ORIGINAL ARTICLE

Parallelising Particle Filters with Butterfly Interactions

KARI HEINE[1] | **NICK WHITELEY**[2] | **A. TAYLAN CEMGIL**[3]

[1]Department of Mathematical Sciences, University of Bath

[2]School of Mathematics, University of Bristol

[3]Department of Computer Engineering, Boğaziçi University

Correspondence
Kari Heine, Department of Mathematical Sciences, University of Bath, Bath, BA2 7AY, UK
Email: k.m.p.heine@bath.ac.uk

Funding information

Bootstrap particle filter (BPF) is the cornerstone of many algorithms used for solving generally intractable inference problems with Hidden Markov models. The long term stability of BPF arises from particle interactions that typically make parallel implementations of BPF nontrivial.

We propose a method whereby the particle interaction is done in several stages. With the proposed method, full interaction can be accomplished even if we allow only pairwise communications between processing elements at each stage. We show that our method preserves the consistency and the long term stability of the BPF, although our analysis suggest that the constraints on the stagewise interactions introduce error leading to a lower convergence rate than standard Monte Carlo. The proposed method also suggests a new, more flexible, adaptive resampling scheme, which according to our numerical experiments is the method of choice, displaying a notable gain in efficiency in certain parallel computing scenarios.

KEYWORDS
hidden Markov model, parallelism, particle filter, particle interaction, sequential Monte Carlo

1 | INTRODUCTION

In modern computing systems an increase in the computational power is primarily obtained by increasing the number of parallel processing elements (PE) rather than by increasing the speed (i.e. the clock rate) of an individual PE (e.g. Pacheco, 2011). While

in many cases such parallel systems have enabled the completion of increasingly complex computational tasks, they can only do so if the task in question admits parallel computations. In this paper we focus on an important class of algorithms lacking such inherent parallelism, namely the sequential Monte Carlo (SMC) methods, or particle filters (Gordon *et al.*, 1993; Doucet *et al.*, 2001).

It is well known (Lee *et al.*, 2010; Murray *et al.*, 2016) that the complications in parallelising SMC methods are due to the same key ingredient that also underpins their popularity: particle interactions, also commonly referred to as resampling. While these interactions stabilise the algorithms in time, and under certain assumptions, enable time uniform approximations (e.g. Del Moral and Guionnet, 2001; Douc *et al.*, 2014), they also imply that in an attempt to speed up the computations by distributing the particles across a number of PEs, we will inevitably introduce some *communication cost*. This cost arises from the need to communicate the particle information between PEs to enable the interaction. In this paper we propose new SMC algorithms that are based on an underlying principle of constraining the particle interactions in a structured way with the aim of reducing the communication cost. The resulting algorithms are studied both theoretically and in practice.

Our theoretical study involves analysing the convergence of the algorithms in the mean of order $r \geq 1$. More specifically, we obtain convergence rates in two specific scenarios:

a. m is fixed and $M \to \infty$,
b. $m \to \infty$ and M is fixed,

where m denotes the number of PEs and M denotes the number particles per PE. In the former case, the proposed algorithms retain the standard Monte Carlo rate $M^{-1/2}$ of convergence, while in the latter case a lower $(\log_2(m)/m)^{1/2}$ rate is obtained.

For the practical study, we compare some of the proposed algorithms empirically in a parallel computation context to a previously proposed SMC algorithm known as the *island particle filter* (IPF) (Vergé *et al.*, 2015) which we regard as the state of the art methodological approach to parallelising SMC. In this paper, we focus on methodology and hence further discussion on more implementation focused approaches, such as those discussed in (Murray *et al.*, 2016), is omitted.

Although the numerical experiments may leave some room for speculation on the optimality of the tested implementations, the proposed methods have two specific properties that can be used to introduce gain in performance as demonstrated by the experiments: they enable a more flexible adaptive resampling scheme — completely unique to the proposed approach — and they allow a straightforward way of reducing the cost of communicating the particle information between PEs.

1.1 | Particle filters and parallelising them

The well-known *bootstrap particle filter* (BPF), introduced by Gordon *et al.* (1993), first simulates an independent and identically distributed (i.i.d.) sample $\boldsymbol{\zeta}_0 := (\zeta_0^1, \cdots, \zeta_0^N)$ from a distribution π_0 defined on a sufficiently regular measurable state space $(\mathbb{X}, \mathcal{X})$. Then, for each $n > 0$, BPF subsequently generates samples $\boldsymbol{\zeta}_n := (\zeta_n^1, \cdots, \zeta_n^N)$ according to

$$\zeta_n^i \overset{\text{iid}}{\sim} \frac{\sum_{j=1}^N g(\zeta_{n-1}^j, y_{n-1}) f(\zeta_{n-1}^j, \cdot)}{\sum_{j=1}^N g(\zeta_{n-1}^j, y_{n-1})}, \qquad 1 \leq i \leq N,$$

where $f : (\mathbb{X}, \mathcal{X}) \to [0, 1]$ is a Markov kernel, and for all $x \in \mathbb{X}$ and some Markov kernel $G : (\mathbb{X}, \mathcal{Y}) \to [0, 1]$, the function $g(x, \cdot)$ is a density of $G(x, \cdot)$ w.r.t. some σ-finite measure on the measurable space $(\mathbb{Y}, \mathcal{Y})$. The samples $(\boldsymbol{\zeta}_n)_{n \geq 0}$ then define

empirical probability measures

$$\pi_n^N := \frac{1}{N}\sum_{i=1}^{N} \delta_{\zeta_n^i}, \qquad n \geq 0,$$

where δ_x denotes a point mass located at $x \in \mathbb{X}$. Many convergence results and central limit theorems exist for these measures, (e.g. Crisan and Doucet, 2002; Del Moral and Guionnet, 1999; Chopin, 2004; Del Moral, 2004), and it is well known that the limiting distribution of π_n^N is the prediction distribution

$$\pi_n(\cdot) := \mathbb{P}(X_n \in \cdot \mid Y_0 = y_0, \ldots, Y_{n-1} = y_{n-1}),$$

where $X := (X_n)_{n\geq 0}$ and $Y := (Y_n)_{n\geq 0}$ are the $\mathbb{X}$ valued signal process and $\mathbb{Y}$ valued observation process, respectively, of the hidden Markov model (HMM)

$$\begin{aligned} X_0 \sim \pi_0, \quad X_n \mid X_{n-1} = x_{n-1} \sim f(x_{n-1}, \cdot) \qquad & n \geq 1, \\ Y_n \mid X_n = x_n \sim g(x_n, \cdot) \qquad & n \geq 0. \end{aligned} \tag{1}$$

BPF can be summarised as shown in Algorithms 1 and 2, where we have also used the notations $\widehat{\boldsymbol{\zeta}}_n := (\widehat{\zeta}_n^1, \ldots, \widehat{\zeta}_n^N)$ and $g_n(\cdot) := g(\cdot, y_n)$ for all $n \geq 0$. We assume that g_n is a strictly positive, bounded and measurable function defined in $\mathbb{X}$. The final loop on lines 5 and 6 of Algorithm 1 we refer to as *the mutation step*.

It is also worth pointing out a notational convention that we will follow throughout the paper: the particles *within* resampling algorithms are denoted by the appropriately decorated Greek letter ξ and the particles *outside* the resampling algorithms are denoted similarly by ζ.

Algorithm 1 Particle filter

1: **for** $i = 1, \ldots, N$ **do**
2: $\quad \zeta_0^i \overset{\text{iid}}{\sim} \pi_0$
3: **for** $n \geq 0$ **do**
4: $\quad \widehat{\boldsymbol{\zeta}}_n \leftarrow \textsc{Resample}(\boldsymbol{\zeta}_n, g_n)$
5: $\quad$ **for** $i = 1, \ldots, N$ **do**
6: $\quad\quad \zeta_{n+1}^i \sim f(\widehat{\zeta}_n^i, \cdot)$

Algorithm 2 Multinomial resampling

1: $(\xi_1^i)_{1\leq i\leq N} = \textsc{Resample}\left((\xi_0^i)_{1\leq i\leq N}, g\right)$
2: **for** $i = 1, \ldots, N$ **do**
3: $\quad \xi_1^i \sim \sum_{j=1}^N g(\xi_0^j)\delta_{\xi_0^i} / \sum_{j=1}^N g(\xi_0^j)$

An obvious starting point for designing parallel SMC algorithms is to assign M particles to m PEs making the total sample size $N = mM$. Most of the calculations in Algorithms 1 and 2 can be done straightforwardly in parallel, except for line 3 in Algorithm 2 where ξ_1^i is generated as a duplicate of a random element of $(\xi_0^1, \ldots, \xi_0^N)$. Due to this step, PEs cannot proceed independently, but are required to exchange information about the particle coordinates ξ_0^i and their associated weights $g(\xi_0^i)$. In this paper we propose new ways of performing this interaction in order to harness the power of parallel computation for more

efficient particle filter algorithms.

One of the most important earlier contributions to the design of parallel SMC algorithms is (Bolić *et al.*, 2005) which introduced a modification of the BPF whereby the particle interactions are constrained by allowing the m PEs to exchange subsets of particles according specific local schemes. The theoretical properties of these popular local exchange particle filters (LEPF) was further investigated in (Míguez, 2007, 2014; Míguez and Vázquez, 2016; Heine and Whiteley, 2017). The analysis of (Míguez, 2014; Míguez and Vázquez, 2016) proved that under specific assumptions, the LEPF was uniformly convergent in time as $m \to \infty$, but interestingly, in addition to the central limit theorem for the LEPF, it was shown in (Heine and Whiteley, 2017) that under some regularity assumptions, LEPF cannot be uniformly convergent in mean of order $r \geq 1$ at rate $m^{-1/2}$. Whether the time uniform convergence holds at any slower rate remains an open question. Although the present paper does not address this question directly, it sheds some light on the matter as we show that particle interactions can indeed be constrained in a manner which preserves the time uniform convergence at a slower rate.

Whiteley *et al.* (2016) showed that in order to ensure time uniform convergence of particle filters, full interaction may not be needed at each iteration but it is sufficient to have enough interaction to ensure that the *effective sample size* remains above some predetermined threshold. Although their analysis techniques are similar to those used in this paper, the methods presented here cannot be regarded as instances of the αSMC framework of (Whiteley *et al.*, 2016). Both αSMC and LEPF, of which the latter can be regarded as an instance of αSMC as discussed in (Heine and Whiteley, 2017), constrain the interactions to subsets of particles. Here we propose methods that allow interactions between all particles, but in a constrained manner. The way of constraining the interactions presented in this paper is completely novel which is also manifested by the unprecedented slower than standard Monte Carlo convergence rate.

A more recent development towards parallelising particle filters is the island particle filter (IPF) proposed by Vergé *et al.* (2015). IPF is based on a two stage implementation of the resampling step. At the first stage one resamples the particle islands, or PEs, to duplicate and redistribute the PE specific particle sets according to some, e.g. multinomial, resampling scheme without considering particles individually. At the second stage, each PE then performs particle level resampling independent of each other. Del Moral *et al.* (2017) provides proofs of convergence in probability, central limit theorem and large deviations for the IPF algorithm. The methods we propose in the present work are reminiscent to IPF and can be thought of as a result of combining IPF with concepts originating from computer network topologies.

1.2 | Augmented resampling

The particle filter algorithms presented in this paper are all based on a novel *augmented resampling* algorithm which is a multi-stage resampling algorithm parametrised by two positive integers N and S that are assumed to satisfy:

Assumption A1 *$N, S \in \{1, 2, \dots\}$ are such that $N = mM$ and $S = \log_2(m)$ for some $m, M \in \{1, 2, \dots\}$.*

We retain the interpretation of m being the number of PEs, M the number of particles per PE, and N being the total number of particles. The parameter S is specific to the augmented resampling algorithm and it denotes the number of resampling stages. For given matrices $A_1, \dots, A_S \in \mathbb{R}^{N \times N}$, to be specified later, augmented resampling proceeds as described in Algorithm 3.

A key characteristic of augmented resampling is that by means of the matrices $A_1, \dots, A_S$, we can control which PEs are allowed to interact at each stage $1 \leq s \leq S$. While our theory allows for defining these matrices in various ways, we will only focus on a specific definition which implies *pairwise* interactions between PEs at each stage $1 \leq s \leq S$. Formally

$$A_s := \mathbf{I}_{2^{S-s}} \otimes \mathbf{1}_{1/2} \otimes \mathbf{I}_{2^{s-1}} \otimes \mathbf{1}_{1/M}, \tag{2}$$

where $\otimes$ denotes the Kronecker product and for any $k > 0$, $\mathbf{I}_k$ is size k identity matrix, and the abusive notation $\mathbf{1}_{1/k}$ is used for

Algorithm 3 Augmented resampling

1: $(\xi_S^i)_{1\leq i\leq N} = \text{AUGMENTEDRESAMPLE}\left((\xi_0^i)_{1\leq i\leq N}, g\right)$
2: **for** $i = 1, \dots, N$ **do**
3: $\quad V_0^i \leftarrow g(\xi_0^i)$
4: **for** $s = 1, \dots, S$ **do**
5: $\quad$ **for** $i = 1, \dots, N$ **do**
6: $\quad\quad V_s^i \leftarrow \sum_{j=1}^N A_s^{ij} V_{s-1}^j$
7: $\quad\quad \xi_s^i \sim (V_s^i)^{-1} \sum_{j=1}^N A_s^{ij} V_{s-1}^j \delta_{\xi_{s-1}^j}$

FIGURE 1 HERE

a size k matrix of ones multiplied by $1/k$.

Figure 1 illustrates the matrices $A_1, \dots, A_S$ and how they determine the pairs of interacting PEs at different stages. Each node in the graph represents an individual particle at a specific stage. An edge between ξ_{s-1}^i and ξ_s^j is equivalent to $A_s^{ij} \neq 0$ and hence the particle ξ_s^i is sampled with replacement among the particles ξ_{s-1}^j where j is such that $A_s^{ij} \neq 0$. At stage s the PE containing particle ξ_s^i is thus required to communicate only with the PE containing the elements of $(\xi_{s-1}^1, \dots, \xi_{s-1}^N)$ connected to ξ_s^i and, as demonstrated in Figure 1, only pairwise interactions between PEs are required; at the first stage the interacting pairs are (PE1,PE2) and (PE3,PE4) and at the second stage (PE1,PE3) and (PE2,PE4). This *radix-2 butterfly diagram* structure of Figure 1 is perhaps better known from the context of Cooley-Tuckey fast Fourier transformation (e.g. Oppenheim, 1975), and a formal definition of this structure will be given Section 2.

There are two motivations for our interest in studying augmented resampling in the particle filtering context. The first is related to the communication pattern between PEs and the second is related to adaptive resampling schemes.

Regarding the communication pattern, let us assume an idealised computer architecture in which a PE can communicate with at most one other PE at a time and different pairs of PEs can communicate perfectly in parallel. Moreover, we assume that the time required to perform the communication is constant over the pairs of PEs. We acknowledge that in reality these assumptions are only approximate as computer architectures involve various types of PEs (e.g. networks of computers or cores within processors) interconnected by various network topologies (e.g. hypercubes or data buses).

Now suppose that there are four PEs (PE1, PE2, PE3, and PE4) and only the sample contained in a single PE, say PE1, has an effectively non-zero weight. In this case, without augmented resampling, PE1 would have to disseminate its sample to all other PEs; first to PE2, then to PE3 and finally to PE4. This suggests that $m-1$ sequential communication steps are needed. With augmented resampling, as in Figure 1, PE1 would first send its sample to PE2, after which PE1 would send the sample to PE3 while *at the same* time PE2 could send its sample (just received from PE1) to PE4 thereby accomplishing complete dissemination of PE1's sample in only $\log_2(m)$ sequential communication steps, making augmented resampling apparently more efficient in terms of communication. It should be made clear that we have presented the above reasoning under the assumption of an idealised computer architecture simply to motivate our interest in the augmented resampling and not to argue for its superiority. We will base our final conclusions on the numerical experiments.

