, 2, \ldots, M$ such that the IWs become ordered, namely $\tilde{\mathrm{w}}_k^{i_1} \geq \tilde{\mathrm{w}}_k^{i_2} \geq \cdots \geq \tilde{\mathrm{w}}_k^{i_M}$. The clipping transformation $\mathcal{T}_M$, with parameter $1 \leq M_c \leq \sqrt{M}$, flattens the M_c largest IWs and makes them equal to the M_c-th non-normalised IW, $\tilde{\mathrm{w}}_k^{i_{M_c}}$. Specifically, for each $j = 1, \ldots, M$, we obtain

$$\hat{\mathrm{w}}_k^j = \mathcal{T}_M\left(j, \{\tilde{\mathrm{w}}_k^l\}_{l=1}^M\right) = \begin{cases} \tilde{\mathrm{w}}_k^{i_{M_c}}, & \text{if } \tilde{\mathrm{w}}_k^j \geq \tilde{\mathrm{w}}_k^{i_{M_c}}, \\ \tilde{\mathrm{w}}_k^j, & \text{if } \tilde{\mathrm{w}}_k^j < \tilde{\mathrm{w}}_k^{i_{M_c}}, \end{cases} \tag{15}$$

Other choices of $\mathcal{T}_M$ are possible. For example, a tempering transformation of the form $\hat{\mathrm{w}}_k^i = \left(\tilde{\mathrm{w}}_k^i\right)^{\epsilon_k}$, where $0 < \epsilon_k < 1$, is also investigated in [8]. However, this transformation makes the algorithm harder to tune. In particular, the performance of the NPMC scheme with tempering is sensitive to the choice of the sequence of exponents $\epsilon_0, \epsilon_1, \ldots, \epsilon_k, \ldots$, a problem similar to the "scheduling" of simulated annealing algorithms [21]. Moreover, this choice is model dependent (a certain sequence of ϵ_k's may work well for one state space model and not for others). The clipping transformation, on the other hand, is easier to tune. Any value $M_c \leq \sqrt{M}$ guarantees convergence and, in practice, performance has been found to be stable across a broad range of values of this parameter [45]. The choice of Gaussian proposals (in step 1 of the *Iteration*) is made merely for simplicity. Other (more efficient) possibilities exist, but we stick to this formulation as it is sufficient for the purpose of this paper.

Given $A \subseteq \mathsf{S}$, being S the support set of the parameter vector θ described in Section 2, let $\mu_{\mathbf{y}}(A) = \int_A p(\theta|\mathbf{y})\mathrm{d}\theta$ denote the posterior probability measure (conditional on the observed data $\mathbf{y}$) associated to the parameter vector θ. This measure yields the full probabilistic description of θ given the available observations. At each iteration k, the NPMC method yields a random, discrete approximation of the form

$$\mu_{\mathbf{y},k}^M(\mathrm{d}\theta) = \sum_{i=1}^M \mathrm{w}_k^i \delta_{\theta_k^i}(\mathrm{d}\theta), \tag{16}$$

where $\delta_{\theta_k^i}$ denotes the Dirac delta measure centred at θ_k^i. Given (16) it is possible to approximate any moments of $\mu_{\mathbf{y}}$. For example, the posterior mean,

$$\hat{\theta}_* = \int_{\mathsf{S}} \theta \mu_{\mathbf{y}}(\mathrm{d}\theta) \approx \hat{\theta}_k^M = \sum_{i=1}^M \theta_k^i \mathrm{w}_k^i, \tag{17}$$

or its associated mean square error (MSE),

$$\mathrm{MSE}(\hat{\theta}_*) = \int_{\mathsf{S}} \|\theta - \hat{\theta}_*\|^2 \mu_{\mathbf{y}}(\mathrm{d}\theta) \approx \mathrm{MSE}(\hat{\theta}_k^M) = \sum_{i=1}^M \mathrm{w}_k^i \|\theta_k^i - \hat{\theta}_k^M\|^2. \tag{18}$$

In the next section we analyse the convergence of the approximate measure $\mu_{\mathbf{y},k}^M$ as $M \to \infty$ in a single iteration (i.e., for a given

[2] An alternative approach to problems involving intractable likelihoods, where IS or MCMC schemes are not easy to use, is the so-called approximate Bayesian computation (ABC) method [41]. In most scenarios where ABC has been applied, the intractability stems from the lack of a suitable model for the data (hence *there is no likelihood*). In such case, the ABC approach involves (i) proposing candidate values for θ, (ii) simulating synthetic data $\tilde{\mathbf{y}}$ for each candidate θ and (iii) comparing them to the actual data $\mathbf{y}$ by means of a heuristically-defined distance. Compared to IS methods, there are no weights and the use of a this heuristic distance replaces the model-based likelihood $\ell(\mathbf{y}|\theta)$.

k) when the number of particles N used to approximate the likelihood via the BF (i.e., the estimate $\ell^N(\mathbf{y}|\theta)$ of $\ell(\mathbf{y}|\theta)$) is kept constant and finite.

4. Analysis

Consider a single iteration k in the NPMC algorithm, with a fixed importance density $q_k \equiv q$. We refer to the random measure $\mu^M_{\mathbf{y},k}(\mathrm{d}\theta) = \sum_{i=1}^M \mathrm{w}^i_k \delta_{\theta^i_k}(\mathrm{d}\theta)$ computed via the TIWs w^i_k, $i = 1, .., M$, as a nonlinear importance sampling (NIS) approximation of $\mu_{\mathbf{y}}(\mathrm{d}\theta)$. Our aim in this section is to assess whether $\mu^M_{\mathbf{y},k}(\mathrm{d}\theta)$ converges towards the true measure $\mu_{\mathbf{y}}(\mathrm{d}\theta)$ or not as $M \to \infty$. To do this, there are two issues that need to be handled and make the analysis more difficult compared to a conventional IS method (that relies on the standard IWs, rather than the TIWs). These issues are:

(i) the distortion in the Monte Carlo approximation due to the clipping of the weights, which introduces additional bias (compared to the use of standard IWs); and
(ii) the impossibility to compute the IWs, and hence the TIWs, exactly, since the likelihood $\ell(\mathbf{y}|\theta)$ is intractable and we work with the particle approximation $\ell^N(\mathbf{y}|\theta)$ instead.

In [8] it was proved that, when the IWs can be computed exactly, the NIS approximation converges almost surely (a.s.) towards the target probability measure as $M \to \infty$, which accounts for (i) above.[3] The problem of the approximate computation of the weights was partially addressed in [46], for a relatively simple case where the errors in the IWs where assumed deterministic and bounded. However, the estimation problem studied in [46] (parameter estimation for α-stable distributions using iid data) did not involve any dynamics and the convergence analysis only showed an upper bound for the approximation errors that included a deterministic constant, namely a non-vanishing term proportional to the approximation error of the IWs.

Here, we show stronger analytical results that ensure the almost sure convergence of the NIS approximation when $M \to \infty$ and the likelihood function can only be estimated as $\ell^N(\mathbf{y}|\theta)$, i.e., using a BF with a finite and fixed number of particles N. Under assumptions which are standard in the classical IS theory, we prove that integrals of the form $\int f(\theta)\mu^M_{\mathbf{y},k}(\mathrm{d}\theta)$ converge towards $\int f(\theta)\mu_{\mathbf{y},k}(\mathrm{d}\theta)$ a.s. as $M \to \infty$ and provide explicit error rates.

4.1. Notation

Since we focus our attention in the NIS scheme alone, i.e., a single iteration of the proposed algorithm, in the remaining of this section we drop the iteration index k. Hence, we assume a fixed importance density $q(\theta)$, from where M independent Monte Carlo samples, $\theta^1, \theta^2, \ldots, \theta^M$, are drawn. Since the observations $\mathbf{y}$ are assumed arbitrary but fixed, we drop them from the likelihood notation and write

$$\ell(\theta) \triangleq \ell(\mathbf{y}|\theta) \quad \text{and} \quad \ell^N(\theta) \triangleq \ell^N(\mathbf{y}|\theta). \tag{19}$$

Similarly, we simplify the notation for the posterior pdf and write $p(\theta) = p(\theta|\mathbf{y})$ and $\mu(\mathrm{d}\theta) = \mu_{\mathbf{y}}(\mathrm{d}\theta)$. Then, the non-normalised IWs are approximated as

$$\tilde{\mathrm{w}}^i = g^N(\theta^i) \triangleq \frac{\ell^N(\theta^i)p_0(\theta^i)}{q(\theta^i)}, \tag{20}$$

where we have introduced the weight function $g^N \triangleq \ell^N p_0/q$ as a shorthand. This weight function is a random approximation of the deterministic function $g = \ell p_0/q$. The support of g is the same as the support of q, ℓ and p_0, denoted $\mathsf{S} \subseteq \mathbb{R}^{d_\theta}$. We assume that $g(\theta) > 0$ for every $\theta \in \mathsf{S}$ as well (a standard assumption in classical IS). It is also apparent that $p \propto gq$, where p is the posterior pdf, and the proportionality constant is independent of θ.

The non-normalised TIWs computed via the clipping function (15) are denoted

$$\hat{\mathrm{w}}^i = [\mathcal{T}_M \circ g^N](\theta^i), \tag{21}$$

where $\circ$ represents function composition and we omit the index argument of (15) for conciseness (its value is clear from the notation in any case). The normalised TIWs are $\mathrm{w}^i = \frac{\hat{\mathrm{w}}^i}{\sum_{j=1}^M \hat{\mathrm{w}}^j}$, and they are used to compute the approximate measure $\mu^M(\mathrm{d}\theta) = \sum_{i=1}^M \delta_{\theta^i}(\mathrm{d}\theta)\mathrm{w}^i$.

4.2. Assumptions and preliminary results

Let the state sequence $\{\mathbf{x}_n\}_{n\geq 0}$ take values on $\mathcal{X} \subseteq \mathbb{R}^{d_x}$. We make the following classical assumptions on the conditional pdf of the observations $\mathbf{y}_n$, $n = 1, 2, \ldots, R$, the prior density of the parameters, $p_0(\theta)$, and the importance function $q(\theta)$.

Assumption 1. (A.1) The observation sequence $\mathbf{y}_{1:R}$ is arbitrary but fixed. The functions $l_n(\mathbf{y}_n|\cdot) : \mathcal{X} \to (0, \infty)$, $n = 1, 2, \ldots, R$, are uniformly bounded, i.e., there exists a finite and positive constant $\|l\|_\infty$ such that

$$\|l\|_\infty = \sup_{n\geq 1, \mathbf{x}_n \in \mathcal{X}, \theta \in \mathsf{S}} l_{n,\theta}(\mathbf{y}_n|\mathbf{x}_n) < \infty. \tag{22}$$

Assumption 2. (A.2) The ratio of pdf's $\frac{p_0(\theta)}{q(\theta)}$ is bounded on S, i.e., there exists a positive and finite constant $\left\|\frac{p_0}{q}\right\|_\infty$ such that

$$\left\|\frac{p_0}{q}\right\|_\infty = \sup_{\theta\in\mathsf{S}} \left|\frac{p_0(\theta)}{q(\theta)}\right| < \infty. \tag{23}$$

Remark 1. Assumption A.1 states that the likelihood function $l_{n,\theta}(\cdot|\mathbf{x}_n)$ should be uniformly bounded and assumption A.2 is classical in IS as it states that the proposal function should have heavier tails that the target density. In many practical problems it is natural to determine a priori a closed range of values for the parameters of interest (e.g., the minimum radio-signal power that a sensor can measure is not arbitrarily low, due to obvious physical limitations) and this often translates into a support set S that is compact. In that case, the regularity assumptions A.1 and A.2 are more easily satisfied. For example, A.2 holds whenever both $p_0(\theta) < \infty$ and $q(\theta) > 0$ in S, irrespective of the decay of their tails. In other cases, one may also choose S to exclude singular points of $l_{n,\theta}(\cdot|\mathbf{x}_n)$ and ensure that A.1 holds.

The following lemma plays a key role in the asymptotic convergence analysis of the approximation $\mu^M(\mathrm{d}\theta)$. It states that $\ell^N(\theta)$ is an unbiased estimator of the likelihood $\ell(\theta)$ and enables us to show that the NIS scheme converges when $M \to \infty$, even if the number of particles N in the approximation $\ell^N(\theta)$ remains finite and constant.

Lemma 1. *If Assumption 1 holds then*

$$\max\{\ell(\theta), \ell^N(\theta)\} \leq \|l\|^R_\infty < \infty \quad \text{and} \quad E\left[\ell^N(\theta)\right] = \ell(\theta) \tag{24}$$

independently of N.

Proof. From the definition of $\ell(\theta)$ in Eq. (10) and its estimator $\ell^N(\theta)$ in Eq. (12), it is clear that both $\ell(\theta) \leq \|l\|^R_\infty$ and $\ell^N(\theta) \leq \|l\|^R_\infty$ when R is the number of available observations.

The equality $E\left[\ell^N(\theta)\right] = \ell(\theta)$ (i.e., $\ell^N(\theta)$ is unbiased) has been shown to hold in [39, Theorem 7.4.2]. See also [44, Lemma 2] for an alternative proof that does not rely on the Feynmann-Kac framework of [39]. □

[3] The analysis of [8] does not provide an error rate, though. Such rate is explicitly derived in this paper

The result below is adapted from [47, Lemma 4.1]. When α^M is a random approximation of some real magnitude α, it enables us to guarantee the convergence of the random error $|\alpha^M - \alpha|$ whenever $E[|\alpha - \alpha^M|^p] \to 0$ sufficiently fast and for sufficiently large p, where $E[\cdot]$ denotes expectation w.r.t. the distribution of the r.v. α^M.

Lemma 2. *Let $\{\alpha^M\}_{M\geq 1}$ be a sequence of non-negative r.v.'s such that, for every $p \geq 4$,*

$$E\left[\left(\alpha^M\right)^p\right] \leq \frac{c}{M^{\frac{p}{2}-\nu}} \tag{25}$$

where $c < \infty$ and $0 \leq \nu < 1$ are constants independent of M. Then, for every $\epsilon \in (0, \frac{1}{2})$ there exists an a.s. finite r.v. U_ϵ independent of M such that

$$\theta^M \leq \frac{U_\epsilon}{M^{\frac{1}{2}-\epsilon}}.$$

Proof. See Appendix A. The proof follows closely the argument in [47, Lemma 4.1], although the statement in that paper differs from Lemma 2 in the exponent of M on the right hand side of inequality (25). □

4.3. Asymptotic convergence, error rates and exact approximation

In the sequel we look into the approximation of integrals of the form

$$(f, \mu) \triangleq \int_{\mathsf{S}} f(\theta)\mu(\mathrm{d}\theta), \tag{26}$$

where f is a bounded real function on the parameter space S. We use $\|f\|_\infty \triangleq \sup_{\theta\in\mathsf{S}} |f(\theta)| < \infty$ to denote the supremum norm of a bounded function, while the set of bounded functions on S is denoted $B(\mathsf{S})$. The approximations of interest are

$$(f, \mu) \approx (f, \mu^M) = \sum_{i=1}^{M} f(\theta^i)\mathrm{w}^i, \tag{27}$$

for any $f \in B(\mathsf{S})$.

The following theorem yields an explicit upper bound for the (random) approximation error $|(f, \mu^M) - (f, \mu)|$. The bound is proportional to $M^{-\frac{1}{2}+\epsilon}$ (for an arbitrarily small $\epsilon > 0$) and, therefore, it vanishes as $M \to \infty$, independently of the number of particles N used in the approximate likelihoods $\ell^N(\theta^i)$.

Theorem 1. *Assume that A.1 and A.2 hold, $M_c \leq \sqrt{M}$ and $\int_{\mathsf{S}} \ell(\theta)p_0(\theta)\mathrm{d}\theta = (\ell, p_0) > 0$. Then, for every $\epsilon \in \left(0, \frac{1}{2}\right)$ (arbitrarily small) and every $f \in B(\mathsf{S})$ there exists a positive and a.s. finite r.v. $V_{f,\epsilon}$, independent of M and M_c, such that*

$$|(f, \mu^M) - (f, \mu)| \leq \frac{V_{f,\epsilon}}{M^{\frac{1}{2}-\epsilon}}. \tag{28}$$

In particular, $\lim_{M\to\infty} |(f, \mu^M) - (f, \mu)| = 0$ a.s.

Proof. Recall the intractable weight function $g = \ell p_0/q$ and its random estimator $g^N = \ell^N p_0/q$. The integral of any $f \in B(\mathsf{S})$ w.r.t. the posterior measure $\mu(\mathrm{d}\theta) \propto \ell(\theta)p_0(\theta)\mathrm{d}\theta$ can be written as

$$(f, \mu) = \frac{(fg, q)}{(g, q)} \tag{29}$$

by simply noting that $g(\theta)q(\theta) = \ell(\theta)p_0(\theta)$. Similarly, for the random measure $\mu^M(\mathrm{d}\theta)$ we can write

$$(f, \mu^M) = \frac{(f[\mathcal{T}_M \circ g^N], q^M)}{(\mathcal{T}_M \circ g^N, q^M)} \tag{30}$$

where $q^M(\mathrm{d}\theta) = \frac{1}{M}\sum_{i=1}^M \delta_{\theta^i}(\mathrm{d}\theta)$ is the Monte Carlo approximation of the proposal distribution (with pdf $q(\theta)$) and $\circ$ denotes composition of functions, hence $[\mathcal{T}_M \circ g^N](\theta^i) = \mathcal{T}_M(g^N(\theta^i))$ is the transformed weight associated to θ^i.