The second motivation for augmented resampling is its additional flexibility in adaptive resampling schemes (Liu and Chen, 1998). It is well known that although resampling is the enabling factor for long term stability of SMC methods, it does introduce error as well as additional computational cost and should only be done when necessary. In adaptive resampling, prior to performing the actual resampling, one first evaluates the effective sample size (ESS) (Liu and Chen, 1995) which for the BPF

can be formally expressed as

$$\mathcal{E}_n = \frac{\left(N^{-1}\sum_{i=1}^N g_n(\zeta_n^i)\right)^2}{N^{-1}\sum_{i=1}^N g_n^2(\zeta_n^i)} \in \left[\frac{1}{N}, 1\right],$$

and executes the resampling step only if $\mathcal{E}_n$ is below some predetermined threshold $\theta \in (1/N, 1]$. This means that every filter iteration with adaptive resampling involves a dichotomous decision to either allow the full interaction of all particles or allow no interaction at all. Augmented resampling enables this decision to be refined so that the decision is made between finer levels of interaction, and hence it may be possible to find a better balance between long term stability, resampling error, and computational cost. In practice this is accomplished by evaluating the ESS after every stage of augmented resampling and, based on the ESS, deciding whether to proceed to the next resampling stage or to skip the remaining resampling stages and move on to the next time step.

This more flexible adaptation of resampling is based on the ideas presented in (Whiteley *et al.*, 2016) and it will lead to an increase in efficiency if sufficiently few resampling stages in total are executed. It is also worth noting that the evaluation of the ESS at every stage introduces some additional communication, but our numerical experiments suggest that the net effect of this *fully adapted* resampling scheme is a notable gain in efficiency. The rigorous theoretical analysis of the convergence properties of this method is left beyond the scope of this paper.

This paper is organised as follows. Section 2 is dedicated to the theoretical properties of augmented resampling outside the particle filtering context and it presents our main convergence result for augmented resampling, namely Proposition 1. In Section 3 we apply the augmented resampling algorithm in the particle filter context and present our main convergence result, Theorem 1. Section 4 introduces a modified augmented resampling scheme reminiscent to that used in IPF and the convergence of the resulting particle filter is proved in Section 5. Section 6 concludes the paper with some results of numerical experiments showing the potential of the proposed algorithms and a brief discussion on the conclusions. Most of the more technical proofs are housed in the appendices.

1.3 | Notations

We let $\mathcal{B}(\mathbb{X})$ denote the bounded and measurable $\mathbb{R}$ valued functions defined on $(\mathbb{X}, \mathcal{X})$. Throughout the paper, we define $\|\varphi\| := \sup_{x\in\mathbb{X}} |\varphi(x)|$ and $\operatorname{osc}(\varphi) := \sup_{x,y\in\mathbb{X}} |\varphi(x) - \varphi(y)|$ for any $\varphi \in \mathcal{B}(\mathbb{X})$. We define two specific subsets of $\mathcal{B}(\mathbb{X})$, $\mathcal{B}_+(\mathbb{X}) := \{\varphi \in \mathcal{B}(\mathbb{X}) : \varphi > 0\}$ and $\mathcal{B}_1(\mathbb{X}) := \{\varphi \in \mathcal{B}(\mathbb{X}) : \|\varphi\| \leq 1\}$. For a sequence of square matrices $(A_s)_{1\leq s\leq S}$ where $S \in \mathbb{N}_+$ we write $\prod_{s=1}^S A_s = A_S \cdots A_1$. For any $N, S \in \{1, 2, \dots\}$ we use the shorthand notation $\sum_{(i_0,\dots,i_S)} := \sum_{i_0=1}^N \cdots \sum_{i_S=1}^N$. We also define $\lceil x \rceil := \min\{z \in \mathbb{Z} : z \geq x\}$ and $\lfloor x \rfloor := \max\{z \in \mathbb{Z} : z \leq x\}$ and $(x \bmod z) := x - z\lfloor x/z \rfloor$. Throughout the remainder of this paper $\mathbb{E}$ and $\mathbb{P}$ refer to the expectation and probability with respect to the probability space charactering the randomness of the algorithm only. The observations of the underlying HMM are assumed fixed.

2 | AUGMENTED RESAMPLING

We start with a study of Algorithm 3 outside the filtering context by applying it to an arbitrary $\mathbb{X}^N$ valued random sample $\boldsymbol{\xi}_0 = (\xi_0^1, \dots, \xi_0^N)$ and an arbitrary weighting function $g \in \mathcal{B}_+(\mathbb{X})$. We have the following result:

Proposition 1 *Assume* (A1) *and let* $g \in \mathcal{B}_+(\mathbb{X})$. *Then for any* $\mathbb{X}^N$ *valued random variable* $\boldsymbol{\xi}_0$ *and for any* $\varphi \in \mathcal{B}(\mathbb{X})$,

$$\mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}\varphi(\xi_S^i)\middle|\boldsymbol{\xi}_0\right] = \frac{\sum_{i=1}^{N} g(\xi_0^i)\varphi(\xi_0^i)}{\sum_{i=1}^{N} g(\xi_0^i)}, \tag{3}$$

and for any $r \geq 1$ *there exists a finite constant* B_r, *depending only on* r, *such that no matter what the distribution of* $\boldsymbol{\xi}_0$ *is, we have*

$$\mathbb{E}\left[\left|\left(\frac{1}{N}\sum_{i=1}^{N} g(\xi_0^i)\right)\left(\frac{1}{N}\sum_{i=1}^{N}\varphi(\xi_S^i)\right) - \frac{1}{N}\sum_{i=1}^{N} g(\xi_0^i)\varphi(\xi_0^i)\right|^r\right]^{\frac{1}{r}} \leq B_r\sqrt{\frac{S}{N}}\,\|g\|\,\mathrm{osc}\,(\varphi)\,. \tag{4}$$

An immediate consequence of Proposition 1 is that if, for example, S is some non-decreasing function $S(N)$ of N such that $\sum_{N=1}^{\infty}(S(N)/N)^{r/2} < \infty$ for some $r \geq 1$, then

$$\left(\frac{1}{N}\sum_{i=1}^{N} g(\xi_0^i)\right)\left(\frac{1}{N}\sum_{i=1}^{N}\varphi(\xi_S^i) - \frac{\sum_{i=1}^{N} g(\xi_0^i)\varphi(\xi_0^i)}{\sum_{i=1}^{N} g(\xi_0^i)}\right) \xrightarrow[N\to\infty]{\text{a.s.}} 0, \tag{5}$$

without requiring any convergence of $N^{-1}\sum_{i=1}^{N} g(\xi_0^i)$ or $N^{-1}\sum_{i=1}^{N} g(\xi_0^i)\varphi(\xi_0^i)$.

The more technical proofs of the results stated in this section are housed in Appendix A.

2.1 | Properties of augmented resampling

The matrices $A_1, \ldots, A_S$ play an important role in augmented resampling and to a large extent they determine its statistical properties. We present first the following result which, although not in its entirety required to prove Proposition 1, summarises some key properties of $A_1, \ldots, A_S$ and also makes it formally explicit, how the structure of the diagram in Figure 1 is obtained.

Lemma 1 *Assume* (A1). *Then for all* $1 \leq s \leq S$, A_s *is symmetric, idempotent, and doubly stochastic. Moreover, for any* $1 \leq i \leq m$

$$\left\{j \in \{1,\ldots,m\} : (\mathbf{I}_{2^{S-s}} \otimes \mathbf{1}_{1/2} \otimes \mathbf{I}_{2^{s-1}})^{ij} \neq 0\right\} = \left\{\left((i-1) \bmod 2^{s-1}\right) + (q-1)2^{s-1} + 2^s\left\lfloor\frac{i-1}{2^s}\right\rfloor + 1 : q \in \{1,2\}\right\}, \tag{6}$$

and for all $1 \leq i \leq m$, $(\mathbf{I}_{2^{S-s}} \otimes \mathbf{1}_{1/2} \otimes \mathbf{I}_{2^{s-1}})^{ii} = 1/2$.

Equation (6) formalises the radix-2 butterfly structure seen in Figure 1 by giving explicit expression for the nonzero elements of $\mathbf{I}_{2^{S-s}} \otimes \mathbf{1}_{1/2} \otimes \mathbf{I}_{2^{s-1}} \in \mathbb{R}^{m\times m}$. By considering $A_s \in \mathbb{R}^{mM\times mM}$ as an m-by-m matrix of M-by-M blocks, the element (i,j) of $\mathbf{I}_{2^{S-s}} \otimes \mathbf{1}_{1/2} \otimes \mathbf{I}_{2^{s-1}}$ is nonzero if and only if the block (i,j) of A_s is the full matrix $\frac{1}{2}\mathbf{1}_{1/M}$.

From Algorithm 3 we obtain the definitions

$$V_0^i := g(\xi_0^i), \qquad V_s^i := \sum_{j=1}^{N} A_s^{ij} V_{s-1}^j, \qquad 1 \leq i \leq N,\ 1 \leq s \leq S \tag{7}$$

for the particle weights at each stage. For the proof of Proposition 1 it is crucial that after finishing all S resampling stages, Algorithm 3 returns an *unweighted* sample in a manner similar to conventional multinomial resampling, i.e. that $V_S^i = V_S^j$ for all $i, j \in \{1, \ldots, N\}$. The proof of this unweighted property is essentially due to the following key result on $A_1, \ldots, A_S$.

Lemma 2 *Assume* (A1). *Then* $\prod_{s=1}^{S} A_s = \mathbf{1}_{1/N}$.

Lemma 2 enables us to establish the following result which, in addition to the unweighted property, states some other facts about the weights V_s^i that are required for the proof of Proposition 1.

Lemma 3 *Assume* (A1). *For any* $1 \leq i \leq N$ *and* $0 \leq s \leq S$,

a. V_s^i *is measurable w.r.t.* $\sigma(\xi_0)$,
b. $V_s^i \leq \|g\|$.
c. $V_S^i = N^{-1} \sum_{j=1}^{N} g(\xi_0^j)$.

Remark 1 *Although we work throughout the paper with the definition* (2) *of* A_s, *the specific definition of* $(A_s)_{1\leq s\leq S}$ *is irrelevant for the proof of Proposition 1 as long as the matrices satisfy Lemma 2 and are doubly stochastic. Different definitions of* $(A_s)_{1\leq s\leq S}$ *for which Lemma 2 still holds can be easily devised (see the unpublished work (Heine* et al.*, 2014) for alternative definitions). Above, the double stochasticity follows from Lemma 1.*

Intuitively speaking, Lemma 2 can be understood as a property of $(A_s)_{1\leq s\leq S}$ *which ensures that full, although constrained, interaction between all particles can be achieved. This is pivotal to the time uniform convergence and indeed, Heine and Whiteley (2017) discuss an interaction framework which lacks such a guarantee for full interaction and consequently the resulting algorithms also lack the time uniform convergence properties typically associated with SMC algorithms.*

2.2 | Proof of Proposition

The proof of Proposition 1 is based on expressing the error term on the left hand side of (5) as a martingale to which we then apply the Burkholder inequality. For the required martingale construction, we observe another important property of Algorithm 3; for all $1 \leq s \leq S$ the random samples $\boldsymbol{\xi}_s := (\xi_s^1, \ldots, \xi_s^N)$ satisfy a *one step conditional independence* property, i.e. the particles $\xi_s^1, \ldots, \xi_s^N$ are conditionally independent given $\boldsymbol{\xi}_0, \ldots, \boldsymbol{\xi}_{s-1}$. We also see that for each $1 \leq i \leq N$ and $B \in \mathcal{X}$, we have

$$\mathbb{P}\left(\xi_s^i \in B \,\middle|\, \boldsymbol{\xi}_0, \ldots, \boldsymbol{\xi}_{s-1}\right) = \frac{1}{V_s^i} \sum_{j=1}^{N} A_s^{ij} V_{s-1}^j \mathbb{I}_B(\xi_{s-1}^j), \tag{8}$$

where $\mathbb{I}_B$ denotes the indicator function of the set $B \in \mathcal{X}$.

To construct the required martingale via a martingale difference, we define a sequence $\mathcal{M} := \{(X_\rho, \mathcal{F}_\rho);\, 0 \leq \rho \leq SN\}$ where $X_0 := 0$, $\mathcal{F}_0 := \sigma(\boldsymbol{\xi}_0)$ and for all $0 < \rho \leq SN$, we define

$$X_\rho := \frac{V_{s_N(\rho)}^{i_N(\rho)}}{\sqrt{SN}} \left(\overline{\varphi}(\xi_{s_N(\rho)}^{i_N(\rho)}) - \frac{1}{V_{s_N(\rho)}^{i_N(\rho)}} \sum_{j=1}^{N} A_{s_N(\rho)}^{i_N(\rho)j} V_{s_N(\rho)-1}^{j} \overline{\varphi}(\xi_{s_N(\rho)-1}^{j}) \right), \qquad \mathcal{F}_\rho := \mathcal{F}_{\rho-1} \vee \sigma(\xi_{s_N(\rho)}^{i_N(\rho)})$$

where

$$\overline{\varphi} := \varphi - \frac{\sum_{i=1}^{N} g(\xi_0^i)\varphi(\xi_0^i)}{\sum_{i=1}^{N} g(\xi_0^i)}, \qquad \varphi \in \mathcal{B}(\mathbb{X}),$$

and for any $k \in \mathbb{N}$

$$i_k(\rho) := ((\rho - 1) \bmod k) + 1 \qquad s_k(\rho) := \left\lceil \frac{\rho}{k} \right\rceil. \tag{9}$$

The purpose of (9) is simply to define a bijective index map taking a one dimensional index ρ in the range $\{1, \dots, Sk\}$ into a pair of indices $s_k(\rho) \in \{1, \dots, S\}$ and $i_k(\rho) \in \{1, \dots, k\}$. The following proposition establishes the required martingale properties of $\mathcal{M}$.

Proposition 2 *Assume* (A1). *The following statements hold:*

a. X_ρ *is* $\mathcal{F}_\rho$*-measurable for all* $0 \le \rho \le SN$*;*
b. $\mathbb{E}\left[X_\rho \mid \mathcal{F}_{\rho-1}\right] = 0$ *(a.s.) for all* $0 < \rho \le SN$*;*
c. $\left|X_\rho\right| \le \|g\| \operatorname{osc}(\varphi) / \sqrt{SN}$ *for all* $0 \le \rho \le SN$*;*
d. *and we have the identities*

$$\sqrt{\frac{S}{N}} \sum_{\rho=1}^{SN} X_\rho = \frac{1}{N} \sum_{i=1}^{N} V_S^i \overline{\varphi}(\xi_S^i) \tag{10}$$

$$= \left(\frac{1}{N} \sum_{i=1}^{N} g(\xi_0^i) \right) \left(\frac{1}{N} \sum_{i=1}^{N} \varphi(\xi_S^i) \right) - \frac{1}{N} \sum_{i=1}^{N} g(\xi_0^i) \varphi(\xi_0^i) \tag{11}$$

By Proposition 2 the proof of Proposition 1 is obtained readily as follows.

Proof of Proposition 1 The lack of bias (3) follows by Proposition 2(b), (10), (11), and the tower property of conditional expectations. Bound (4) follows by Burkholder-Davis-Gundy inequality and Proposition 2(c) by writing

$$\mathbb{E}\left[\left|\sum_{\rho=1}^{SN} X_\rho\right|^r\right] \le B_r^r \mathbb{E}\left[\left|\sqrt{\sum_{\rho=1}^{SN} \left|X_\rho\right|^2}\right|^r\right] \le B_r^r \|g\|^r \operatorname{osc}(\varphi)^r .$$

3 | PARTICLE FILTER WITH AUGMENTED RESAMPLING

We now turn to analysing the implications of replacing Algorithm 2 in BPF with Algorithm 3. The following mild regularity condition on the underlying HMM is assumed to hold.