Given Eqs. (29) and (30) it is straightforward to show that

$$\begin{aligned}(f, \mu^M) - (f, \mu) = &\frac{(f[\mathcal{T}_M \circ g^N], q^M) - (fg, q)}{(g, q)} \\ &+(f, \mu^M)\frac{(g, q) - (\mathcal{T}_M \circ g^N, q^M)}{(g, q)}.\end{aligned} \tag{31}$$

Since $(f, \mu^M) \leq \|f\|_\infty < \infty$ and $(g, q) = (\ell, p_0)$, where $(\ell, p_0) > 0$ by assumption, Eq. (31) readily yields

$$\begin{aligned}|(f, \mu^M) - (f, \mu)| \leq &\frac{1}{(\ell, p_0)}\left|(f[\mathcal{T}_M \circ g^N], q^M) - (fg, q)\right| \\ &+\frac{\|f\|_\infty}{(\ell, p_0)}\left|(\mathcal{T}_M \circ g^N, q^M) - (g, q)\right|\end{aligned} \tag{32}$$

and, therefore, the problem of calculating bounds for $|(f, \mu^M) - (f, \mu)|$ reduces to the problem of computing bounds for errors of the form

$$|(b[\mathcal{T}_M \circ g^N], q^M) - (bg, q)|, \tag{33}$$

for $b \in B(\mathsf{S})$.

Choose any $b \in B(\mathsf{S})$. A simple triangle inequality yields

$$\begin{aligned}|(b[\mathcal{T}_M \circ g^N], q^M) - (bg, q)| \leq &|(b[\mathcal{T}_M \circ g^N], q^M) - (bg^N, q^M)| \\ &+|(bg^N, q^M) - (bg, q)|.\end{aligned} \tag{34}$$

It is straightforward to obtain an upper bound for the first term on the right hand side of the inequality (34). Indeed, by construction of $\mathcal{T}_M$ (see Eq. (15)) we readily obtain

$$\begin{aligned}|(b[\mathcal{T}_M \circ g^N], q^M) - (bg^N, q^M)| &= \left|\frac{1}{M}\sum_{r=1}^{M_c} b(\theta^{i_r})\left[g^N(\theta^{i_{M_c}}) - g^N(\theta^{i_r})\right]\right| \\ &\leq 2\|l\|_\infty^R \left\|\frac{p_0}{q}\right\|_\infty \|b\|_\infty \frac{M_c}{M}\end{aligned} \tag{35}$$

where the inequality follows from the bound $g^N \leq \|l\|_\infty^R \left\|\frac{p_0}{q}\right\|_\infty$, which is a straightforward consequence of assumptions A.1 and A.2 and the definition of the estimate ℓ^N produced by the BF (see Eq. (12)).

Finding a suitable bound for the second term on the right hand side of the inequality (34) takes some more effort. Choose, again, any $b \in B(\mathsf{S})$. A simple triangle inequality yields

$$|(bg^N, q^M) - (bg, q)| \leq |(bg^N, q^M) - (bg, q^M)| + |(bg, q^M) - (bg, q)|. \tag{36}$$

Since $q^M = \frac{1}{M}\sum_{i=1}^M \delta_{\theta^i}$, for the second term on the right hand side of (36) we can write

$$\mathbb{E}\left[|(bg, q^M) - (bg, q)|^p\right] = \mathbb{E}\left[\left|\frac{1}{M}\sum_{i=1}^{M} Z^i\right|^p\right], \tag{37}$$

where the r.v.'s

$$Z^i = b(\theta^i)g(\theta^i) - (bg, q), \quad i = 1, \ldots, M,$$

are independent, with zero mean (recall the $\theta^{(i)}$'s are i.i.d. draws from q) and bounded, because b is bounded and A.1 and A.2 imply that $g < \|l\|_\infty^R \times \left\|\frac{p_0}{q}\right\|_\infty < \infty$. Therefore, it is an exercise in combinatorics to show that

$$\mathbb{E}\left[\left|\frac{1}{M}\sum_{i=1}^{M} Z^{(i)}\right|^p\right] \leq \frac{\tilde{c}^p \|l\|_\infty^{Rp}\left\|\frac{p_0}{q}\right\|_\infty^p \|b\|_\infty^p}{M^{\frac{p}{2}}}, \tag{38}$$

where $\tilde{c}$ is a constant independent of M and q. Combining (38) with (37) readily yields

$$\|(bg, q^M) - (bg, q)\|_p \leq \frac{\tilde{c}\|l\|_\infty^R \left\|\frac{p_0}{q}\right\|_\infty \|b\|_\infty}{\sqrt{M}}. \tag{39}$$

The inequality (39) holds for every $p \geq 1$ and it implies, via Lemma 2 with $\nu = 0$, that there exists an a.s. finite r.v. $\tilde{U}_{b,\epsilon} > 0$ such that

$$|(bg, q^M) - (bg, q)| \leq \frac{\tilde{U}_{b,\epsilon}}{M^{\frac{1}{2}-\epsilon}}, \tag{40}$$

where $0 < \epsilon < \frac{1}{2}$ is an arbitrarily small constant independent of M.

If we expand the first term on the right hand side of (36) we arrive at

$$\left|(bg^N, q^M) - (bg, q^M)\right| = \left|\frac{1}{M}\sum_{i=1}^{M} b(\theta^i)\left(g^N(\theta^i) - g(\theta^i)\right)\right| = \left|\frac{1}{M}\sum_{i=1}^{M} Z_N^i\right|, \tag{41}$$

where the r.v.'s $Z_N^i = \frac{b(\theta^i)p_0(\theta^i)}{q(\theta^i)}\left(\ell^N(\theta^i) - \ell(\theta^i)\right)$, $i = 1, 2, \ldots, M$, are independent (because the samples $\theta^1, \ldots, \theta^M$ are independent) and zero mean, as a result of Lemma 1.[4] Since they are also bounded, namely $|Z_N^i| \leq \|b\|_\infty \|l\|_\infty^R \left\|\frac{p_0}{q}\right\|_\infty$ as a consequence of A.1 and A.2, it is again an exercise to show that (41) implies

$$E\left[\left|(bg^N, q^M) - (bg, q^M)\right|^p\right] \leq \frac{\bar{c}^p \|l\|_\infty^{Rp} \left\|\frac{p_0}{q}\right\|_\infty^p \|b\|_\infty^p}{M^{\frac{p}{2}}} \tag{42}$$

in the same manner as we obtained the inequality (38). Resorting again to Lemma 2, from (42) we deduce that there exists an a.s. finite r.v. $\bar{U}_{b,\epsilon} > 0$, independent of M, such that

$$|(bg^N, q^M) - (bg, q^M)| \leq \frac{\bar{U}_{b,\epsilon}}{M^{\frac{1}{2}-\epsilon}}, \tag{43}$$

where $0 < \epsilon < \frac{1}{2}$ is an arbitrarily small constant independent of M.

Taking together (36), (40) and (43) we arrive at

$$|(bg^N, q^M) - (bg, q)| \leq \frac{U_{b,\epsilon}}{M^{\frac{1}{2}-\epsilon}}, \tag{44}$$

where $U_{b,\epsilon} = \tilde{U}_{b,\epsilon} + \bar{U}_{b,\epsilon} \geq 0$ is an a.s. finite r.v. independent of M, and $\epsilon \in \left(0, \frac{1}{2}\right)$ can be chosen to be arbitrarily small.

Substituting the inequalities (35) and (44) back into the relation (34) we arrive at the bound

$$|(b[\mathcal{T}_M \circ g^N], q^M) - (bg, q)| \leq 2\|l\|_\infty^R \left\|\frac{p_0}{q}\right\|_\infty \|b\|_\infty \frac{M_c}{M} + \frac{U_{b,\epsilon}}{M^{\frac{1}{2}-\epsilon}} \leq \frac{\tilde{V}_{b,\epsilon}}{M^{\frac{1}{2}-\epsilon}} \tag{45}$$

where the second inequality follows from the assumption $M_c \leq \sqrt{M}$ and choosing $\tilde{V}_{b,\epsilon} = 2\|l\|_\infty^R \left\|\frac{p_0}{q}\right\|_\infty \|b\|_\infty + U_{b,\epsilon}$. Since the r.v. $U_{b,\epsilon}$ is a.s. finite, $\tilde{V}_{b,\epsilon} < \infty$ a.s. as well.

To conclude the proof, we substitute the inequality (45) twice into the relation (32). To be precise, we choose $b = f$ first and use (45) to obtain a bound for the first term on the right hand side of (32). Then, we choose $b = 1$ and apply (45) again to find a bound for the second term on the right hand side of (32). As a result, we arrive at

$$|(f, \mu^M) - (f, \mu)| \leq \frac{\tilde{V}_{f,\epsilon}}{(\ell, p_0)} \times \frac{1}{M^{\frac{1}{2}-\epsilon}} + \frac{\|f\|_\infty \tilde{V}_{1,\epsilon}}{(\ell, p_0)} \times \frac{1}{M^{\frac{1}{2}-\epsilon}}. \tag{46}$$

Since $(\ell, p_0) > 0$ by assumption of Theorem 1, taking

$$V_{f,\epsilon} = \frac{1}{(\ell, p_0)}\left(\tilde{V}_{f,\epsilon} + \|f\|_\infty \tilde{V}_{1,\epsilon}\right) < \infty \quad \text{a.s.} \tag{47}$$

leads to the desired result and concludes the proof. □

[4] Note that $E\left[Z_N^i|\theta^i\right] = \frac{b(\theta^i)p_0(\theta^i)}{q(\theta^i)} E\left[\ell^N(\theta^i) - \ell(\theta^i)\right] = 0$, because $\ell^N(\theta^i)$ is an unbiased estimator of $\ell(\theta^i)$, hence $E\left[Z_N^i\right] = E\left[E\left[Z_N^i|\theta^i\right]\right] = 0$.

4.4. Discussion

The proof of Theorem 1 relies on assumptions A.1 and A.2, as a well as a property of the clipping transformation that is applied to obtain the inequality (35). The assumptions have already been discussed (see Remark 1); they are relatively mild and amount to the classical assumptions in the analysis of standard IS algorithms.

The inequality (35) holds for the clipping scheme and any other transformation of the IWs that can be restricted to a subset of $M_c < M$ samples. For example, Theorem 1 and its proof would still hold exactly if we apply a "soft clipping" transformation where the $M_c \leq \sqrt{M}$ samples with the highest weights are assigned different TIWs (compared to the equal-weighting of (15)). If the transformation $\mathcal{T}_M$ is such that *all* the IWs can be transformed, i.e., $\tilde{w}_k^i \neq \hat{w}_k^i$ in general for every i, then we can still have the same type of error bound as in Theorem 1 if we can guarantee that, for some constant $c < \infty$,

$$\frac{D(M)}{M} \leq \frac{c}{\sqrt{M}}, \quad \text{where} \quad D(M) = \sum_{i=1}^{M} |\tilde{w}_k^i - \hat{w}_k^i|. \tag{48}$$

Weaker types of convergence can be obtained provided that $\lim_{M\to\infty} \frac{M_c}{M} = 0$ for the clipping transformation (see the results in [8]) or assuming $\lim_{M\to\infty} \frac{D(M)}{M} = 0$ for more general classes of non-linearities.

Remark 2. Finally, we draw attention to the fact that the error $|(f, \mu^M) - (f, \mu)|$ vanishes a.s. when $M \to \infty$ even if the number of particles N in the BF remains fixed and, hence, ℓ^N does not converge to ℓ. This property has been coined "exact approximation" in the MCMC literature (see [7]).

5. Computer simulations

5.1. State-space model

In order to illustrate the performance of the NPMC algorithm and the exact approximation property granted by Theorem 1 we have carried out computer simulations for the estimation of the unknown parameters in a problem consisting of the tracking of a target moving over a region monitored by a network of sensors.

5.1.1. Target dynamics

The target moves over a closed rectangular region $\mathcal{R} = [-20, +20] \times [-10, +10]$. When it hits the border of $\mathcal{R}$, the target bounces back in according to the law of reflection [48]. The state of the system at time n is $\mathbf{x}_n = \begin{bmatrix}\mathbf{r}_n \\ \mathbf{v}_n\end{bmatrix} \in \mathbb{R}^4$, where $\mathbf{r}_n \in \mathcal{R}$ is the target position and $\mathbf{v}_n$ is its velocity. At time $n = 0$, we assume a uniform prior on $\mathcal{R}$ for the position and a zero-mean Gaussian distribution for the velocity. To be specific, the prior probability measure is defined as

$$\mathcal{K}_0(\mathrm{d}\mathbf{x}_0) = \mathcal{U}(\mathcal{R}) \times \mathcal{N}(\mathbf{0}, \frac{1}{20} \times \mathbf{I}_2) \tag{49}$$

where $\mathbf{I}_2$ is the 2×2 identity matrix, $\mathcal{U}(\mathcal{R})$ is the uniform distribution on $\mathcal{R}$ and $\mathcal{N}(\mathbf{m}, \mathbf{C})$ denotes the Gaussian distribution with mean $\mathbf{m}$ and covariance matrix $\mathbf{C}$.

At time $n > 0$, the state vector $\mathbf{x}_n$ evolves according to a linear-Gaussian equation if the target position remains within the bounded region $\mathcal{R}$ but it "reflects" back in when the target reaches a border of $\mathcal{R}$. Specifically, let

$$\tilde{\mathbf{x}}_n = \begin{bmatrix}\mathbf{I}_2 & \kappa\mathbf{I}_2 \\ 0 & \mathbf{I}_2\end{bmatrix}\mathbf{x}_{n-1} + \mathbf{u}_n, \tag{50}$$

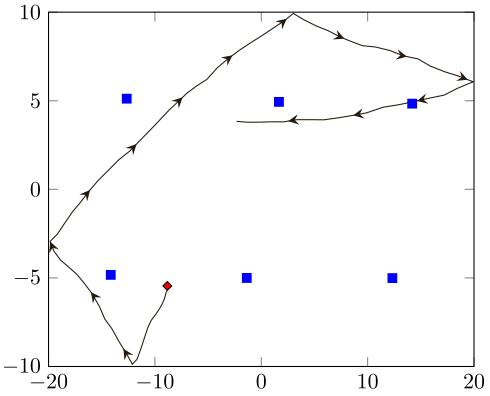

Fig. 1. Wireless sensors network with a sample trajectory overimposed. The blue squares mark the positions of the sensors, and the red diamond indicates the starting point of the trajectory, which is depicted as a black solid line. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

where $\mathbf{u}_n \sim \mathcal{N}(\mathbf{0}, \mathbf{C})$ is a Gaussian noise term with 0-mean and covariance matrix

$$\mathbf{C} = \begin{bmatrix} (\kappa\sigma_u^2 + \sigma_z^2)\mathbf{I}_2 & 0 \\ 0 & \sigma_u^2\mathbf{I}_2 \end{bmatrix}, \tag{51}$$

κ is a time-discretisation step (we assume $\kappa = 1$ in our simulations), σ_u^2 is a velocity variance parameter, and σ_z^2 is a position variance parameter. The latter are assumed known and identical, $\sigma_u^2 = \sigma_z^2 = 10^{-2}$. If $\tilde{\mathbf{x}}_n$ generated in this way is inside $\mathcal{R}$, $\tilde{\mathbf{x}}_n \in \mathcal{R}$, then $\mathbf{x}_n = \tilde{\mathbf{x}}_n$, otherwise $\mathbf{x}_n = f(\mathbf{x}_{n-1})$, where f is the reflection function detailed in Appendix B. Note that we do not provide an expression for the kernel $\mathcal{K}_n(\mathrm{d}\mathbf{x}_n|\mathbf{x}_{n-1})$ but have just described how to draw samples from it instead. This is enough for the implementation of the bootstrap filter and the PMC-like algorithms to be evaluated in this section.

For illustration, Fig. 1 depicts the region $\mathcal{R}$ and a sample trajectory (i.e., a sequence of positions $\mathbf{r}_0, \mathbf{r}_1, \ldots$) which hits the borders of $\mathcal{R}$ and is reflected back in at four different times. In the figure, the starting target position is represented by a red diamond, the direction of motion is indicated by arrows and the blue squares represent the position of the sensors used to monitor the target motion.