Assumption A2 *For all* $n \ge 0$, $g_n \in \mathcal{B}_+(\mathbb{X})$.

Under (A2), we show that the resulting particle filter is convergent in mean (of order $r \ge 1$). In order to establish uniform in time convergence in mean, the following strong but standard regularity assumption is made (Whiteley *et al.*, 2016; Del Moral, 2004).

Assumption A3 *There exists* $\delta \ge 1$ *and* $\epsilon \in (0, 1)$ *such that*

$$\sup_{n \ge 0} \sup_{x,y} \frac{g_n(x)}{g_n(y)} \le \delta, \qquad \text{and} \qquad f(x, \cdot) \ge \epsilon f(y, \cdot), \ \forall x, y \in \mathbb{X}^2.$$

Theorem 1 *Fix* N *and* S *and assume* (A1) *and* (A2). *If the measures* $(\pi_n^N)_{n \ge 0}$ *are calculated by Algorithm 1 deploying Algorithm 3, then we have the following:*

a. *For all $n \geq 0$ and $r \geq 1$, there exists $C_{n,r} \in \mathbb{R}_+$ such that*

$$\sup_{\varphi \in \mathcal{B}_1(\mathbb{X})} \mathbb{E}\left[\left|\pi_n^N(\varphi) - \pi_n(\varphi)\right|^r\right]^{\frac{1}{r}} \leq C_{n,r}\sqrt{\frac{S}{N}}. \tag{12}$$

b. *If in addition* (A3) *holds, then for all $r \geq 1$ there exists $C_r \in \mathbb{R}_+$ such that*

$$\sup_{n \geq 0} \sup_{\varphi \in \mathcal{B}_1(\mathbb{X})} \mathbb{E}\left[\left|\pi_n^N(\varphi) - \pi_n(\varphi)\right|^r\right]^{\frac{1}{r}} \leq C_r\sqrt{\frac{S}{N}}.$$

Although Theorem 1 resembles many existing results on BPF, and its variations, the interpretation is somewhat different. The result is stated under the assumption (A1) which leaves the convergence rate ambiguous. However, if we write the r.h.s. of (12) in terms of m and M, we observe that by fixing m, (12) yields the standard $M^{-1/2}$ rate of convergence, and by fixing M, a slower $\sqrt{\log_2(m)/m}$ rate is obtained. The rate is slower due to the numerator term $\sqrt{S} = \sqrt{\log_2(m)}$ which can be intuitively interpreted to trace back to the resampling errors introduced at each stage of augmented resampling. In both cases, by Borel-Cantelli argument, Theorem 1 also yields the law of large numbers, i.e. that $\pi_n^{mM}(\varphi) - \pi_n(\varphi) \to 0$ almost surely as $m \to \infty$ (resp. $M \to \infty$) and M (resp. m) is kept fixed.

While the convergence rate $M^{-1/2}$ that we obtain for fixed m is known to be optimal, the analysis that we carry out to prove Theorem 1 does not explicitly imply that also the $\sqrt{\log_2(m)/m}$ rate, obtained for fixed M, is optimal, and not an artefact of our analysis. However, we conjecture this to be the case. First, the term $\sqrt{S/N}$, arises quite naturally and inevitably from the martingale construction of Proposition 2 and second, the unpublished work (Heine *et al.*, 2014) proves a CLT for a similar, but not identical, algorithm with the same smaller scaling factor. In (Heine *et al.*, 2014) the slower convergence rate also arises from the radix-2 butterfly structure. Last, we refer the reader to Section 6.3 where we also provide some numerical evidence to support this conjecture.

In the following subsections we go through the steps of proving Theorem 1. The more technical proofs are postponed to Appendix B.

3.1 | Preliminary results

The proof of Theorem 1(a) is by induction. The following lemma, whose primary purpose is to initialise the induction, is a special instance of the more general result proved in (Del Moral, 2004, Lemma 7.3.3) and hence we omit the proof.

Lemma 4 *Let $(\zeta^1, \ldots, \zeta^N)$ be an i.i.d. sample from some distribution π defined on $(\mathbb{X}, \mathcal{X})$. Then there exists a constant $C_r^* \in \mathbb{R}_+$ depending only on r such that*

$$\sup_{\varphi \in \mathcal{B}_1(\mathbb{X})} \mathbb{E}\left[\left|\frac{1}{N}\sum_{i=1}^N \varphi(\zeta^i) - \pi(\varphi)\right|^r\right]^{\frac{1}{r}} \leq \frac{C_r^*}{\sqrt{N}}.$$

We also frequently use the following result to bound the error introduced by the mutation step of the particle filter.

Lemma 5 *Let $(\widehat{\zeta}^1, \ldots, \widehat{\zeta}^N)$ be a $\mathbb{X}^N$ valued random variable and let*

$$(\zeta^1, \ldots, \zeta^N) \mid (\widehat{\zeta}^1, \ldots, \widehat{\zeta}^N) \sim \prod_{i=1}^N f(\widehat{\zeta}^i, \cdot). \tag{13}$$

Then there exists a constant $B_r \in \mathbb{R}_+$ *such that for all* $N > 0$

$$\sup_{\varphi \in \mathcal{B}_1(\mathbb{X})} \mathbb{E}\left[\left|\frac{1}{N}\sum_{i=1}^{N}\varphi(\zeta^i) - \frac{1}{N}\sum_{i=1}^{N} f(\varphi)(\widehat{\zeta}^i)\right|^r\right]^{\frac{1}{r}} \leq \frac{2B_r}{\sqrt{N}}.$$

Instead of a proof by induction, the proof of Theorem 1(b) is based on the proof of Theorem 7.4.4 in (Del Moral, 2004). For any probability measure μ on $(\mathbb{X}, \mathcal{X})$ and any $\varphi \in \mathcal{B}(\mathbb{X})$, we define

$$\Phi_0(\pi_{-1}^N) := \pi_0, \qquad \text{and} \qquad \Phi_n(\mu)(\varphi) := \frac{\mu(g_{n-1} f(\varphi))}{\mu(g_{n-1})}, \qquad n > 0.$$

We note that Φ_n is the mapping which generates the sequence of exact measures $(\pi_n)_{n\geq 0}$ by the recursion

$$\pi_n = \Phi_n(\pi_{n-1}), \qquad n \geq 0.$$

By using these notations, we have the following corollary of the proof of Theorem 7.4.4 in (Del Moral, 2004).

Lemma 6 *Assume* (A3). *Then for all* $0 \leq n$, $0 \leq p \leq n$ *and* $\varphi \in \mathcal{B}_1(X)$, *there exists* $\alpha_{p,n} \in \mathbb{R}_+$ *and* $\varphi_{p,n,\varphi} \in \mathcal{B}_1(\mathbb{X})$ *such that*

$$\left|\pi_n^N(\varphi) - \pi_n(\varphi)\right| \leq \sum_{p=0}^{n} \alpha_{p,n}\left|\left(\pi_p^N - \Phi_p(\pi_{p-1}^N)\right)(\varphi_{p,n,\varphi})\right|$$

and $\sum_{p=0}^{n} \alpha_{p,n} \leq \delta/\epsilon^3$.

3.2 | Convergence

Before the proof of Theorem 1, we introduce an intermediate result, Proposition 3 below, consisting of two parts. The first part establishes the induction step needed for the proof of Theorem 1(a). The second part is used in the proof of Theorem 1(b) and it establishes a uniform bound for the local error terms $\pi_n^N - \Phi_n(\pi_{n-1}^N)$ appearing in Lemma 6. For the brevity of notation we introduce the following probability measures

$$\widehat{\pi}_n^N := \frac{1}{N}\sum_{i=1}^{N} \delta_{\widehat{\zeta}_n^i}, \qquad \widehat{\pi}_n(\mathrm{d}x) := \frac{g_n(x)\pi_n(\mathrm{d}x)}{\pi_n(g_n)}, \qquad n \geq 0, \tag{14}$$

where $\widehat{\pi}_n$ is the exact filtering distribution associated with the HMM (1).

Proposition 3 *Assume* (A1) *and* (A2).

a. *If for some* $n \geq 0$ *and some* $r \geq 1$ *there exists* $C_{n,r} \in \mathbb{R}_+$ *such that*

$$\sup_{\varphi \in \mathcal{B}_1(\mathbb{X})} \mathbb{E}\left[\left|\pi_n^N(\varphi) - \pi_n(\varphi)\right|^r\right]^{\frac{1}{r}} \leq C_{n,r}\sqrt{\frac{S}{N}}, \tag{15}$$

then there also exists $\widehat{C}_{n,r} \in \mathbb{R}$ *such that*

$$\sup_{\varphi \in \mathcal{B}_1(\mathbb{X})} \mathbb{E}\left[\left|\widehat{\pi}_n^N(\varphi) - \widehat{\pi}_n(\varphi)\right|^r\right]^{\frac{1}{r}} \leq \widehat{C}_{n,r}\sqrt{\frac{S}{N}}.$$

b. *If in addition* (A3) *holds, then for all* $r \geq 1$ *there exists* $\widehat{C}_r \in \mathbb{R}_+$ *such that*

$$\sup_{n\geq 0}\sup_{\varphi \in \mathcal{B}_1(\mathbb{X})} \mathbb{E}\left[\left|\pi_n^N(\varphi) - \Phi_n(\pi_{n-1}^N)(\varphi)\right|^r\right]^{\frac{1}{r}} \leq \widehat{C}_r\sqrt{\frac{S}{N}}.$$

Part a) introduces the precondition (15) to bound the local error which effectively leads to the proof of Theorem 1(a) being by induction. Under the assumption (A3) in part b) such condition is not needed and the analysis becomes somewhat simpler.

Proof of Theorem 1 Fix $r \geq 1$. The proof of part a) is by induction in $n \geq 0$. The induction is initialised by observing that at rank $n = 0$, (12) holds by Lemma 4. Suppose now that (12) holds at some rank $n \geq 0$. By Minkowski's inequality, and the fact that $\pi_{n+1}(\varphi) = \widehat{\pi}_n(f(\varphi))$, we have

$$\sup_{\varphi \in \mathcal{B}_1(\mathbb{X})} \mathbb{E}\left[\left|\pi_{n+1}^N(\varphi) - \pi_{n+1}(\varphi)\right|^r\right]^{\frac{1}{r}} \leq \sup_{\varphi \in \mathcal{B}_1(\mathbb{X})} \mathbb{E}\left[\left|\pi_{n+1}^N(\varphi) - \widehat{\pi}_n^N(f(\varphi))\right|^r\right]^{\frac{1}{r}} + \sup_{\varphi \in \mathcal{B}_1(\mathbb{X})} \mathbb{E}\left[\left|\widehat{\pi}_n^N(f(\varphi)) - \widehat{\pi}_n(f(\varphi))\right|^r\right]^{\frac{1}{r}}.$$

By applying Lemma 5 and Proposition 3, respectively, to the first and the second term on the r.h.s., we obtain the bound

$$\sup_{\varphi \in \mathcal{B}_1(\mathbb{X})} \mathbb{E}\left[\left|\pi_{n+1}^N(\varphi) - \pi_{n+1}(\varphi)\right|^r\right]^{\frac{1}{r}} \leq 2B_r\sqrt{\frac{S}{N}} + \widehat{C}_{n,r}\sqrt{\frac{S}{N}},$$

and thus (12) holds at rank $n+1$ with $C_{n+1,r} = 2B_r + \widehat{C}_{n,r}$.

For part b) we have by Lemma 6

$$\mathbb{E}[|\pi_n^N(\varphi) - \pi_n(\varphi)|^r]^{\frac{1}{r}} \leq \sum_{p=0}^{n} \alpha_{p,n}\mathbb{E}\left[\left|\left(\pi_p^N - \Phi_p(\pi_{p-1}^N)\right)\left(\varphi_{p,n,\varphi}\right)\right|^r\right]^{\frac{1}{r}} \leq \widehat{C}_r\frac{\delta}{\epsilon^3}\sqrt{\frac{S}{N}}, \tag{16}$$

where the second inequality follows from Proposition 3(b) and Lemma 6.

4 | AUGMENTED RESAMPLING FOR PARTICLE ISLANDS

So far we have seen that by replacing the multinomial resampling (Algorithm 2) in the BPF with the augmented resampling (Algorithm 3), we obtain a convergent approximation of $(\pi_n)_{n\geq 0}$. However, the proposed algorithm has some shortcomings in efficiency which will address in this section.

First, we observe that at each stage of Algorithm 3, each PE resamples M particles out of $2M$ particles, which is in general more computationally expensive than resampling M out of M particles. Second, we observe that in order to do the resampling, a PE must receive the M individual particle weights from the paired PE, which may imply a notable communication cost, especially for large M. In this section we propose a modification which addresses both of these sources of computation and communication cost; in the proposed method each PE resamples M particles out of M particles and communicates only a single weight with its paired PE at each stage.

The proposed modification is reminiscent to the IPF algorithm of Vergé *et al.* (2015) with the exception that the between

island (i.e. between PE) resampling is done in multiple stages by means of augmented resampling. We dub the algorithm *augmented island resampling particle filter (AIRPF)* and it is described in Algorithm 4 below, where we also use the shorthand notations,

$$\breve{\boldsymbol{\zeta}}_n := (\breve{\zeta}_n^1, \dots, \breve{\zeta}_n^N) \qquad \text{and} \qquad \breve{\boldsymbol{W}}_n := (\breve{W}_n^1, \dots, \breve{W}_n^N), \qquad n \geq 0,$$

where $\breve{\zeta}_n^i$ and $\breve{W}_n^i$ for all $1 \leq i \leq N$ will be defined below by Algorithms 4 and 5.

Algorithm 4 Augmented Island Resampling Particle Filter

for $i = 1, \dots, N$ **do**
  $\zeta_0^i \sim \pi_0$
for $n \geq 1$ **do**
  $(\breve{\boldsymbol{\zeta}}_{n-1}, \breve{\boldsymbol{W}}_{n-1}) \leftarrow$ WITHINISLANDRESAMPLE$(\boldsymbol{\zeta}_{n-1}, g_{n-1})$
  $\breve{g}_{n-1}(\cdot) \leftarrow \sum_{i=1}^N \breve{W}_{n-1}^i \mathbb{I}[\cdot = \breve{\zeta}_{n-1}^i]$
  $\widehat{\boldsymbol{\zeta}}_{n-1} \leftarrow$ AUGMENTEDISLANDRESAMPLE$(\breve{g}_{n-1}, \breve{\boldsymbol{\zeta}}_{n-1})$
  for $i = 1, \dots, N$ **do**
    $\zeta_n^i \sim f(\widehat{\zeta}_{n-1}^i, \cdot)$

Essentially AIRPF has two resampling steps. First is the within island (i.e. within PE) resampling step which preforms multinomial resampling of M particles within each PE. Subsequently, in the second step, the m groups of M particles per PE are resampled by duplicating the entire samples of size M without selecting individual particles within the samples. These two resampling subroutines will be analysed theoretically in the following two sections. The analysis is analogous to that conducted for the augmented resampling algorithm in Section 2. The more technical proof are postponed to Appendix C.

4.1 | Within island resampling

For a formal description of WITHINISLANDRESAMPLE we define $A \in \mathbb{R}^{mM \times mM}$ as

$$A := \mathbf{I}_m \otimes \mathbf{1}_{1/M},$$

and notations

$$\boldsymbol{\xi}_{\text{in}} := (\xi_{\text{in}}^1, \dots, \xi_{\text{in}}^N), \qquad \boldsymbol{W}_{\text{out}} := (W_{\text{out}}^1, \dots, W_{\text{out}}^N), \qquad \text{and} \qquad \boldsymbol{\xi}_{\text{out}} := (\xi_{\text{out}}^1, \dots, \xi_{\text{out}}^N).$$

The within island resampling then proceeds as described in Algorithm 5.