5.1.2. Observations

There are J sensors deployed in $\mathcal{R}$ and, at time n, each sensor collects a measurement of the power of the radio signal transmitted by the target. To be specific, the observation recorded by sensor j at time n has the form

$$y_{j,n} = 10\log\left(\frac{P_t}{||\mathbf{r}_n - \mathbf{s}_j||^{\nu}} + \rho\right) + \epsilon_{j,n} \tag{52}$$

where P_t is the power of the transmitted radio signal, $\mathbf{s}_j$ is the location of the jth sensor, $||\mathbf{r}_n - \mathbf{s}_j||$ is the distance at time n between the target and the sensor, $\nu > 0$ is the path loss exponent, ρ is the sensitivity of the sensor, i.e., the minimum power it can measure (note that $y_{j,n} \to 10\log(\rho) + \epsilon_{j,n}$ when $||\mathbf{r}_n - \mathbf{s}_j|| \to \infty$) and $\epsilon_{j,n} \sim \mathcal{N}(0, \sigma_\epsilon^2)$ is a Gaussian term accounting for observational errors. We assume $\sigma_\epsilon^2 = 1$ is a known parameter.

At each time instant n, a vector of J observations $\mathbf{y}_n = [y_{1,n}, y_{2,n}, \ldots, y_{J,n}]^\top \in \mathbb{R}^J$ is collected. The target is observed over m time instants, and hence the available dataset is $\mathbf{y} = \mathbf{y}_{1:m}$. We set $m = 80$ for our computer simulations.

5.1.3. Problem statement

Given the state space model described in Sections 5.1.1 and 5.1.2 above, we aim at estimating the unknown parameters P_t, ν and ρ. All other parameters (namely the discretisation period κ and the relevant variances) are assumed known. For all computer simulations we have set ground truth values $P_t = 0.8$, $\nu = 3$ and $\rho = 10^{-5}$ for the parameters to be estimated.

The three physical parameters P_t, ν, and ρ are necessarily positive and, therefore, it is more convenient to address the estimation of the *logarithms* of these parameters. Specifically, we apply the NPMC algorithm (together with competing algorithms to be described below) to approximate the posterior probability measure $\mu_{\mathbf{y}}(\mathrm{d}\theta)$ of the vector of unknowns $\theta = [\log P_t, \log\nu, \log\rho]^\top \in \mathbb{R}^3$ and we assume prior distributions of the form $\log P_t \sim \mathcal{N}(-0.11, 0.22)$, $\log\nu \sim \mathcal{N}(0.4, 0.56)$ and $\log\rho \sim \mathcal{N}(-11.02, 0.4)$. In natural units, the prior mean and variance of the parameters are 1 and 0.25, respectively, for P_t; 2 and 3 for ν and 2×10^{-5} and 2×10^{-10} for ρ.

The likelihood $\ell(\mathbf{y}|\theta)$ for the model does not have a closed form and, therefore, it is estimated using a BF, for the state space model described in Sections 5.1.1 and 5.1.2, to yield the approximation $\ell^N(\mathbf{y}|\theta)$ detailed in Section 2.2.

5.2. Competing methods

We have tackled this problem using

- the NPMC method described in Section 3;
- a standard PMC procedure, that only differs from the above NPMC method in that TIWs are not computed, and hence all approximations rely on the conventional IWs;
- a particle Metropolis-Hastings (pMH) algorithm [7];
- the "Adaptive Multiple Importance Sampling" (AMIS) algorithm described in [26] that differs from the standard PMC scheme in that, at every iteration, the samples from previous iterations are recycled by recomputing their weights with respect to the current proposal (and *all* samples are used for estimation);
- the AMIS algorithm using TIWs (with the same clipping transformation as in the NPMC scheme) in order to compute the proposals, labeled NAMIS.

The proposal of the PMC-like algorithms at each iteration k is a (multivariate) Gaussian distribution whose mean vector and covariance matrix are computed using the samples and weights obtained from the $(k-1)$-th iteration (these weights are the standard IWs for PMC and AMIS, and the TIWs for the NPMC and NAMIS schemes). If the Gaussian distribution turns out to be degenerate, i.e., the covariance matrix is all-zeros, then the resulting proposal function is a Dirac delta. This is not a valid proposal for IS and, in our simulations, we handle this scenario by using the same mean vector but the prior covariance matrix of the parameters.

The pMH is a representative of the class of particle MCMC methods [7] that have become popular in the past two years. It generates a Markov chain on the space of the unknown parameter vector θ according to the following procedure:

1. Draw $\theta_0 \sim p_0(\theta)$ from the prior distribution of the parameters
2. At the r-th iteration, and given the previous element θ_{r-1}:
   (a) Draw a tentative new element $\tilde{\theta}_r \sim \mathcal{N}(\theta_{r-1}, \frac{2}{10}\mathbf{C})$, where both $\mathbf{C} = \mathrm{diag}[\mathbf{0.22}, \mathbf{0.56}, \mathbf{0.4}]$ and the scale factor $\frac{2}{10}$ have been empirically chosen to optimise the performance of the algorithm.
   (b) Compute the (approximate) likelihood $\ell^N(\mathbf{y}|\tilde{\theta}_r)$ and prior density $p_0(\tilde{\theta}_r)$. The acceptance probability for $\tilde{\theta}_r$ is

$$\alpha_r = \min\left(1, \frac{\ell^N(\mathbf{y}|\tilde{\theta}_r)\, p_0(\tilde{\theta}_r)}{\ell^N(\mathbf{y}|\theta_{r-1})\, p_0(\theta_{r-1})}\right) \tag{53}$$

   (c) Draw $u_r \sim \mathcal{U}(0,1)$. If $u_r < \alpha_r$ then $\theta_r = \tilde{\theta}_r$, else $\theta_r = \theta_{r-1}$.

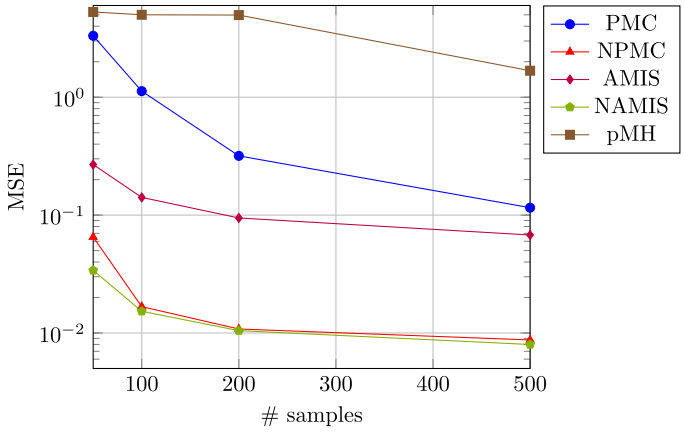

Fig. 2. MSE for several values of the number of samples per iteration of the PMC-like algorithms, M. The PMC, NPMC, AMIS and NAMIS algorithms are iterated $K = 10$ times. The pMH scheme generates a chain of length $L = M \times K$. The curves are averaged over 1000 independent simulation runs.

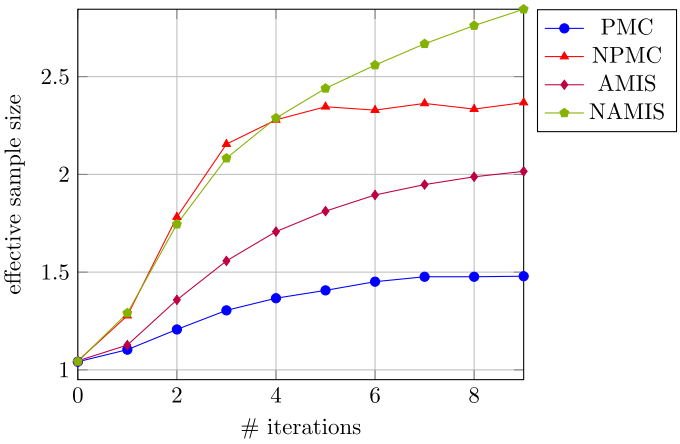

Fig. 3. Evolution of the effective sample size (ESS) across iterations of the PMC-based algorithms. The horizontal axis represents the itration number, k. The number of samples per iteration is $M = 500$. The curves have been averaged over 1000 independent simulation runs.

When we generate a chain of length L using the procedure above we set a burn-in period of $\frac{L}{2}$, hence estimates are computed from the samples $\theta_{\lfloor \frac{L}{2} \rfloor + 1}, \ldots, \theta_L$ in the chain.

To compare the pMH and PMC-like algorithms on a fair basis, we let $L = M \times K$, where K is the number of iterations of the PMC, NPMC, AMIS and NAMIS algorithms, and M is the number of samples generated per iteration.

All five methods (PMC, NPMC, AMIS, NAMIS, pMH) rely on a BF with N particles for the computation of $\ell^N(\mathbf{y}|\theta)$. The value of N is set for all algorithms as $N = 100$ unless explicitly stated otherwise.