Algorithm 5 Within island resample

1: $(\boldsymbol{\xi}_{\text{out}}, \boldsymbol{W}_{\text{out}}) =$ WITHINISLANDRESAMPLE $\left(\boldsymbol{\xi}_{\text{in}}, g\right)$
2: **for** $i = 1, \dots, N$ **do**
3: $\quad W_{\text{out}}^i \leftarrow \sum_{j=1}^N A^{ij} g(\xi_{\text{in}}^i)$
4: $\quad \xi_{\text{out}}^i \sim (W_{\text{out}}^i)^{-1} \sum_{j=1}^N A^{ij} g(\xi_{\text{in}}^i) \delta_{\xi_{\text{in}}^j}$

From Algorithm 5 and the definition of A we obtain the following expression for the weights $\boldsymbol{W}_{\text{out}}$ returned by Algorithm 5

$$W_{\text{out}}^i := \frac{1}{M}\sum_{j=1}^{M} g(\xi_{\text{in}}^{(k-1)M+j}), \qquad (k-1)M < i \leq kM, \qquad 1 \leq k \leq m. \tag{17}$$

Note in particular that for any $1 \leq k \leq m$ the weights with indices in $\{(k-1)M+1, \ldots, kM\}$ are equal.

Proposition 4 below is our main result for Algorithm 5, and it is analogous to Proposition 1; part a) establishes a result similar to Proposition 1 for the entire sample $\boldsymbol{\xi}_{\text{out}}$ while part b) establishes a similar result for individual PEs, i.e. the sub-samples $(\xi_{\text{out}}^{(k-1)M+1}, \ldots, \xi_{\text{out}}^{kM})$, where $1 \leq k \leq m$.

Proposition 4 *Assume* (A1). *If $\boldsymbol{\xi}_{\text{in}}$ is any $\mathbb{X}^N$ valued random variable, $\boldsymbol{\xi}_{\text{out}}$ is generated according to Algorithm 5, $g \in \mathcal{B}_+(\mathbb{X})$ and we define probability measures*

$$\pi^N := \frac{1}{N}\sum_{i=1}^{N} \delta_{\xi_{\text{in}}^i}, \qquad \pi^{M,k} := \frac{1}{M}\sum_{i=1}^{M} \delta_{\xi_{\text{in}}^{(k-1)M+i}}, \qquad 1 \leq k \leq m, \tag{18}$$

and

$$\breve{\pi}^N := \frac{\sum_{i=1}^N W_{\text{out}}^i \delta_{\xi_{\text{out}}^i}}{\sum_{i=1}^N W_{\text{out}}^i} \qquad \breve{\pi}^{M,k} := \frac{\sum_{i=1}^M W_{\text{out}}^{(k-1)M+i} \delta_{\xi_{\text{out}}^{(k-1)M+i}}}{\sum_{i=1}^M W_{\text{out}}^{(k-1)M+i}} \qquad 1 \leq k \leq m,$$

then

a. *there exists $B_r \in \mathbb{R}_+$ such that for any $\varphi \in \mathcal{B}(\mathbb{X})$*

$$\mathbb{E}\left[\left|\pi^N(g)\breve{\pi}^N(\varphi) - \pi^N(g\varphi)\right|^r\right]^{\frac{1}{r}} \leq \frac{B_r \|g\| \operatorname{osc}(\varphi)}{\sqrt{N}}, \tag{19}$$

b. *there exists $B_r \in \mathbb{R}_+$ such that for any $\varphi \in \mathcal{B}(\mathbb{X})$ and any $1 \leq k \leq m$*

$$\mathbb{E}\left[\left|\pi^{M,k}(g)\left(\breve{\pi}^{M,k}(\varphi) - \frac{\pi^{M,k}(g\varphi)}{\pi^{M,k}(g)}\right)\right|^r\right]^{\frac{1}{r}} \leq \frac{B_r \|g\| \operatorname{osc}(\varphi)}{\sqrt{M}}. \tag{20}$$

4.2 | Augmented island resampling

Theoretically the augmented resampling for particle islands is very similar to the augmented resampling, Algorithm 3. With appropriate notational conventions the theoretical analysis becomes nearly identical to that of Section 2 with the exception that for any $1 \leq s \leq S$, we replace individual particles ξ_s^i by particle islands

$$\boldsymbol{\xi}_s^i := (\xi_s^{(i-1)M+1}, \ldots, \xi_s^{iM})$$

and set $M = 1$ to signify the fact that there is only one particle island per PE. Following the convention that $M = 1$, we define matrices $\overline{A}_1, \dots, \overline{A}_S$ analogously to (2) as

$$\overline{A}_s := \mathbf{I}_{2^{S-s}} \otimes \mathbf{1}_{1/2} \otimes \mathbf{I}_{2^{s-1}}.$$

The resulting algorithm is described in Algorithm 6.

Algorithm 6 Augmented island resampling

$\boldsymbol{\xi}_{\text{out}} = \text{AUGMENTEDISLANDRESAMPLE}\left(g, \boldsymbol{\xi}_{\text{in}}\right)$
for $i = 1, \dots, m$ **do**
  for $j = 1, \dots, M$ **do**
    $\xi_0^{(i-1)M+j} \leftarrow \xi_{\text{in}}^{(i-1)M+j}$
  $\overline{V}_0^i \leftarrow M^{-1} \sum_{j=1}^M g(\xi_0^{(i-1)M+j})$
for $s = 1, \dots, S$ **do**
  for $i = 1, \dots, m$ **do**
    $\overline{V}_s^i \leftarrow \sum_{j=1}^m \overline{A}_s^{ij} \overline{V}_{s-1}^j$
    $\boldsymbol{\xi}_s^i \sim (\overline{V}_s^i)^{-1} \sum_{j=1}^m \overline{A}_s^{ij} \overline{V}_{s-1}^j \delta_{\boldsymbol{\xi}_{s-1}^j}$
for $i = 1, \dots, N$ **do**
  $\xi_{\text{out}}^i \leftarrow \xi_S^i$.

Proposition 5 *Assume* (A1). *If $\boldsymbol{\xi}_{\text{in}}$ is any $\mathbb{X}^N$ valued random variable, $\boldsymbol{\xi}_{\text{out}}$ is computed according to Algorithm 6 and $g \in \mathcal{B}_+(\mathbb{X})$, then for any $r \geq 1$ there exists $B_r \in \mathbb{R}_+$ such that for any $\varphi \in \mathcal{B}(\mathbb{X})$,*

$$\mathbb{E}\left[\left|\left(\frac{1}{m}\sum_{i=1}^m \boldsymbol{g}(\boldsymbol{\xi}_{\text{in}}^i)\right)\left(\frac{1}{m}\sum_{i=1}^m \boldsymbol{\varphi}(\boldsymbol{\xi}_{\text{out}}^i)\right) - \frac{1}{m}\sum_{i=1}^m \boldsymbol{g}(\boldsymbol{\xi}_{\text{in}}^i)\boldsymbol{\varphi}(\boldsymbol{\xi}_{\text{in}}^i)\right|^r\right]^{\frac{1}{r}} \leq B_r \|g\| \sqrt{\frac{S}{m}} \operatorname{osc}(\varphi), \tag{21}$$

where

$$\boldsymbol{g}(\boldsymbol{x}) := \frac{1}{M}\sum_{i=1}^M g(x^i), \qquad \textit{and} \qquad \boldsymbol{\varphi}(\boldsymbol{x}) := \frac{1}{M}\sum_{i=1}^M \varphi(x^i),$$

for all $\boldsymbol{x} = (x^1, \dots, x^M) \in \mathbb{X}^M$.

Due to the similarity of the proof of Proposition 5 to that of Proposition 1 we will only outline the proof. First we construct a sequence $\overline{\mathcal{M}} := \{(\overline{X}_\rho, \overline{\mathcal{F}}_\rho); 0 \leq \rho \leq Sm\}$ such that $\overline{X}_0 := 0$ and $\overline{\mathcal{F}}_0 := \sigma(\boldsymbol{\xi}_{\text{in}})$ and for all $1 \leq \rho \leq Sm$

$$\overline{X}_\rho := \frac{\overline{V}_{s_m(\rho)}^{i_m(\rho)}}{\sqrt{Sm}}\left(\overline{\boldsymbol{\varphi}}(\boldsymbol{\xi}_{s_m(\rho)}^{i_m(\rho)}) - \frac{\sum_{j=1}^m \overline{A}_{s_m(\rho)}^{i_m(\rho)j} \overline{V}_{s_m(\rho)-1}^j \overline{\boldsymbol{\varphi}}(\boldsymbol{\xi}_{s_m(\rho)-1}^j)}{\overline{V}_{s_m(\rho)}^{i_m(\rho)}}\right), \qquad \overline{\mathcal{F}}_\rho := \overline{\mathcal{F}}_{\rho-1} \vee \sigma(\boldsymbol{\xi}_{s_m(\rho)}^{i_m(\rho)}),$$

FIGURE 2 HERE

where $i_m(\rho)$ and $s_m(\rho)$ are as defined in (9), and

$$\overline{\varphi}(\xi_s^i) := \varphi(\xi_s^i) - \frac{\sum_{j=1}^m g(\xi_0^j)\varphi(\xi_0^j)}{\sum_{j=1}^m g(\xi_0^j)}, \qquad 1 \le i \le m,\ 1 \le s \le S. \tag{22}$$

With these notations we obtain the following result, analogous to Proposition 2. The proof is essentially identical to that of Proposition 2 and hence omitted.

Proposition 6 *Assume* (A1). *The following statements hold:*

a. $\overline{X}_\rho$ *is* $\overline{\mathcal{F}}_\rho$*-measurable for all* $0 \le \rho \le Sm$*;*
b. $\mathbb{E}\left[\overline{X}_\rho \mid \overline{\mathcal{F}}_{\rho-1}\right] = 0$ *(a.s.) for all* $0 < \rho \le Sm$*;*
c. $\left|\overline{X}_\rho\right| \le \|g\|\mathrm{osc}(\varphi)/\sqrt{Sm}$ *for all* $0 \le \rho \le Sm$*;*
d. *and we have the identities*

$$\sqrt{\frac{S}{m}}\sum_{\rho=1}^{Sm}\overline{X}_\rho = \frac{1}{m}\sum_{i=1}^m \overline{V}_S^i\overline{\varphi}(\xi_S^i) = \left(\frac{1}{m}\sum_{i=1}^m g(\xi_0^i)\right)\left(\frac{1}{m}\sum_{i=1}^m \varphi(\xi_S^i)\right) - \frac{1}{m}\sum_{i=1}^m g(\xi_0^i)\varphi(\xi_0^i)$$

Proposition 5 then follows by Burkholder-Davis-Gundy inequality similarly as Proposition 1.

4.3 | Modified augmented resampling

In the introduction we stated that augmented resampling enables a straightforward way to further control the communication of particle information between PEs. We will now address this claim more closely.

By Lemma 1 we know that $\overline{A}_s$ is symmetric and that for any $1 \le s \le S$, each row of $\overline{A}_s$ has exactly two nonzero elements of which one is on the diagonal. This implies that the pairs of indices of the nonzero columns for each row of $\overline{A}_s$ form a partition of $\{1, \dots, m\}$ into $m/2$ pairs of indices

$$P_{i,s} := (\ell_s^i, r_s^i), \qquad 1 \le i \le m/2,$$

for which, by (6), we can obtain explicit expressions as

$$\begin{aligned}(\ell_s^1, \dots, \ell_s^{m/2}) &:= ((2^s(i-1) + (j-1) + 1)_{1\le j\le 2^{s-1}})_{1\le i\le 2^{S-s}},\\ (r_s^1, \dots, r_s^{m/2}) &:= ((2^s(i-1) + (j-1) + 1 + 2^{s-1})_{1\le j\le 2^{s-1}})_{1\le i\le 2^{S-s}}.\end{aligned}$$

If, for any $1 \le s \le S$ we associate the subsample ξ_s^i with PE i, as we have done so far, then the pair $P_{i,s}$ has the interpretation of representing the indices of PEs that are paired up for communication at stage s, and they are illustrated in Figure 2.

Remark 2 *For our purposes, the indexing of pairs* $(P_{1,s}, \dots, P_{m/2,s})$ *where* $1 \le s \le S$ *is fixed could be replaced with any permutation of* $\{1, \dots, m/2\}$.

Now consider the PE ℓ_s^i at stage s for some $1 \le i \le m/2$. By line 9 of Algorithm 6 and the discussion above we now

see that when drawing the sample $\boldsymbol{\xi}_s^{\ell_s^i}$, PE ℓ_s^i essentially randomly decides whether to keep the sample $\boldsymbol{\xi}_{s-1}^{\ell_s^i}$ from the previous stage or assume the sample $\boldsymbol{\xi}_{s-1}^{r_s^i}$ of its paired PE. Simultaneously the paired PE r_s^i makes randomly and independently a similar decision between $\boldsymbol{\xi}_{s-1}^{\ell_s^i}$ or $\boldsymbol{\xi}_{s-1}^{r_s^i}$.

It may be the case, in particular if $\overline{V}_{s-1}^{\ell_s^i} \approx \overline{V}_{s-1}^{r_{i,s}}$, that after the random sampling on line 9 of Algorithm 6 at stage s, one has $\left(\boldsymbol{\xi}_s^{\ell_s^i}, \boldsymbol{\xi}_s^{r_s^i}\right) = \left(\boldsymbol{\xi}_{s-1}^{r_s^i}, \boldsymbol{\xi}_{s-1}^{\ell_s^i}\right)$ i.e. the paired PEs have simply exchanged their particles. Intuitively, it seems that performing this exchange is unnecessary, as the purpose of resampling is to duplicate particles appropriately many, possibly zero, times and hence only the number of duplicates is expected to matter, not the order in which they are allocated to the PEs. Thus to reduce the time spent on the communication between PEs, it seems advisable to avoid the above-mentioned exchange. Algorithm 7 describes a simple modification of Algorithm 6 designed to avoid this seemingly redundant particle exchange.

Algorithm 7 Modified Augmented Island Resampling

1: $\boldsymbol{\xi}_{\text{out}} = \text{AUGMENTEDISLANDRESAMPLE}\left(g, \boldsymbol{\xi}_{\text{in}}\right)$
2: **for** $i = 1, \ldots, m$ **do**
3: **for** $j = 1, \ldots, M$ **do**
4: $\xi_0^{(i-1)M+j} \leftarrow \xi_{\text{in}}^{(i-1)M+j}$
5: $\overline{V}_0^i \leftarrow g(\xi_0^i)$
6: **for** $s = 1, \ldots, S$ **do**
7: **for** $i = 1, \ldots, m$ **do**
8: $\overline{V}_s^i \leftarrow \sum_{j=1}^m \overline{A}_s^{ij} \overline{V}_{s-1}^j$
9: $\widetilde{\xi}_s^i \sim (\overline{V}_s^i)^{-1} \sum_{j=1}^m \overline{A}_s^{ij} \overline{V}_{s-1}^j \delta_{\xi_{s-1}^j}$
10: **for** $i = 1, \ldots, m/2$ **do**
11:

$$\left(\boldsymbol{\xi}_s^{r_s^i}, \boldsymbol{\xi}_s^{\ell_s^i}\right) = \begin{cases} \left(\widetilde{\boldsymbol{\xi}}_s^{\ell_s^i}, \widetilde{\boldsymbol{\xi}}_s^{r_s^i}\right), & \text{if} \quad \left(\widetilde{\boldsymbol{\xi}}_s^{r_s^i}, \widetilde{\boldsymbol{\xi}}_s^{\ell_s^i}\right) = \left(\widetilde{\boldsymbol{\xi}}_{s-1}^{\ell_s^i}, \widetilde{\boldsymbol{\xi}}_{s-1}^{r_s^i}\right) \\ \left(\widetilde{\boldsymbol{\xi}}_s^{r_s^i}, \widetilde{\boldsymbol{\xi}}_s^{\ell_s^i}\right), & \text{otherwise.} \end{cases}$$

12: **for** $i = 1, \ldots, N$ **do**
13: $\xi_{\text{out}}^i \leftarrow \xi_S^i$.

The modification on lines 10 and 11 of Algorithm 7 changes slightly the statistical behaviour of the algorithm and hence Propositions 5 and 6 are not immediately valid for Algorithm 7. However, similar results with an appropriate martingale difference construction can be obtained.

We define a sequence $\widetilde{\mathcal{M}} := \{(\widetilde{X}_\rho, \widetilde{\mathcal{F}}_\rho); 0 \leq \rho \leq Sm/2\}$ such that $\widetilde{X}_0 = 0$ and $\widetilde{\mathcal{F}}_0 = \sigma(\boldsymbol{\xi}_0)$, and for all $0 < \rho \leq Sm/2$,

$$\widetilde{X}_\rho := \widetilde{X}_{s(\rho)}^{r(\rho)} + \widetilde{X}_{s(\rho)}^{\ell(\rho)}, \qquad \widetilde{\mathcal{F}}_\rho := \widetilde{\mathcal{F}}_{\rho-1} \vee \sigma(\boldsymbol{\xi}_{s(\rho)}^{r(\rho)}) \vee \sigma(\boldsymbol{\xi}_{s(\rho)}^{\ell(\rho)}), \tag{23}$$

where $r(\rho) := r_{s_{m/2}(\rho)}^{i_{m/2}(\rho)}$, $\ell(\rho) := \ell_{s_{m/2}(\rho)}^{i_{m/2}(\rho)}$, $s(\rho) := s_{m/2}(\rho)$ and $i_{m/2}(\rho)$ and $s_{m/2}(\rho)$ are as defined in (9), and for all $1 \leq i \leq m/2$

and $1 \leq s \leq S$

$$\widetilde{X}_s^i = \frac{\overline{V}_s^i}{\sqrt{Sm}} \left(\overline{\varphi}(\xi_s^i) - \frac{1}{\overline{V}_s^i} \sum_{j=1}^m \overline{A}_s^{ij} \overline{V}_{s-1}^j \overline{\varphi}(\xi_{s-1}^j) \right).$$

Function $\overline{\varphi}$ is as defined in (22).

Proposition 7 *Assume (A1). We have the following*

a. $\widetilde{X}_\rho$ *is* $\widetilde{\mathcal{F}}_\rho$*-measurable for all* $0 \leq \rho \leq Sm/2$.
b. $\mathbb{E}[\widetilde{X}_\rho \mid \widetilde{\mathcal{F}}_{\rho-1}] = 0$ *a.s. for all* $0 < \rho \leq Sm/2$.
c. $\widetilde{X}_\rho \leq 2\|g\|\mathrm{osc}(\varphi)/\sqrt{Sm}$*, for all* $0 \leq \rho \leq Sm/2$.
d. *and we have the identities*

$$\sqrt{\frac{S}{m}} \sum_{\rho=1}^{Sm/2} \widetilde{X}_\rho = \frac{1}{m} \sum_{i=1}^m \overline{V}_S^i \overline{\varphi}(\xi_S^i) \tag{24}$$

$$= \left(\frac{1}{m} \sum_{i=1}^m g(\xi_0^i) \right) \left(\frac{1}{m} \sum_{i=1}^m \varphi(\xi_S^i) \right) - \frac{1}{m} \sum_{i=1}^m g(\xi_0^i)\varphi(\xi_0^i) \tag{25}$$

Proposition 8 *Assume* (A1) *and let* $g \in \mathcal{B}_+(\mathbb{X})$*. Then for any* $r \geq 1$ *there exists a finite constant* $\widetilde{B}_r$*, depending only on* r*, such that no matter what the distribution of* ξ_0 *is, we have*

$$\mathbb{E}\left[\left| \left(\frac{1}{m} \sum_{i=1}^m g(\xi_{\mathrm{in}}^i) \right) \left(\frac{1}{m} \sum_{i=1}^m \varphi(\xi_{\mathrm{out}}^i) \right) - \frac{1}{m} \sum_{i=1}^m g(\xi_{\mathrm{in}}^i)\varphi(\xi_{\mathrm{in}}^i) \right|^r \right]^{\frac{1}{r}} \leq \tilde{B}_r \sqrt{\frac{S}{m}} \|g\| \mathrm{osc}(\varphi).$$

Proof The claim follows by applying the Burkholder-Davis-Gundy inequality to the expectation

$$\mathbb{E}\left[\left| \sqrt{\frac{S}{m}} \sum_{\rho=1}^{Sm/2} \widetilde{X}_\rho \right|^r \right].$$

5 | AUGMENTED ISLAND RESAMPLING PARTICLE FILTER

We will now analyse the convergence properties of Algorithm 4. The analysis is somewhat more complicated than the analogous analysis in Section 3. Additional complications arise due to Proposition 5 being independent of M. This implies that the two regimes identified earlier, i.e. m fixed, $M \to \infty$ (regime 1) and $m \to \infty$, M fixed (regime 2), cannot be covered by one overarching analysis as before, but the scenarios have to be studied separately. The more technical proofs are postponed to Appendix D.

5.1 | Convergence when m is fixed and $M \to \infty$

We introduce the following two PE specific empirical measure approximations

$$\pi_n^{M,k} := \frac{1}{M}\sum_{j=1}^{M} \delta_{\zeta_n^{(k-1)M+j}}, \quad \widehat{\pi}_n^{M,k} := \frac{1}{M}\sum_{j=1}^{M} \delta_{\widehat{\zeta}_n^{(k-1)M+j}}, \quad 1 \le k \le m, \tag{26}$$

for π_n and $\widehat{\pi}_n$, respectively, based on the PE specific subsamples

$$\zeta_n^k := (\zeta_n^{(k-1)M+1}, \dots, \zeta_n^{kM}), \quad \text{and} \quad \widehat{\zeta}_n^k := (\widehat{\zeta}_n^{(k-1)M+1}, \dots, \widehat{\zeta}_n^{kM}) \qquad 1 \le k \le m.$$

With these notations we have the following analogue of Proposition 3.

Proposition 9 *Assume* (A1) *and* (A2). *If for some $n \ge 0$, there exists $C_{n,r} \in \mathbb{R}_+$ such that for all $1 \le k \le m$*

$$\sup_{\varphi \in \mathcal{B}_1(\mathbb{X})} \mathbb{E}\left[\left|\pi_n^{M,k}(\varphi) - \pi_n(\varphi)\right|^r\right]^{\frac{1}{r}} \le \frac{C_{n,r}}{\sqrt{M}}, \tag{27}$$

then there exists $\widehat{C}_{n,r} \in \mathbb{R}_+$ such that for all $1 \le k \le m$

$$\sup_{\varphi \in \mathcal{B}_1(\mathbb{X})} \mathbb{E}\left[\left|\widehat{\pi}_n^{M,k}(\varphi) - \widehat{\pi}_n(\varphi)\right|^r\right]^{\frac{1}{r}} \le \frac{\widehat{C}_{n,r}}{\sqrt{M}}. \tag{28}$$

Proposition 9 enables us to proof the convergence of Algorithm 4 by induction according to the following theorem.

Theorem 2 *Assume* (A1) *and* (A2). *If π_n^N is computed according to Algorithm 4, then for all $n \ge 0$ there exists $C_{n,r} \in \mathbb{R}_+$ such that*

$$\sup_{\varphi \in \mathcal{B}_1(\mathbb{X})} \mathbb{E}\left[\left|\pi_n^{N}(\varphi) - \pi_n(\varphi)\right|^r\right]^{\frac{1}{r}} \le \frac{C_{n,r}}{\sqrt{M}}.$$

Proof The proof is by induction, the assumption being that for some $n \ge 0$

$$\sup_{\varphi \in \mathcal{B}_1(\mathbb{X})} \sup_{1 \le k \le m} \mathbb{E}\left[\left|\pi_n^{M,k}(\varphi) - \pi_n(\varphi)\right|^r\right]^{\frac{1}{r}} \le \frac{C_{n,r}}{\sqrt{M}}. \tag{29}$$

The induction is started by observing that (29) holds for $n = 0$ by Lemma 4. Now suppose (29) holds for some $n \ge 0$. By Minkowski's inequality

$$\sup_{\varphi \in \mathcal{B}_1(\mathbb{X})} \mathbb{E}[|\pi_{n+1}^{M,k}(\varphi) - \pi_{n+1}(\varphi)|^r]^{1/r} \le \sup_{\varphi \in \mathcal{B}_1(\mathbb{X})} \mathbb{E}[|\pi_{n+1}^{M,k}(\varphi) - \widehat{\pi}_n^{M,k}(f(\varphi))|^r]^{1/r} + \sup_{\varphi \in \mathcal{B}_1(\mathbb{X})} \mathbb{E}[|\widehat{\pi}_n^{M,k}(f(\varphi)) - \widehat{\pi}_n(f(\varphi))|^r]^{1/r}.$$

Similarly as in the proof of Theorem 1 we can bound the two terms on the r.h.s. by using Lemma 5 and Proposition 9, respectively, to see that (29) holds for $n+1$ with $C_{n+1,r} = 2B_r + \widehat{C}_{n,r}$.

5.2 | Convergence when $m \to \infty$ and M is fixed

For regime 2 we have the following analogues of Proposition 9 and Theorem 2.

Proposition 10 *Assume* (A1) *and* (A2). *If for some* $n \geq 0$ *there exists* $C_{n,r} \in \mathbb{R}_+$ *such that*

$$\sup_{\varphi \in \mathcal{B}_1(\mathbb{X})} \mathbb{E}\left[\left|\pi_n^N(\varphi) - \pi_n(\varphi)\right|^r\right]^{\frac{1}{r}} \leq C_{n,r}\sqrt{\frac{S}{m}}, \tag{30}$$

then there exists $\widehat{C}_{n,r} \in \mathbb{R}$ *such that*

$$\sup_{\varphi \in \mathcal{B}_1(\mathbb{X})} \mathbb{E}\left[\left|\widehat{\pi}_n^N(\varphi) - \widehat{\pi}_n(\varphi)\right|^r\right]^{\frac{1}{r}} \leq \widehat{C}_{n,r}\sqrt{\frac{S}{m}},$$

where $\widehat{\pi}_n^N = N^{-1}\sum_{i=1}^N \delta_{\widehat{\xi}_n^i}$ *for all* $n \geq 0$.

Theorem 3 *Assume* (A1) *and* (A2). *If* π_n^N *is computed according to Algorithm 4, then for all* $n \geq 0$ *there exists* $C_{n,r} \in \mathbb{R}_+$ *such that*

$$\sup_{\varphi \in \mathcal{B}_1(\mathbb{X})} \mathbb{E}\left[\left|\pi_n^N(\varphi) - \pi_n(\varphi)\right|^r\right]^{\frac{1}{r}} \leq C_{n,r}\sqrt{\frac{S}{m}}.$$

Proof The proof follows by induction analogously to the proof of Theorem 2 by using Lemmata 4 and 5 and Proposition 10.

Remark 3 *We can extend Proposition 10 and Theorem 3 straightforwardly to Algorithm 4 deploying the modified augmented resampling algorithm (Algorithm 7) presented in Section 4.3. This follows from observing that at every step of the respective proofs we can replace Proposition 5 with Proposition 8.*

6 | NUMERICAL EXPERIMENTS

We have seen that AIRPF, as well as BPF where we have replaced the resampling step with an augmented resampling step, are valid algorithms in the sense that they are convergent. However, we have also seen in Theorems 1 and 3 that augmented resampling imposes constraints on the interactions that introduce error and this error is manifested as a slower convergence rate. This raises the practically important question whether these algorithms are not only faster than the existing methods, but is the speedup significant enough to outweigh the introduced error. We now aim to shed some light to this question with numerical experiments.

6.1 | Experimental setup

In order to obtain accurate error estimates, we chose to run the experiments on a simple random walk HMM which admits exact numerical calculation by Kalman filter recursions (Kalman, 1960). The model we used is

$$\begin{aligned}
X_0 &\sim \mathcal{N}(\mathbf{0}_d, \mathbf{I}_d), \\
X_n &= X_{n-1} + \eta_n,\ \eta_n \sim \mathcal{N}(\mathbf{0}_d, \mathbf{I}_d), && n > 0, \\
Y_n &= X_n + \varepsilon_n, \quad \varepsilon_n \sim \mathcal{N}\left(\mathbf{0}_d, \frac{1}{4}\mathbf{I}_d\right), && n \geq 0,
\end{aligned}$$

where $\mathbf{0}_d$ denotes a vector of zeros in $\mathbb{R}^d$ and $\mathcal{N}(\mu, \sigma)$ denotes a Gaussian distribution with mean μ and covariance σ.

The approximation error was taken to be the mean squared error

$$\mathrm{MSE} := \frac{1}{J}\sum_{j=1}^{J}\sum_{n=0}^{n_{\max}-1}\sum_{i=1}^{d}\left(\overline{x}_{n,j}^{N,i} - \overline{x}_n^i\right)^2.$$

where J is the number of independent runs, $n_{\max}$ is the length of the time series, $\overline{x}_n := (\overline{x}_n^1, \ldots, \overline{x}_n^d)$ is the mean of the true filtering distribution, and $\overline{x}_{n,j}^N := (\overline{x}_{n,j}^{N,1}, \ldots, \overline{x}_{n,j}^{N,d})$, where $1 \le j \le J$, denotes the approximation of $\overline{x}_n$ at the jth run.

We used this model with $d = 7$, to generate a data set of length $n_{\max} = 8000$ iterations and the error was calculated of $J = 5$ independent runs. The choice of dimension $d = 7$ is largely arbitrary although very low dimensions were intentionally avoided to introduce some pressure for the ESS to take low values which in turn emphasises the role of adaptive resampling scheme.

In addition to this tractable model, we also considered another, more practically motivated model, namely the *prokaryotic autoregulation model* as presented by Golightly and Kypraios (2018) and the references therein. In this case the exact filter is intractable and hence, instead of studying the error to the exact filter, we studied the variance of the filter mean integrated over time as our performance measure. This variance was evaluated from $J = 80$ realisations of the filter over a data sequence of length $n_{\max} = 10000$.

The algorithms were implemented in C and the parallelism was implemented using Intel MPI. The experiments were conducted on the high performance computing system Balena at the University of Bath using 16 computing nodes each capable of running 16 processes simultaneously. The code is available at `https://github.com/heinekmp/AIRPF`

6.2 | Algorithms

For reference, two versions of IPF were implemented. The first version was our implementation of the original IPF which performed the between island resampling first and the within island resampling second (IPF1). A slight gain in efficiency is expected if the order of these resampling steps is reversed as this would mean that PEs would not have to communicate the individual particles but only a single weight per PE. The IPF with this reversed resampling order we call IPF2.

For the algorithms proposed in this paper, we implemented AIRPF with the modified augmented island resampling algorithm, Algorithm 7 (AIRPF1). In order to make the comparison against IPF as fair as possible, also both versions of IPF deployed a modification analogous to Algorithm 7; if the sample of any PE was duplicated at the between islands resampling stage, then the PE in question was ensured to keep one copy of the sample set.