5.3. Results

Fig. 2 shows the evolution of the MSE of the posterior-mean estimators of the parameters produced by the PMC, NPMC, AMIS, NAMIS and pMH algorithms as the number of samples per iteration, M, is increased. If the posterior-mean estimate is $\hat{\theta}_k^M = [\hat{\theta}_{1,k}^M, \hat{\theta}_{2,k}^M, \hat{\theta}_{3,k}^M]^\top$ (at the k-th iteration), then we recover parameter estimates in natural units, namely, $\hat{P}_{t,k}^M = \exp\{\hat{\theta}_{1,k}^M\}$, $\hat{\nu}_k^M = \exp\{\hat{\theta}_{2,k}^M\}$ and $\hat{\rho}_k^M = \exp\{\hat{\theta}_{3,k}^M\}$, and then compute the MSE, which is averaged over a number of independent simulation runs (1,000 for this particular computer experiment).

The MSE for the NPMC algorithm in Fig. 2 is at least one order of magnitude below the errors of the conventional PMC and the pMH algorithms for every tested value of M. For $M = 200$ samples, for example, the MSE attained by the NPMC is $\approx 1.08 \times 10^{-2}$, while for the standard PMC and pMH algorithms the errors are $\approx 3.18 \times 10^{-1}$ and ≈ 4.97, respectively. The clipping transformation also yields a performance improvement in the case of the AMIS algorithm. In this example, the MSE is reduced from $\approx 9.45 \times 10^{-2}$, for the standard AMIS method, to 1.05×10^{-2} when TIWs are used for the computation of the proposal distributions.

We have also evaluated the effective sample size (ESS) of the PMC-based algorithms (including the AMIS and NAMIS schemes) as they are iterated. The ESS provides a measure of the number of samples with non-negligible weight, and it is usually defined as [33]

$$\mathrm{ESS} = \frac{1}{\sum_{i=1}^M \left(\mathrm{w}^i\right)^2}$$

for the k-th iteration.

Fig. 3 shows the ESS as a function of the iteration counter k (with $k = 0$ corresponding to the first iteration, i.e., using the prior as a proposal). The algorithms are meant to be be compared pairwise, i.e., PMC versus NPMC, and AMIS versus NAMIS. Although all algorithms operate with a low ESS (because sampling from the prior is inefficient and yields a poor ESS$= 1$ for $k = 0$), in both cases we see that the nonlinear transformation over the IWs yields a significant increase in the ESS of the algorithms. After the 9-th iteration, the ESS of the NPMC algorithm is $\approx 60\%$ larger than the ESS of the PMC scheme. Similarly, the ESS of the NAMIS method improves by $\approx 40\%$ over the ESS of the AMIS algorithm.

Fig. 4 shows estimates of the marginal posterior pdf's of the three unknown parameters (P_t, ν and ρ) computed from the random measures $\mu_{\mathbf{y},K}^M$ generated by the NPMC and NAMIS algorithms, with $M = 200$ and $K = 10$, in a typical simulation run. Each plot in Fig. 4 shows the prior for the parameter (in natural units, not logarithmic) as a dashed black line, the true value as a solid vertical line and the kernel density estimates (KDEs) of the marginal posterior pdf for the NPMC and the NAMIS algorithms as coloured solid lines. The KDEs are computed using a common Gaussian kernel with bandwidth 1.2. Both KDEs display modes which are well aligned with the true parameters, including the path-loss exponent ν, for which the true value is on the tail of the prior (see Fig. 4c). Next, we investigate the length of the chain, L, which is required for the pMH algorithm to attain the same performance, in terms of MSE, as the NPMC algorithm. Fig. 5 shows the MSE of the pMH method for different chain lengths. Since L is the total number of generated samples, this should be compared with the number $M \times K$ of samples generated by the NPMC algorithm through K iterations. Specifically, the performance of the NPMC algorithm for $M = 500$ samples and $K = 10$ iterations ($500 \times 10 = 5,000$ Monte Carlo samples overall) is also indicated in the plot. It can be seen that, in the pMH algorithm, chains that are around 500,000 samples long are required to attain the same MSE as the NPMC algorithm (a 100-fold increase of the computational cost). While the parameters of the pMH scheme may be further tuned to improve this performance, the gap between the algorithms is large enough to conclude that the NMPC method is more efficient in this example.

Finally, we examine the exact approximation property of the NPMC scheme stated by Theorem 1. Fig. 6 shows the MSE of the NPMC algorithm versus the number of Monte Carlo samples, M, for different values of N (the number of particles used by the BF to approximate the IWs). While Theorem 1 guarantees that the approximation errors vanish as $M \to \infty$, even if N is fixed, it is reasonable to expect that for a fixed $M < \infty$, greater values of N lead to better performance. This is shown, indeed, by Fig. 6. Note, however, that the difference in performance is very small. For $M = 1,000$,

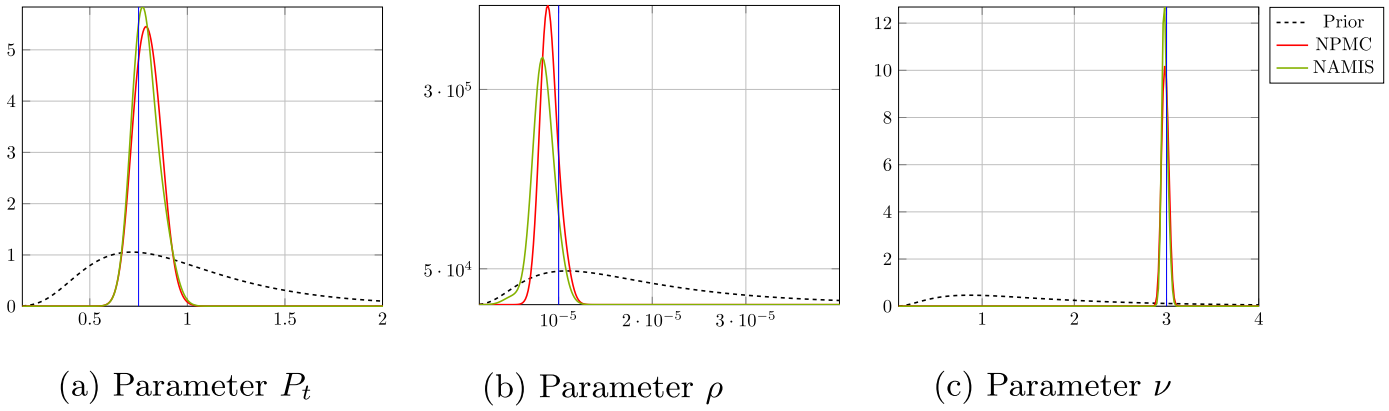

Fig. 4. Kernel density estimates of the marginal posterior pdf's of the unknown parameters obtained using the NPMC and NAMIS algorithms. The true parameter values are indicated as vertical solid lines. The prior densities are depicted as dashed curves. The densities were estimated using a Gaussian kernel with bandwidth 1.2.

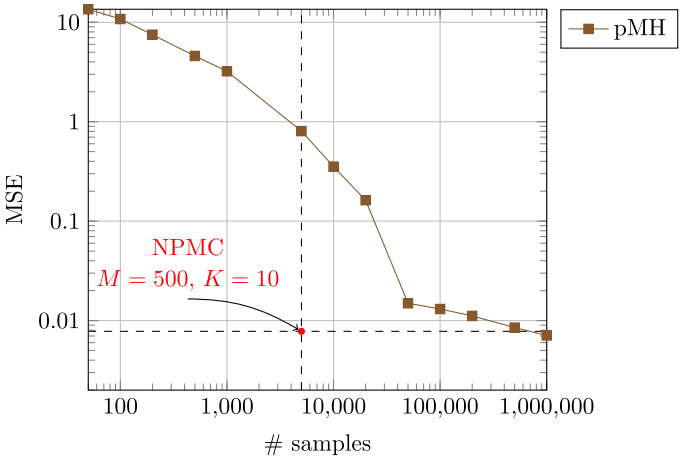

Fig. 5. MSE for different values of the chain lengths, L, of the pMH algorithm. These results have been averaged over 100 independent simulation runs.