In accordance with our discussion in Section 1.2, we also implemented AIRPF with the fully adapted resampling scheme (AIRPF2). This means that for each stage $0 \le s \le S$ in Algorithm 6 we evaluate the ESS

$$\frac{\left(m^{-1}\sum_{i=1}^{m}\overline{V}_s^i\right)^2}{m^{-1}\sum_{i=1}^{m}\left(\overline{V}_s^i\right)^2}$$

and proceed to the next stage only if this value is less than some predetermined threshold θ. If all stages are completed, the particles will remain unweighted, but otherwise, if the last executed resampling stage is $0 \le s < S$, the output of Algorithm 6, $\widehat{\zeta}_{n-1}$ in Algorithm 4, will be associated with weights

$$\widehat{W}_{n-1}^{(k-1)M+j} = \overline{V}_s^k, \quad 1 \le j \le M, \quad 1 \le k \le m. \tag{31}$$

FIGURE 3 HERE

FIGURE 4 HERE

These weights need to be taken into account also in the WITHINISLANDRESAMPLE step of the next iteration, which means that the lines (3) and (4) in Algorithm 5 are replaced with

$$W^i_{\text{out}} \leftarrow \sum_{j=1}^{N} A^{ij} g(\xi^i_{\text{in}}) W^i_{\text{in}} \qquad \text{and} \qquad \xi^i_{\text{out}} \sim (W^i_{\text{out}})^{-1} \sum_{j=1}^{N} A^{ij} g(\xi^i_{\text{in}}) W^i_{\text{in}} \delta_{\xi^j_{\text{in}}},$$

where $(W^1_{\text{in}}, \dots, W^N_{\text{in}})$ are the weights $(\widehat{W}^1_{n-1}, \dots, \widehat{W}^N_{n-1})$ from the previous iteration as defined in (31).

6.3 | Results

For the tractable model, the algorithms were run with 22 roughly equally spaced values of M in the interval $\{200, \dots, 4200\}$ and $m \in \{64, 128, 256\}$. The resampling threshold θ took values in $\{.1, .2, .4, .6, .8, 1\}$. Each computing node used in our experiments always used its full capacity of 16 processes.

If we fix m and M and let θ vary, we obtain four MSE vs. time curves; one curve for each algorithm. Figure 3a illustrates such sets of four curves for four different values of $M \in \{200, 400, 600, 800\}$. To clarify which curves are obtained with the same value of M, the curves obtained with $M = 200$ are highlighted by a rectangle in Figure 3a. In order to improve the visualisation by reducing the overlap of the curves, we calculated the lower envelope curves for each algorithm as shown in Figure 3b. The horizontal dashed line at level 2396 denotes the worst case MSE which is obtained by taking the raw observations as the estimates of the filtering mean.

Figure 4a shows the lower envelopes of MSE vs. time curves for the entire range of M and m. Differences between IPF and AIRPF are more pronounced for larger values of m and, in particular, for moderate values of M. For large M, the differences vanish as the resampling that takes place within each PE independently begins to dominate the execution time. For moderate values of M, communication cost plays a more significant role, and in this case, AIRPF is more efficient than IPF. Also a notable gain in performance can be observed due to the fully adapted resampling. Figure 4b summarises the lower envelopes for AIRPF2 and IPF2 for the whole range of M and m.

To study the error introduced by the augmented resampling and the consequently lower convergence rate, Figure 5a shows the square root of the mean square error (RMSE) scaled by $\sqrt{\log_2(m)/m}$ and $\sqrt{m}$ for the tractable model using $m \in \{2, 4, 8, 16, 32, 64, 128, 256$ $M = 1000$ and $\theta = 1$. As the curve corresponding to the scaling by $\sqrt{\log_2(m)/m}$ is nearly constant, we conclude that the results are consistent with the convergence rate of Theorem 3.

For the prokaryotic autoregulation model, we ran AIRPF2 and IPF2 with $m = 256$, $M \in \{25, 50, 75, 100, 150\}$ and $\theta \in \{.1, .25, .5, .75\}$. The results are summarised in Figure 5b which shows the integrated variance for one state dimension vs. execution time. A pattern similar to that in Figure 4 is observed, with the exception that AIRPF2 and IPF2 differ from each other only for relatively small values of M. This can be attributed to the fact that for the prokaryotic autoregulation model, the state space dimension is $d = 4$ while for the tractable model we have $d = 7$. This means that communicating the particles between PEs is faster and therefore the time spent on resampling will dominate the communication time for smaller M.

FIGURE 5 HERE

6.4 | Conclusions

Based on the results reported above, it appears that AIRPF shows the best potential in scenarios where the execution time is critical but the best possible accuracy with given resources is required. Consider for example Figure 4b with the maximum execution time being limited to 20 seconds. In this case, increasing M is not an option as it implies longer computation time. Also, the speedup by increasing the computer clock rate, which in turn would enable larger values of M, is not an option as modern computers have essentially reached their limit in clock rate. Therefore the only option is to increase m, and in the case of IPF2, Figure 4b suggest that also this may ineffective as the error of IPF2 is essentially the same with $m = 128$ and $m = 256$. Although AIRPF2 will presumably reach a similar limit for large enough m, Figure 4b suggests that the error will decrease from the current value of $m = 256$ for substantially larger values of m in which case AIRPF could be the method of choice.

Both AIRPF1 and AIRPF2 use the modification proposed in Section 4.3. We also experimented with AIRPF without this modification, but the performance in that case was notably worse in terms of execution time. This leads us to attribute the efficiency of AIRPF that we have seen above, to the ideas presented in Section 4.3 rather than to the assumptions about the idealised computer architecture of Section 1.2. We therefore believe that the performance of AIRPF could be further improved in certain situations. A scenario which seems particularly well suited for AIRPF is a computer network with a hypercube topology which matches exactly the radix-2 butterfly diagram structure of the AIRPF resampling step. In a hypercube network, PEs are not connected to all other PEs but to only those PEs that they are required to communicate with in the radix-2 augmented resampling. Clearly, all these communication channels are completely separate and hence our assumptions about the idealised architecture seem more reasonable. However, we believe that this too, can only be verified empirically. The experiments presented above were not executed in a hypercube architecture and hence the performance of AIRPF2 is mostly attributed to the ideas presented in Section 4.3.

We also believe that the performance of AIRPF can be further improved with the following algorithmic modification. In the current fully adapted AIRPF the augmented resampling always begins at stage $s = 1$, but presumably fewer resampling stages would have to be executed in total if the resampling was started at $s + 1$, s being the last executed resampling stage of the previous iteration that included resampling. The rationale for this modification is simple. By starting the resampling always at stage $s = 1$ we introduce a bias towards the pairwise interactions associated with stage $s = 1$ but by rotating the first resampling stage this bias is removed and more complete interactions are obtained which in turn is expected to increase ESS and lead to fewer resampling stages being executed in total.

The augmented resampling framework is a generalisation of multinomial resampling in the sense that a single stage augmented resampling is equivalent to multinomial resampling, but other resampling methods exist, e.g. stratified, systematic and residual resampling. The methods proposed in this paper lend themselves immediately to experiments with these alternative resampling algorithms; for example the multinomial resampling of Algorithm 5 can be replaced with any of the alternative methods, and if a radix parameter greater than 2 is considered, the different stages of Algorithm 6 become essentially applications of multinomial resampling which again may be replaced with the above-mentioned alternatives. However, it does not seem obvious, how the theory presented in this paper could be extended to generalise these alternative resampling methods the same way we have now generalised multinomial resampling.

For future theoretical research the convergence properties of the fully adapted resampling AIRPF remain to be analysed. Although inspired by (Whiteley *et al.*, 2016), the existing results do not immediately apply to AIRPF2 due to the constrained interactions. However, our conjecture is that similar uniform convergence results as in (Whiteley *et al.*, 2016) can be obtained by ensuring a control on the ESS. The results are also expected to extend to versions of AIRPF whereby an arbitrary subset of resampling stages are selected for execution at each iteration while ensuring the control of the ESS, but we believe that the most efficient heuristic is to choose the stage subsequent to the previously executed stage as suggested above.

We finish on a more practical note by pointing out that the validity of the proposed algorithms will hold for various definitions

of $A_1, \ldots, A_S$. In the context of the present paper, we have only considered the radix-2 structure as it limits the communication between PEs to pairwise interactions and offers an immediate connection with the hypercube network topology. An obvious question is the impact of the radix parameter on the efficiency. The radix parameter introduces an obvious trade-off between the number of simultaneously communicating PEs and the number of resampling stages. One possibility is to consider a different radix parameter for each stage as proposed in (Heine *et al.*, 2014). It seems impossible to make a statement claiming the superiority of one design of $A_1, \ldots, A_S$ over another in general, as we believe that the optimal design will inevitably depend on the physical computing system in question.

This research made use of the Balena High Performance Computing (HPC) Service at the University of Bath.

REFERENCES

Bolić, M., Djurić, P. M. and Hong, S. (2005) Resampling algorithms and architectures for distributed particle filters. *IEEE Trans. Signal Process.*, **53**, 2442–2450.

Chopin, N. (2004) Central limit theorem for sequential Monte Carlo methods and its application to Bayesian inference. *Ann. Statist.*, **32**, 2385–2411.

Crisan, D. and Doucet, A. (2002) A survey of convergence results on particle filtering methods for practitioners. *IEEE Trans. Signal Process.*, **50**, 736–746.

Del Moral, P. (2004) *Feynman-Kac Formulae. Genealogical and interacting particle systems with applications*. Probab. Appl. New York: Springer Verlag.

Del Moral, P. and Guionnet, A. (1999) Central limit theorem for nonlinear filtering and interacting particle systems. *Ann. Appl. Probab.*, **9**, 275–297.

— (2001) On the stability of interacting processes with applications to filtering and genetic algorithms. *Ann. Inst. Henri Poincaré Probab. Stat.*, **37**.

Del Moral, P., Moulines, E., Olsson, J. and Vergé, C. (2017) Convergence properties of weighted particle islands with application to the double bootstrap algorithm. *Stoch. Syst.*, **6**, 367–419.

Douc, R., Moulines, E. and Olsson, J. (2014) Long-term stability of sequential Monte Carlo methods under verifiable conditions. *Ann. Appl. Probab.*, **24**.

Doucet, A., De Freitas, N. and Gordon, N. (eds.) (2001) *Sequential Monte Carlo methods in practice*. New York: Springer.

Golightly, A. and Kypraios, T. (2018) Efficient SMC^2 schemes for stochastic kinetic models. *Stat. Comput.*, **28**, 1215–1230.

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

Heine, K. and Whiteley, N. (2017) Fluctuations, stability and instability of a distributed particle filter with local exchange. *Stochastic Process. Appl.*, **127**, 2508–2541.

Heine, K., Whiteley, N., Cemgil, A. T. and Güldaş, H. (2014) Butterfly resampling: Asymptotics for particle filters with constrained interactions. *arXiv:1411.5876*.

Horn, R. A. and Johnson, C. R. (1991) *Topics in matrix analysis*. Cambridge University Press.

Kalman, R. E. (1960) A new approach to linear filtering and prediction problems. *J. Basic Eng*, **82**.

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

Liu, J. S. and Chen, R. (1995) Blind deconvolution via sequential imputation. *J. Amer. Statist. Assoc.*, **90**.

— (1998) Sequential Monte Carlo methods for dynamic systems. *J. Amer. Statist. Assoc.*, **93**.

Míguez, J. (2007) Analysis of parallelizable resampling algorithms for particle filtering. *Signal Processing*, **87**, 3155–3174.

— (2014) On the uniform asymptotic convergence of a distributed particle filter. In *IEEE 8th Sensor Array and Multichannel Signal Processing Workshop (SAM)*, 241–244. IEEE.

Míguez, J. and Vázquez, M. A. (2016) A proof of uniform convergence over time for a distributed particle filter. *Signal Processing*, **122**, 152–163.

Murray, L., Lee, A. and Jacob, P. (2016) Parallel resampling in the particle filter. *J. Comput. Graph. Statist.*, **25**.

Oppenheim, A. (1975) *Digital Signal Processing*. Prenctice-Hall.

Pacheco, P. (2011) *An Introduction to Parallel Programming*. Morgan Kaufmann Publishers Inc., 1st edn.

Vergé, C., Dubarry, C., Del Moral, P. and Moulines, E. (2015) On parallel implementation of sequential Monte Carlo methods: the island particle model. *Stat. Comput.*, **25**, 243–260.

Whiteley, N., Lee, A. and Heine, K. (2016) On the role of interaction in sequential Monte Carlo algorithms. *Bernoulli*, **22**, 494–529.

A | PROOFS FOR SECTION 2

Proof of Lemma 1 First we recall the *mixed product property* of Kronecker product: for any matrices A, B, C and D, such that the products AC and BD are defined, one has (e.g. Horn and Johnson, 1991)

$$(A \otimes B)(C \otimes D) = (AC) \otimes (BD). \tag{32}$$

Also we note that for any two square matrices $A \in \mathbb{R}^{p\times p}$ and $B \in \mathbb{R}^{q\times q}$ we have the element-wise formula:

$$(A \otimes B)^{ij} = A^{\left\lfloor \frac{i-1}{q} \right\rfloor+1,\left\lfloor \frac{j-1}{q} \right\rfloor+1} B^{((i-1) \bmod q)+1,((j-1) \bmod q)+1}, \tag{33}$$

where $i, j \in \{1, \dots, pq\}$. By (33), $A \otimes B$ is symmetric whenever A and B are symmetric, proving the symmetry. Associativity of the Kronecker product and repeated applications of (32) to the definition of A_s in (2) yield $A_s A_s = A_s$ proving the idempotence. Also, by associativity and repeated applications of (33) to (2), we have

$$A_s^{ij} = \mathbf{I}_{2^{S-s}}^{\left\lfloor \frac{i-1}{2^s M} \right\rfloor+1,\left\lfloor \frac{j-1}{2^s M} \right\rfloor+1} \mathbf{1}_{1/2}^{\left(\left\lfloor \frac{i-1}{2^{s-1} M} \right\rfloor \bmod 2\right)+1,\left(\left\lfloor \frac{j-1}{2^{s-1} M} \right\rfloor \bmod 2\right)+1} \mathbf{I}_{2^{s-1}}^{\left(\left\lfloor \frac{i-1}{M} \right\rfloor \bmod 2^{s-1}\right)+1,\left(\left\lfloor \frac{j-1}{M} \right\rfloor \bmod 2^{s-1}\right)+1} \mathbf{I}_{1/M}^{((i-1) \bmod M)+1,((j-1) \bmod)+1}. \tag{34}$$

From this we see immediately that $A_s^{ij} \in \{0, (2M)^{-1}\}$.

By the idempotence, symmetry and the facts that by (34), $A_s^{ii} = (2M)^{-1}$ and $A_s^{ij} \in \{0, (2M)^{-1}\}$ one has

$$\frac{1}{2M} = A_s^{ii} = (A_s A_s)^{ii} = (A_s^T A_s)^{ii} = \sum_{j=1}^{mM} (A_s^{ij})^2 = \frac{u}{(2M)^2} \iff u = 2M,$$

where u is the number of non-zero elements on the ith column of A_s. Hence double stochasticity follows by symmetry.