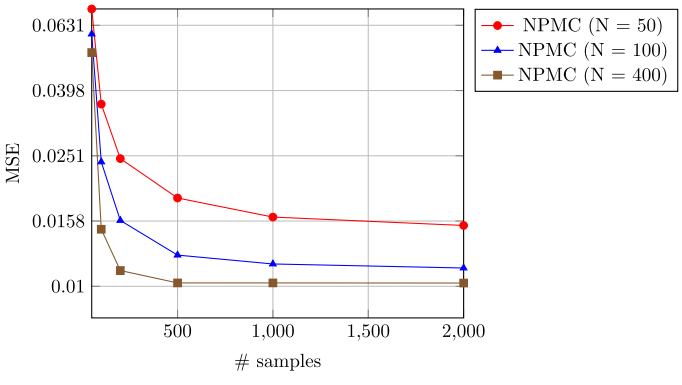

Fig. 6. MSE vs. number of samples, M, attained by the NPMC algorithm with different choices of the number of particles in the BF, N. The curves are averaged over 100 independent simulation runs.

the gap between the MSE of the NPMC scheme with $N = 400$ and the NPMC scheme with $N = 50$ is $\approx 6 \times 10^{-3}$.

6. Conclusion

We have rigorously proved, under mild assumptions, that nonlinear importance samplers with clipped IWs converge a.s. with optimal Monte Carlo error rates even when the weights can only be estimated (and have a positive, non-vanishing variance) as long as these estimates are unbiased. Therefore, nonlinear importance samplers can perform exact approximation in the same manner as, e.g., particle MCMC schemes. Besides the theoretical contribution, we have numerically shown that the proposed algorithm can be more efficient than a particle Metropolis-Hastings algorithm of the same complexity for inference on a target tracking model.

Acknowledgments

This research has been partially supported by the Spanish Ministry of Economy and Competitiveness (projects TEC2015-69868-C2-1-R ADVENTURE and FIS2013-40653-P), the Spanish Ministry of Education, Culture and Sport (mobility award PRX15/00378) and the Office of Naval Research (ONR) Global (Grant Award no. N62909-15-1-2011).

Appendix A. Proof of Lemma 2

Let us choose an arbitrary constant $\psi \in (\nu,\ 1)$ and define the r.v. $U^{\psi,\,p} = \sum_{M=1}^{\infty} M^{\frac{p}{2}-1-\psi}(\alpha^M)^p$. If (25) holds, then the expectation $E[U^{\psi,\,p}]$ is finite, as we prove in the sequel. Indeed, from Fatou's lemma,

$$E\left[U^{\psi,p}\right] \leq \sum_{M=1}^{\infty} M^{\frac{p}{2}-1-\psi} E\left[(\alpha^M)^p\right] \tag{54}$$

$$\leq c \sum_{M=1}^{\infty} M^{\nu-\psi-1}, \tag{55}$$

where (55) follows from substituting (25) into (54). Since we have chosen $\psi \in (\nu,\ 1)$, then it follows that $-1 < \nu - \psi < 0$ and $\nu - \psi - 1 < -1$, which ensures that $\sum_{M=1}^{\infty} M^{\nu-\psi-1} < \infty$ and, therefore, $E[U^{\psi,\,p}] < \infty$. Since $E[U^{\psi,\,p}] < \infty$, then $U^{\psi,\,p} < \infty$ a.s.

For any given value of M, it is apparent from the definition of $U^{\psi,\,p}$ that

$$M^{\frac{p}{2}-1-\psi}(\theta^M)^p \leq U^{\psi,p}$$

and, as a consequence,

$$\theta^M \leq \frac{(U^{\psi,p})^{\frac{1}{p}}}{M^{\frac{1}{2}-\frac{1+\psi}{p}}} = \frac{U_\epsilon}{M^{\frac{1}{2}-\epsilon}} \tag{56}$$

where the equality in (56) follows from defining $\epsilon \triangleq \frac{1+\psi}{p}$ and $U_\epsilon \triangleq (U^{\psi,p})^{\frac{1}{p}}$. Since $\psi < 1$, it is sufficient to choose $p \geq 4$ to ensure that $\epsilon = \frac{1+\psi}{p} < \frac{1}{2}$. Also, since p can actually be chosen as large as we wish, it follows that (56) holds for $\epsilon > 0$ as small as needed.

Appendix B. Definition of function $f(\cdot)$

Let us denote the upper right, upper left, lower left and lower right vertices of the monitored region by, respectively, $\mathbf{c}_0$, $\mathbf{c}_1$, $\mathbf{c}_2$

and $\mathbf{c}_3$. The sides of the rectangle, obtained by joining adjacent vertices, are denoted $\mathbf{l}_0 = \overline{\mathbf{c}_1\mathbf{c}_0}$ (top), $\mathbf{l}_1 = \overline{\mathbf{c}_1\mathbf{c}_2}$ (left), $\mathbf{l}_2 = \overline{\mathbf{c}_2\mathbf{c}_3}$ (bottom) and $\mathbf{l}_3 = \overline{\mathbf{c}_3\mathbf{c}_0}$ (right). With this notation, Algorithm 1 can be

Algorithm 1 Generation of a sample $\mathbf{x}_n \in \mathcal{R}$, conditional on $\mathbf{x}_{n-1}$.

1: Generate $\tilde{\mathbf{x}}_n = [\tilde{\mathbf{r}}_n^\top, \tilde{\mathbf{v}}_n^\top]^\top$, conditional on $\mathbf{x}_{n-1}$, using Eq. (50).
2: If $\tilde{\mathbf{r}}_n \in \mathcal{R}$ then return $\mathbf{x}_n = f(\mathbf{x}_{n-1}) = \tilde{\mathbf{x}}_n$. Otherwise, continue.
3: Compute the vectors

$$\mathbf{q}_j = \mathbf{c}_j - \mathbf{r}_{n-1}, \quad j = 0, 1, 2, 3, \quad \text{and} \quad \mathbf{s} = \tilde{\mathbf{r}}_n - \mathbf{r}_{n-1}$$

and the corresponding angles

$$\Theta_s = \angle(\mathbf{s}), \quad \Theta_j = \angle(\mathbf{q}_j), \quad j = 0, 1, 2, 3,$$

i.e., the angles of vectors $\mathbf{s}$ and $\mathbf{q}_j$, respectively, w.r.t. the horizontal axis
4: Find $j \in \{0, 1, 2, 3\}$ such that $\Theta_j < \Theta_s < \Theta_{(j+1) \bmod 4}$ and decompose $\tilde{\mathbf{r}}_n$ as

$$\tilde{\mathbf{r}}_n = \mathbf{r}_{n-1} + \mathbf{s}' + \mathbf{s}'',$$

where $\mathbf{s}' = \Lambda(j)\mathbf{s}$, $s'' = (1 - \Lambda(j))\mathbf{s}$ and

$$\Lambda(j) = \begin{cases} (\mathbf{c}_j(2) - \mathbf{r}_{n-1}(2))/\mathbf{s}(2), & \text{for } j = 0, 2 \\ (\mathbf{c}_j(1) - \mathbf{r}_{n-1}(1))/\mathbf{s}(1), & \text{for } j = 1, 3 \end{cases}$$

(with $b(j)$ denoting the j-th component of vector b).
5: Compute the vector n_j normal to $\mathbf{l}_j$ (namely $\mathrm{n}_j^\top \mathbf{l}_j = 0$ and $\|\mathrm{n}_j\| = 1$). Compute the new state vector $\mathbf{x}_n = [\check{\mathbf{r}}_n^\top, \check{\mathbf{v}}_n^\top]^\top$, where

$$\check{\mathbf{r}}_n = \mathbf{r}_{n-1} + \mathbf{s}' + \mathbf{s}'' - 2\mathrm{n}_j\mathrm{n}_j^\top\mathbf{s}'', \quad \check{\mathbf{v}}_n = \frac{\mathbf{s}'' - 2\mathrm{n}_j\mathrm{n}_j^\top\mathbf{s}''}{\|\mathbf{s}'' - 2\mathrm{n}_j\mathrm{n}_j^\top\mathbf{s}''\|}\|\tilde{\mathbf{v}}_n\|$$

6: **return** $f(\mathbf{x}_{n-1}) = \mathbf{x}_n$

used at time n to generate a sample $\mathbf{x}_n = [\mathbf{r}_n^\top, \mathbf{v}_n^\top]^\top$ from $\mathbf{x}_{n-1} = [\mathbf{r}_{n-1}^\top, \mathbf{v}_{n-1}^\top]^\top$. It accounts for the scenario in which the target hits one of the walls and deals with it by means of the law of reflection [48].

We are implicitly assuming that $\mathbf{r}_n \in \mathcal{R}$ in step 5 above. If this is not the case, i.e., $r_n \notin \mathcal{R}$, then steps 3–5 can be run again to implement a second reflection.