To prove (6) we observe that by setting $M = 1$ in (34) we have

$$(\mathbf{I}_{2^{S-s}} \otimes \mathbf{1}_{1/2} \otimes \mathbf{I}_{2^{s-1}})^{ij} = \mathbf{I}_{2^{S-s}}^{\left\lfloor \frac{i-1}{2^s} \right\rfloor+1,\left\lfloor \frac{j-1}{2^s} \right\rfloor+1} \mathbf{1}_{1/2}^{\left(\left\lfloor \frac{i-1}{2^{s-1}} \right\rfloor \bmod 2\right)+1,\left(\left\lfloor \frac{j-1}{2^{s-1}} \right\rfloor \bmod 2\right)+1} \mathbf{I}_{2^{s-1}}^{\left((i-1) \bmod 2^{s-1}\right)+1,\left((j-1) \bmod 2^{s-1}\right)+1}.$$

From this, by considering only the diagonal elements of the identity matrices, we have readily that the indices of the nonzero columns of the ith row of $\mathbf{I}_{2^{S-s}} \otimes \mathbf{1}_{1/2} \otimes \mathbf{I}_{2^{s-1}}$ are those $1 \le j \le m$ for which

$$\left\lfloor \frac{i-1}{2^s} \right\rfloor = \left\lfloor \frac{j-1}{2^s} \right\rfloor, \quad \text{and} \quad \left((i-1) \bmod 2^{s-1}\right) = \left((j-1) \bmod 2^{s-1}\right).$$

To prove that this is a superset of (6), suppose that

$$j = \left((i-1) \bmod 2^{s-1}\right) + (q-1)2^{s-1} + 2^s \left\lfloor (i-1)/2^s \right\rfloor + 1, \quad q \in \{1, 2\}. \tag{35}$$

It is then simple to check that $\lfloor (j-1)/2^s \rfloor = \lfloor (i-1)/2^s \rfloor$ and

$$\left((j-1) \bmod 2^{s-1}\right) = j - 1 - \left\lfloor \frac{j-1}{2^{s-1}} \right\rfloor 2^{s-1} = \left((i-1) \bmod 2^{s-1}\right).$$

To prove the converse inclusion, suppose that $\lfloor (i-1)/2^s \rfloor = \lfloor (j-1)/2^s \rfloor$ and $\big((i-1) \bmod 2^{s-1}\big) = \big((j-1) \bmod 2^{s-1}\big)$. Then one can check that

$$j - 1 = \big((i-1) \bmod 2^{s-1}\big) + 2^s \left\lfloor \frac{i-1}{2^s} \right\rfloor + 2^{s-1} \left(\left\lfloor \frac{j-1}{2^{s-1}} \right\rfloor \bmod 2 \right),$$

and since $\big(\lfloor (j-1)/2^{s-1} \rfloor \bmod 2\big) + 1 \in \{1, 2\}$, the claim follows.

Proof of Lemma 2 We prove by induction that

$$\prod_{s=1}^{k} A_s = \mathbf{I}_{2^{S-k}} \otimes \mathbf{1}_{1/2^k} \otimes \mathbf{1}_{1/M} \tag{36}$$

holds for all $1 \le k \le S$. Case $k = S$ then yields the claim. For $k = 1$, (36) holds by definition. Then by assuming that (36) holds for some $1 \le k - 1 < S$ we have

$$\begin{aligned}
\prod_{s=1}^{k} A_s &= \big(\mathbf{I}_{2^{S-k}} \otimes \mathbf{1}_{1/2} \otimes \mathbf{I}_{2^{k-1}} \otimes \mathbf{1}_{1/M}\big)\big(\mathbf{I}_{2^{S-k+1}} \otimes \mathbf{1}_{1/2^{k-1}} \otimes \mathbf{1}_{1/M}\big) \\
&= \big(\big(\mathbf{I}_{2^{S-k}} \otimes \mathbf{1}_{1/2}\big)\mathbf{I}_{2^{S-k+1}}\big) \otimes \big(\big(\mathbf{I}_{2^{k-1}} \otimes \mathbf{1}_{1/M}\big)\big(\mathbf{1}_{1/2^{k-1}} \otimes \mathbf{1}_{1/M}\big)\big) \\
&= \big(\mathbf{I}_{2^{S-k}} \otimes \mathbf{1}_{1/2}\big) \otimes \big(\big(\mathbf{I}_{2^{k-1}} \mathbf{1}_{1/2^{k-1}}\big) \otimes \big(\mathbf{1}_{1/M} \mathbf{1}_{1/M}\big)\big) \\
&= \big(\mathbf{I}_{2^{S-k}} \otimes \mathbf{1}_{1/2}\big) \otimes \big(\mathbf{1}_{1/2^{k-1}} \otimes \mathbf{1}_{1/M}\big) \\
&= \mathbf{I}_{2^{S-k}} \otimes \mathbf{1}_{1/2^k} \otimes \mathbf{1}_{1/M},
\end{aligned}$$

where the 2nd and 3rd equalities follow from the mixed product property of the Kronecker product.

Proof of Lemma 3 From (7) we have $V_0^i = g(\xi_0^i)$ and a proof by induction shows that for $1 \le s \le S$,

$$V_s^{i_s} = \sum_{(i_0, \ldots, i_{s-1})} g(\xi_0^{i_0}) \prod_{q=1}^{s} A_q^{i_q i_{q-1}}, \tag{37}$$

from which (a) follows. Since A_s is row-stochastic, (b) follows from (7). In the case $s = S$, (37) together with Lemma 2 gives

$$V_S^{i_S} = \sum_{(i_0, \ldots, i_{S-1})} g(\xi_0^{i_0}) \prod_{q=1}^{S} A_q^{i_q i_{q-1}} = \sum_{i_0=1}^{N} g(\xi_0^{i_0}) \left[\prod_{q=1}^{S} A_q \right]^{i_S i_0} = \frac{1}{N} \sum_{i_0=1}^{N} g(\xi_0^{i_0}).$$

Proof of Proposition 2 By the definitions of X_ρ and $\mathcal{F}_\rho$ we have (a) by Lemma 3(a). Claim (b) follows from the one step conditional independence and (8). Claim (c) follows from Lemma 3(b), (7), and the row-stochasticity of A_s for all $1 \le s \le S$. It remains to prove (10) and (11).

Since $N^{-1} \sum_{i=1}^{N} V_0^i \overline{\varphi}(\xi_0^i) = 0$, we have the decomposition

$$\frac{1}{N} \sum_{i=1}^{N} V_S^i \overline{\varphi}(\xi_S^i) = \sum_{s=1}^{S} \left(\frac{1}{N} \sum_{i_s=1}^{N} V_s^{i_s} \overline{\varphi}(\xi_s^{i_s}) - \frac{1}{N} \sum_{i_{s-1}=1}^{N} V_{s-1}^{i_{s-1}} \overline{\varphi}(\xi_{s-1}^{i_{s-1}}) \right) \tag{38}$$

Because A_s is doubly stochastic we have $\sum_{j=1}^N A_s^{ji} = 1$ and hence

$$\begin{aligned}\frac{1}{N}\sum_{i_s=1}^N V_s^{i_s}\overline{\varphi}(\xi_s^{i_s}) - \frac{1}{N}\sum_{i_{s-1}=1}^N V_{s-1}^{i_{s-1}}\overline{\varphi}(\xi_{s-1}^{i_{s-1}}) &= \frac{1}{N}\sum_{i_s=1}^N V_s^{i_s}\overline{\varphi}(\xi_s^{i_s}) - \frac{1}{N}\sum_{j=1}^N\sum_{i_{s-1}=1}^N A_s^{ji_{s-1}} V_{s-1}^{i_{s-1}}\overline{\varphi}(\xi_{s-1}^{i_{s-1}}) \\ &= \frac{1}{N}\sum_{i_s=1}^N V_s^{i_s}\left(\overline{\varphi}(\xi_s^{i_s}) - \frac{1}{V_s^{i_s}}\sum_{i_{s-1}=1}^N A_s^{i_s i_{s-1}} V_{s-1}^{i_{s-1}}\overline{\varphi}(\xi_{s-1}^{i_{s-1}})\right) \\ &= \sqrt{\frac{S}{N}}\sum_{i=1}^N X_{(s-1)N+i}.\end{aligned}$$

By substituting the last form into (38) we obtain (10). Finally, since by Lemma 3(c) we have that V_S^i is independent of i, we can prove (11) by writing

$$\begin{aligned}\left(\frac{1}{N}\sum_{i=1}^N g(\xi_0^i)\right)\left(\frac{1}{N}\sum_{i=1}^N \varphi(\xi_S^i)\right) - \frac{1}{N}\sum_{i=1}^N g(\xi_0^i)\varphi(\xi_0^i) &= \frac{1}{N}\sum_{i=1}^N V_S^i\varphi(\xi_S^i) - \frac{1}{N}\sum_{i=1}^N g(\xi_0^i)\varphi(\xi_0^i) \\ &= \frac{1}{N}\sum_{i=1}^N V_S^i\left(\varphi(\xi_S^i) - \frac{\sum_{i=1}^N g(\xi_0^i)\varphi(\xi_0^i)}{\sum_{i=1}^N g(\xi_0^i)}\right) \\ &= \frac{1}{N}\sum_{i=1}^N V_S^i\overline{\varphi}(\xi_S^i).\end{aligned}$$

B | PROOFS FOR SECTION 3

Proof of Lemma 5 We define $\mathcal{M}_{\mathrm{M}} := \{(X_\rho, \mathcal{A}_\rho); 1 \le \rho \le N\}$ as

$$X_\rho := \frac{1}{N}\sum_{i=1}^\rho \left(\varphi\left(\zeta^i\right) - f(\varphi)(\widehat{\zeta}^i)\right), \quad \mathcal{A}_\rho := \sigma\left(\widehat{\zeta}^1, \ldots, \widehat{\zeta}^N, \zeta^1, \ldots, \zeta^\rho\right).$$

Clearly, for all $1 \le \rho \le N$, X_ρ is $\mathcal{A}_\rho$-measurable, $|X_\rho| \le \infty$, and, by (13), $\mathbb{E}[X_\rho \mid \mathcal{A}_{\rho'}] = X_{\rho'}$ for any $1 \le \rho, \rho' \le N$ such that $\rho' < \rho$. Hence $\mathcal{M}_{\mathrm{M}}$ is a martingale and

$$X_N = \frac{1}{N}\sum_{i=1}^N \varphi(\zeta^i) - \frac{1}{N}\sum_{i=1}^N f(\varphi)(\widehat{\zeta}^i).$$

The claim then follows from Burkholder-Davis-Gundy inequality.

Proof of Proposition 3 Throughout the proof we assume $\varphi \in \mathcal{B}_1(\mathbb{X})$. To prove part a), we have by Minkowski's inequality

$$\begin{aligned}\mathbb{E}\left[\left|\widehat{\pi}_n^N(\varphi) - \widehat{\pi}_n(\varphi)\right|^r\right]^{\frac{1}{r}} \le \frac{1}{\pi_n(g_n)}\mathbb{E}\left[\left|\pi_n^N(g_n)\widehat{\pi}_n^N(\varphi) - \pi_n^N(g_n\varphi)\right|^r\right]^{\frac{1}{r}} + \mathbb{E}\left[\left|\frac{\pi_n^N(g_n\varphi)}{\pi_n^N(g_n)} - \frac{\pi_n(g_n\varphi)}{\pi_n(g_n)}\right|^r\right]^{\frac{1}{r}} \\ + \frac{1}{\pi_n(g_n)}\mathbb{E}\left[\left|\widehat{\pi}_n^N(\overline{\varphi}_n)\left(\pi_n(g_n) - \pi_n^N(g_n)\right)\right|^r\right]^{\frac{1}{r}}, \qquad (39)\end{aligned}$$

where $\overline{\varphi}_n := \varphi - \pi_n^N(g_n\varphi)/\pi_n^N(g_n)$. For the first term on the r.h.s. we have by Proposition 1

$$\frac{1}{\pi_n(g_n)}\mathbb{E}\left[\left|\pi_n^N(g_n)\widehat{\pi}_n^N(\varphi) - \pi_n^N(g_n\varphi)\right|^r\right]^{\frac{1}{r}} \leq \frac{2B_r\left\|g_n\right\|}{\pi_n(g_n)}\sqrt{\frac{S}{N}}. \tag{40}$$

For the second term we have (similarly as in the proof of Lemma 4 in (Crisan and Doucet, 2002))

$$\begin{aligned}\mathbb{E}\left[\left|\frac{\pi_n^N(g_n\varphi)}{\pi_n^N(g_n)} - \frac{\pi_n(g_n\varphi)}{\pi_n(g_n)}\right|^r\right]^{\frac{1}{r}} &\leq \frac{\left\|g_n\right\|}{\pi_n(g_n)}\mathbb{E}\left[\left|\pi_n^N\left(\frac{g_n}{\left\|g_n\right\|}\right) - \pi_n\left(\frac{g_n}{\left\|g_n\right\|}\right)\right|^r\right]^{\frac{1}{r}} + \frac{\left\|g_n\right\|}{\pi_n(g_n)}\mathbb{E}\left[\left|\pi_n^N\left(\frac{g_n\varphi}{\left\|g_n\right\|}\right) - \pi_n\left(\frac{g_n\varphi}{\left\|g_n\right\|}\right)\right|^r\right]^{\frac{1}{r}} \\ &\leq \frac{2\left\|g_n\right\|C_{n,r}}{\pi_n(g_n)}\sqrt{\frac{S}{N}}\end{aligned} \tag{41}$$

where the last inequality uses the assumption (15). For the third term on the r.h.s. of (39) we have by (15)

$$\frac{1}{\pi_n(g_n)}\mathbb{E}\left[\left|\widehat{\pi}_n^N(\overline{\varphi}_n)\left(\pi_n(g_n) - \pi_n^N(g_n)\right)\right|^r\right]^{\frac{1}{r}} \leq \frac{2\left\|g_n\right\|C_{n,r}}{\pi_n(g_n)}\sqrt{\frac{S}{N}}. \tag{42}$$

From (39)–(42), part a) follows with $\widehat{C}_{n,r} := (2B_r + 4C_{n,r})\left\|g_n\right\|/\pi_n(g_n)$.

For part b), the case $n = 0$ follows from Lemma 4. For the case $n > 0$ we can write

$$\mathbb{E}\left[\left|\pi_{n+1}^N(\varphi) - \Phi_{n+1}(\pi_n^N)(\varphi)\right|^r\right]^{\frac{1}{r}} \leq \mathbb{E}\left[\left|\pi_{n+1}^N(\varphi) - \widehat{\pi}_n^N(f(\varphi))\right|^r\right]^{\frac{1}{r}} + \mathbb{E}\left[\left|\widehat{\pi}_n^N(f(\varphi)) - \frac{\pi_n^N(g_nf(\varphi))}{\pi_n^N(g_n)}\right|^r\right]^{\frac{1}{r}}.$$

For the first term on the r.h.s. we can use Lemma 5 to obtain an upper bound

$$\sup_{\varphi\in\mathcal{B}_1(\mathbb{X})}\mathbb{E}\left[\left|\pi_{n+1}^N(\varphi) - \widehat{\pi}_n^N(f(\varphi))\right|^r\right]^{\frac{1}{r}} \leq 2B_r\sqrt{\frac{S}{N}}.$$

By (A3), $\left\|g_n\right\|/g_n \leq \delta$ implying $\pi_n^N(g_n)/\left\|g_n\right\| \geq \delta^{-1}$. Hence, by Proposition 1 we have for any $\|\varphi\| \leq 1$

$$\mathbb{E}\left[\left|\widehat{\pi}_n^N(\varphi) - \frac{\pi_n^N(g_n\varphi)}{\pi_n^N(g_n)}\right|^r\right]^{\frac{1}{r}} \leq \frac{\delta}{\left\|g_n\right\|}\mathbb{E}\left[\left|\pi_n^N(g_n)\widehat{\pi}_n^N(\varphi) - \pi_n^N(g_n\varphi)\right|^r\right]^{\frac{1}{r}} \leq 2B_r\delta\sqrt{\frac{S}{N}}.$$

Part b) thus holds with $\widehat{C}_r := \max\left(C_r^*, 2B_r(1+\delta)\right)$.

C | PROOFS FOR SECTION 4

Proof of Proposition 4 The proof is similar to that of Proposition 2. We define a sequence $\breve{\mathcal{M}} := \{(\breve{X}_\rho, \breve{\mathcal{F}}_\rho); 0 \leq \rho \leq N\}$ such that $\breve{X}_0 := 0$, $\breve{\mathcal{F}}_0 := \sigma(\boldsymbol{\xi}_{\text{in}})$ and for all $1 \leq \rho \leq N$

$$\breve{X}_\rho := \frac{W_{\text{out}}^\rho}{\sqrt{N}}\left(\overline{\varphi}(\xi_{\text{out}}^\rho) - \frac{1}{W_{\text{out}}^\rho}\sum_{j=1}^N A^{\rho j}g(\xi_{\text{in}}^j)\overline{\varphi}(\xi_{\text{in}}^j)\right), \qquad \breve{\mathcal{F}}_\rho := \breve{\mathcal{F}}_{\rho-1} \vee \sigma(\xi_{\text{out}}^\rho)$$

where $\overline{\varphi}(x) := \varphi(x) - \pi^N(g\varphi)/\pi^N(g)$. We show that $\breve{\mathcal{M}}$ is a martingale difference.