References

[1] M. Jansson, B. Wahlberg, A linear regression approach to state-space subspace system identification, Signal Process. 52 (2) (1996) 103–129.
[2] G. Storvik, Particle filters for state-space models with the presence of unknown static parameters, IEEE Trans. Signal Process. 50 (2) (2002) 281–289.
[3] C. Andrieu, A. Doucet, Online expectation-maximization type algorithms for parameter estimation in general state space models, in: 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), volume 6, IEEE, 2003, pp. VI–69.
[4] J. Ding, Y. Shi, H. Wang, F. Ding, A modified stochastic gradient based parameter estimation algorithm for dual-rate sampled-data systems, Digital Signal Process. 20 (4) (2010) 1238–1247.
[5] F. Ding, Y. Gu, Performance analysis of the auxiliary model-based stochastic gradient parameter estimation algorithm for state-space systems with one-step state delay, Circuits, Syst. Signal Process. 32 (2) (2013) 585–599.
[6] J. Kokkala, S. Särkkä, Combining particle MCMC with Rao-Blackwellized Monte Carlo data association for parameter estimation in multiple target tracking, Digital Signal Process. 47 (2015) 84–95.
[7] C. Andrieu, A. Doucet, R. Holenstein, Particle Markov chain Monte Carlo methods, J. R. Stat. Soc. B 72 (2010) 269–342.
[8] E. Koblents, J. Míguez, A population monte carlo scheme with transformed weights and its application to stochastic kinetic models, Stat. Comput. 25 (2) (2015) 407–425.
[9] D. Crisan, J. Miguez, Nested particle filters for online parameter estimation in discrete-time state-space Markov models. To appear in Bernoulli arXiv:1308.1883v3 [stat.CO].
[10] N. Kantas, A. Doucet, S.S. Singh, J.M. Maciejowski, N. Chopin, On particle methods for parameter estimation in state-space models, Stat. Sci. 30 (2015) 328–351.
[11] J. Olsson, T. Ryden, Rao-Blackwellization of particle Markov chain Monte Carlo methods using forward filtering backward sampling, in: IEEE Transactions on Signal Processing, volume 59, 2011, pp. 4606–4619.
[12] T. Vu, B.-N. Vo, R. Evans, A particle marginal metropolis-hastings multi-target tracker, IEEE Trans. Signal Process. 62 (15) (2014) 3953–3964.
[13] J. Kwon, R. Dragon, L.V. Gool, Joint tracking and ground plane estimation, IEEE Signal Process. Lett. 23 (11) (2016) 1514–1517.
[14] J. Ala-Luhtala, N. Whiteley, K. Heine, R. Piché, An introduction to twisted particle filters and parameter estimation in non-linear state-space models, IEEE Trans. Signal Process. 64 (18) (2016) 4875–4890.
[15] W.J. Fitzgerald, Markov chain Monte Carlo methods with applications to signal processing, Signal Process. 81 (1) (2001) 3–18.
[16] N. Gordon, D. Salmond, A.F.M. Smith, Novel approach to nonlinear and non–Gaussian Bayesian state estimation, IEE Proc.-F 140 (2) (1993) 107–113.
[17] A. Doucet, N. de Freitas, N. Gordon (Eds.), Sequential Monte Carlo Methods in Practice, Springer, New York (USA), 2001.
[18] A. Doucet, S. Godsill, C. Andrieu, On sequential monte carlo sampling methods for bayesian filtering, Stat. Comput. 10 (3) (2000) 197–208.
[19] P.M. Djurić, J.H. Kotecha, J. Zhang, Y. Huang, T. Ghirmai, M.F. Bugallo, J. Míguez, Particle filtering, IEEE Signal Process. Mag. 20 (5) (2003) 19–38.
[20] O. Cappé, S.J. Godsill, E. Moulines, An overview of existing methods and recent advances in sequential Monte Carlo, Proc. IEEE 95 (5) (2007) 899–924.
[21] C.P. Robert, G. Casella, Monte Carlo Statistical Methods, Springer, 2004.
[22] O. Cappé, A. Guillin, J.M. Marin, C.P. Robert, Population monte carlo, J. Comput. Graph. Stat. 13 (4) (2004) 907–929.
[23] N. Chopin, P.E. Jacob, O. Papaspiliopoulos, SMC2: an efficient algorithm for sequential analysis of state space models, J. R. Stat. Soc. (Statistical Methodology).
[24] R. Douc, A. Guillin, J.-M. Marin, C.P. Robert, Convergence of adaptive mixtures of importance sampling schemes, Annals Stat. (2007) 420–448.
[25] O. Cappé, R. Douc, A. Guillin, J.M. Marin, C.P. Robert, Adaptive importance sampling in general mixture classes, Stat. Comput. 18 (4) (2008) 447–459.
[26] J. Cornuet, J.-M. Marin, A. Mira, C.P. Robert, Adaptive multiple importance sampling, Scand. J. Stat. 39 (4) (2012) 798–812.
[27] M. Hong, M.F. Bugallo, P.M. Djuric, Joint model selection and parameter estimation by population Monte Carlo simulation, IEEE J. Selected Topics Signal Process. 4 (3) (2010) 526–539.
[28] L. Martino, V. Elvira, D. Luengo, J. Corander, An adaptive population importance sampler: learning from uncertainty, IEEE Trans. Signal Process. 63 (16) (2015) 4422–4437.
[29] M.F. Bugallo, L. Martino, J. Corander, Adaptive importance sampling in signal processing, Digital Signal Process. 47 (2015) 36–49.
[30] V. Elvira, L. Martino, D. Luengo, M.F. Bugallo, Improving population Monte Carlo: alternative weighting and resampling schemes, Signal Process. 131 (2017) 77–91.
[31] N. Chopin, A sequential particle filter method for static models, Biometrika 89 (3) (2002) 539–552.
[32] P. Del Moral, A. Doucet, A. Jasra, Sequential monte carlo samplers, J. R. Stat. Soc. 68 (3) (2006) 411–436.
[33] A. Kong, J.S. Liu, W.H. Wong, Sequential imputations and bayesian missing data problems, J. Am. Stat. Assoc. 9 (1994) 278–288.
[34] V. Elvira, L. Martino, D. Luengo, M.F. Bugallo, Efficient multiple importance sampling estimators, IEEE Signal Process. Lett. 22 (10) (2015) 1757–1761.
[35] B.D.O. Anderson, J.B. Moore, Optimal filtering, Englewood Cliffs, 1979.
[36] G. Kitagawa, Monte carlo filter and smoother for non-gaussian nonlinear state-space models, J. Comput. Graph. Stat. 1 (1996) 1–25.
[37] R. Douc, O. Cappé, E. Moulines, Comparison of resampling schemes for particle filtering, in: Proceedings of the 4th International Symposium on Image and Signal Processing and Analysis, 2005, pp. 64–69.
[38] A. Bain, D. Crisan, Fundamentals of Stochastic Filtering, Springer, 2008.
[39] P. Del Moral, Feynman–Kac Formulae: Genealogical and Interacting Particle Systems with Applications, Springer, 2004.
[40] J. Míguez, D. Crisan, P.M. Djurić, On the convergence of two sequential Monte Carlo methods for maximum a posteriori sequence estimation and stochastic global optimization, Stat. Comput. 23 (1) (2013) 91–107.
[41] M.G.B. Blum, Approximate Bayesian computation: a nonparametric perspective, J. Am. Stat. Assoc. 105 (491) (2010) 1178–1187.
[42] C. Andrieu, G. Roberts, The pseudo-marginal approach for efficient Monte Carlo computations, Annal. Stat. 37 (2009) 697–725.
[43] A. Doucet, N. de Freitas, N. Gordon, An introduction to sequential Monte Carlo methods, in: A. Doucet, N. de Freitas, N. Gordon (Eds.), Sequential Monte Carlo Methods in Practice, Springer, 2001, pp. 4–14. Ch. 1
[44] D. Crisan, J. Miguez, G. Ríos, A simple scheme for the parallelisation of particle filters and its application to the tracking of complex stochastic systems. ArXiv arXiv:1407.8071v2 [stat.CO].
[45] E. Koblents, Nonlinear population Monte Carlo methods for Bayesian inference, Universidad Carlos III de Madrid, 2015 Ph. d. dissertation.
[46] E. Koblents, J. Miguez, M.A. Rodriguez, A.M. Schmidt, A nonlinear population monte carlo scheme for the Bayesian estimation of parameters of α-stable distributions, Computational Statistics and Data Analysis 95 (2016) 57–74.
[47] D. Crisan, J. Miguez, Particle-kernel estimation of the filter density in state-space models, Bernoulli 20 (4) (2014) 1879–1929.
[48] G. Farin, D. Hansford, Practical Linear Algebra: A Geometry Toolbox, CRC Press, 2013.