Clearly $g(\xi^i_{\text{in}})$ is $\mathcal{F}_0$-measurable for all $1 \le i \le N$ and by (17), also W^ρ_{out} is $\mathcal{F}_0$-measurable for all $1 \le \rho \le N$. By the definition of $\check{X}_\rho$ and $\check{\mathcal{F}}_\rho$, $\check{X}_\rho$ is thus $\check{\mathcal{F}}_\rho$-measurable for all $\rho \ge 0$. The requirement that $\mathbb{E}[\check{X}_\rho \mid \check{\mathcal{F}}_{\rho-1}] \overset{\text{a.s.}}{=} 0$ follows from ξ^i_{out} being conditionally independently distributed according to line 4 in Algorithm 5, given $\sigma(\boldsymbol{\xi}_{\text{in}})$. Finally, by (17), and the fact that $g \in \mathcal{B}_+(\mathbb{X})$ we have $|\check{X}_\rho| \le \|g\|\text{osc}(\varphi)/\sqrt{N}$. From these observations we conclude that $\check{\mathcal{M}}$ is a martingale difference.

Next we establish the connection between $\check{\mathcal{M}}$ and the error term in (19). By the double stochasticity of A and the fact that $N^{-1}\sum_{i=1}^N g(\xi^i_{\text{in}})\overline{\varphi}(\xi^i_{\text{in}}) = 0$ we have

$$\frac{1}{\sqrt{N}}\sum_{\rho=0}^N \check{X}_\rho = \frac{1}{N}\sum_{i=1}^N W^i_{\text{out}}\overline{\varphi}(\xi^i_{\text{out}}) - \frac{1}{N}\sum_{i=1}^N\sum_{j=1}^N A^{ij} g(\xi^j_{\text{in}})\overline{\varphi}(\xi^j_{\text{in}}) = \frac{1}{N}\sum_{i=1}^N W^i_{\text{out}}\overline{\varphi}(\xi^i_{\text{out}}) = \pi^N(g)\check{\pi}^N(\varphi) - \pi^N(g\varphi).$$

where the last equality follows from the fact that, by (17), $N^{-1}\sum_{i=1}^N W^i_{\text{out}} = \pi^N(g)$. Part a) of the claim then follows by applying Burkholder-Davis-Gundy inequality to the martingale $\sum_{\rho=0}^N \check{X}_\rho$.

For part b) we define for all $1 \le k \le m$, $\check{\mathcal{M}}_k := \{(\check{X}^k_\rho, \check{\mathcal{F}}^k_\rho); 0 \le \rho \le M\}$ such that $\check{X}^k_0 := 0$, $\check{\mathcal{F}}^k_0 := \sigma(\boldsymbol{\xi}_{\text{in}})$ and for all $1 \le \rho \le M$

$$\check{X}^k_\rho := \frac{W^{(k-1)M+\rho}_{\text{out}}}{\sqrt{M}}\left(\overline{\varphi}_k\left(\xi^{(k-1)M+\rho}_{\text{out}}\right) - \frac{\sum_{j=1}^N A^{(k-1)M+\rho,j} g(\xi^j_{\text{in}})\overline{\varphi}_k(\xi^j_{\text{in}})}{W^{(k-1)M+\rho}_{\text{out}}}\right), \qquad \check{\mathcal{F}}^k_\rho := \check{\mathcal{F}}^k_{\rho-1} \vee \sigma(\xi^{(k-1)M+\rho}_{\text{out}}),$$

where $\overline{\varphi}_k = \varphi - \pi^{M,k}(g\varphi)/\pi^{M,k}(g)$. Similarly as above, we can check that $\check{\mathcal{M}}_k$ is a martingale difference with terms bounded by $|\check{X}^k_\rho| \le \|g\|\text{osc}(\varphi)/\sqrt{M}$.

By the definition of A and $\overline{\varphi}_n$ we have

$$\sum_{j=1}^N A^{(k-1)M+\rho,j} g(\xi^j_{\text{in}})\overline{\varphi}_k(\xi^j_{\text{in}}) = \frac{1}{M}\sum_{j=1}^M g(\xi^{(k-1)M+j}_{\text{in}})\overline{\varphi}_k(\xi^{(k-1)M+j}_{\text{in}}) = 0,$$

which together with (17) enables us to write

$$\frac{1}{\sqrt{M}}\sum_{\rho=0}^M \check{X}^k_\rho = \frac{1}{M}\sum_{\rho=1}^M W^{(k-1)M+\rho}_{\text{out}}\overline{\varphi}_k(\xi^{(k-1)M+\rho}_{\text{out}}) = \pi^{M,k}(g)\left(\check{\pi}^{M,k}(\varphi) - \frac{\pi^{M,k}(g\varphi)}{\pi^{M,k}(g)}\right).$$

The claim then follows by Burkholder-Davis-Gundy inequality as before.

Proof of Proposition 7 By definition, $\widetilde{X}_\rho$ depends only on $\boldsymbol{\xi}^{r(\rho)}_{s(\rho)}$, $\boldsymbol{\xi}^{\ell(\rho)}_{s(\rho)}$, $\overline{V}^{r(\rho)}_{s(\rho)}$, $\overline{V}^{\ell(\rho)}_{s(\rho)}$, as well as $\overline{V}^i_{s(\rho)-1}$ and $\boldsymbol{\xi}^i_{s(\rho)-1}$ for all $1 \le i \le m$. The measurability then follows similarly by the $\sigma(\boldsymbol{\xi}_0)$-measurability of $\overline{V}^i_s$ for all $1 \le s \le S$ and $1 \le i \le m$ as for Proposition 6 by Lemma 3.

Fix $1 \le \rho \le Sm/2$. By definition

$$\overline{V}^{\ell(\rho)}_{s(\rho)} = \overline{V}^{r(\rho)}_{s(\rho)} = \frac{1}{2}\left(\overline{V}^{\ell(\rho)}_{s(\rho)-1} + \overline{V}^{r(\rho)}_{s(\rho)-1}\right)$$

and we can write

$$\frac{1}{\overline{V}^{\ell(\rho)}_{s(\rho)}}\sum_{j=1}^m \overline{A}^{\ell(\rho)j}_{s(\rho)}\overline{V}^j_{s(\rho)-1}\overline{\varphi}(\boldsymbol{\xi}^j_{s(\rho)-1}) = \frac{1}{\overline{V}^{r(\rho)}_{s(\rho)}}\sum_{j=1}^m \overline{A}^{r(\rho)j}_{s(\rho)}\overline{V}^j_{s(\rho)-1}\overline{\varphi}(\boldsymbol{\xi}^j_{s(\rho)-1}) = p_\ell\overline{\varphi}(\boldsymbol{\xi}^{\ell(\rho)}_{s(\rho)-1}) + p_r\overline{\varphi}(\boldsymbol{\xi}^{r(\rho)}_{s(\rho)-1}) \tag{43}$$

where $p_\ell := \frac{1}{2}\overline{V}^{\ell(\rho)}_{s(\rho)-1}/\overline{V}^{\ell(\rho)}_{s(\rho)}$ and $p_r := \frac{1}{2}\overline{V}^{r(\rho)}_{s(\rho)-1}/\overline{V}^{r(\rho)}_{s(\rho)}$ and hence $p_\ell + p_r = 1$. By (23) and (43)

$$\mathbb{E}\left[\widetilde{X}_\rho \,\middle|\, \widetilde{\mathcal{F}}_{\rho-1}\right] = \frac{\overline{V}}{\sqrt{Sm}}\left(\mathbb{E}\left[\overline{\varphi}(\xi^{\ell(\rho)}_{s(\rho)}) \,\middle|\, \widetilde{\mathcal{F}}_{\rho-1}\right] + \mathbb{E}\left[\overline{\varphi}(\xi^{r(\rho)}_{s(\rho)}) \,\middle|\, \widetilde{\mathcal{F}}_{\rho-1}\right] - 2\left(p_\ell \overline{\varphi}(\xi^{\ell(\rho)}_{s(\rho)-1}) + p_r \overline{\varphi}(\xi^{r(\rho)}_{s(\rho)-1})\right)\right),$$

where $\overline{V} := \overline{V}^{\ell(\rho)}_{s(\rho)} = \overline{V}^{r(\rho)}_{s(\rho)}$ and the conditional expectations can be written explicitly by observing the conditional probabilities

$$\begin{aligned}
\mathbb{P}(\xi^{\ell(\rho)}_{s(\rho)} = \xi^{\ell(\rho)}_{s(\rho)-1} \mid \widetilde{\mathcal{F}}_{\rho-1}) &= p_\ell^2 + 2p_r p_\ell, \qquad \mathbb{P}(\xi^{\ell(\rho)}_{s(\rho)} = \xi^{r(\rho)}_{s(\rho)-1} \mid \widetilde{\mathcal{F}}_{\rho-1}) = p_r^2,\\
\mathbb{P}(\xi^{r(\rho)}_{s(\rho)} = \xi^{r(\rho)}_{s(\rho)-1} \mid \widetilde{\mathcal{F}}_{\rho-1}) &= p_r^2 + 2p_r p_\ell, \qquad \mathbb{P}(\xi^{r(\rho)}_{s(\rho)} = \xi^{\ell(\rho)}_{s(\rho)-1} \mid \widetilde{\mathcal{F}}_{\rho-1}) = p_\ell^2,
\end{aligned}$$

and the fact that

$$\mathbb{P}(\xi^{\ell(\rho)}_{s(\rho)} \notin \{\xi^{\ell(\rho)}_{s(\rho)-1}, \xi^{r(\rho)}_{s(\rho)-1}\} \mid \widetilde{\mathcal{F}}_{\rho-1}) = \mathbb{P}(\xi^{r(\rho)}_{s(\rho)} \notin \{\xi^{\ell(\rho)}_{s(\rho)-1}, \xi^{r(\rho)}_{s(\rho)-1}\} \mid \widetilde{\mathcal{F}}_{\rho-1}) = 0.$$

The claim $\mathbb{E}[\widetilde{X}_\rho | \widetilde{\mathcal{F}}_{\rho-1}] = 0$ of part (b) then follows by straightforward calculation.

Part (c) follows from the boundedness of g by the definition of $\widetilde{X}_\rho$ and part (d) follows from observing that

$$\sqrt{\frac{S}{m}}\sum_{\rho=1}^{Sm/2}\widetilde{X}_\rho = \sqrt{\frac{S}{m}}\sum_{s=1}^{S}\sum_{i=1}^{m}\widetilde{X}^i_s$$

where $\widetilde{X}^{i(\rho)}_{s(\rho)}$ is defined in an otherwise identical manner as $\overline{X}_\rho$, except for the law of $\xi^{i(\rho)}_{s(\rho)}$ being different for $\widetilde{X}^{i(\rho)}_{s(\rho)}$ and $\overline{X}_\rho$, which does not affect the validity of (24) and (25).

D | PROOFS FOR SECTION 5

Proof of Propostition 9 We start the proof by introducing two more empirical approximations of $\breve{\pi}_n$

$$\breve{\pi}^{M,k}_n := \frac{\sum_{i=1}^{M}\breve{W}^{(k-1)M+i}_n \delta_{\breve{\zeta}^{(k-1)M+i}_n}}{\sum_{i=1}^{M}\breve{W}^{(k-1)M+i}_n}, \quad \breve{\pi}^{N}_n := \frac{\sum_{i=1}^{N}\breve{W}^{i}_n \delta_{\breve{\zeta}^{i}_n}}{\sum_{i=1}^{N}\breve{W}^{i}_n} \quad 1 \le k \le m, \tag{44}$$

based on the samples $\breve{\zeta}^k_n := (\breve{\zeta}^{(k-1)M+1}_n, \ldots, \breve{\zeta}^{kM}_n)$ and $\breve{\zeta}_n := (\breve{\zeta}^1_n, \ldots, \breve{\zeta}^N_n)$, respectively. Note that $\breve{\pi}^N_n$ is the approximation of $\widehat{\pi}_n$ based on the entire sample after the first within island resampling while $\breve{\pi}^{M,k}_n$ is the PE specific approximation based on the sample contained in kth PE.

By Minkowski's inequality we have

$$\mathbb{E}\left[\left|\widehat{\pi}^{M,k}_n(\varphi) - \widehat{\pi}_n(\varphi)\right|^r\right]^{\frac{1}{r}} \le \mathbb{E}\left[\left|\widehat{\pi}^{M,k}_n(\varphi) - \breve{\pi}^{M,k}_n(\varphi)\right|^r\right]^{\frac{1}{r}} + \mathbb{E}\left[\left|\breve{\pi}^{M,k}_n(\varphi) - \frac{\pi^{M,k}_n(g_n\varphi)}{\pi^{M,k}_n(g_n)}\right|^r\right]^{\frac{1}{r}} + \mathbb{E}\left[\left|\frac{\pi^{M,k}_n(g_n\varphi)}{\pi^{M,k}_n(g_n)} - \frac{\pi_n(g_n\varphi)}{\pi_n(g_n)}\right|^r\right]^{\frac{1}{r}} \tag{45}$$

For the last term on the r.h.s. we have by using (27), similarly as in equation (41) in the proof of Proposition 3,

$$\mathbb{E}\left[\left|\pi_n^{M,k}(g_n\varphi)/\pi_n^{M,k}(g_n) - \pi_n(g_n\varphi)/\pi_n(g_n)\right|^r\right]^{\frac{1}{r}} \leq \frac{2\left\|g_n\right\| C_{n,r}}{\pi_n(g_n)\sqrt{M}}. \tag{46}$$

The remainder of the proof hinges on expressing the first two terms in the r.h.s. of (45) in terms of expectations such as the one in (46) as well as expectations of the form $\mathbb{E}[|\breve{\pi}_n^{M,i}(\overline{\varphi}_n^i)|^r]$ where $\overline{\varphi}_n^i := \varphi - \pi_n^{M,i}(g_n\varphi)/\pi_n^{M,i}(g_n)$ and $\mathbb{E}[|\breve{\pi}_n^N(\overline{\varphi}_n)|^r]$ where $\overline{\varphi}_n := \varphi - \pi_n^N(g_n\varphi)/\pi_n^N(g_n)$. The last two types of expectations can be bounded by using the identities

$$\begin{aligned}
\breve{\pi}_n^{M,i}(\overline{\varphi}_n) &= \frac{\pi_n^{M,i}(g_n)\left(\breve{\pi}_n^{M,i}(\varphi) - \frac{\pi_n^{M,i}(g_n\varphi)}{\pi_n^{M,i}(g_n)}\right)}{\pi_n(g_n)} + \frac{\breve{\pi}_n^{M,i}(\overline{\varphi}_n)}{\pi_n(g_n)}\left(\pi_n(g_n) - \pi_n^{M,i}(g_n)\right) \\
\breve{\pi}_n^{N}(\overline{\varphi}_n) &= \frac{\pi_n^{N}(g_n)\breve{\pi}_n^{N}(\varphi) - \pi_n^{N}(g_n\varphi)}{\pi_n(g_n)} + \frac{\breve{\pi}_n^{N}(\overline{\varphi}_n)}{\pi_n(g_n)}\left(\pi_n(g_n) - \pi_n^{N}(g_n)\right)
\end{aligned} \tag{47}$$

together with Proposition 4 and assumption (27).

Clearly, the second term on the r.h.s. of (45) is readily of the desired form and therefore we only need to consider the first term. Again, by applying Minkowski's inequality, we have

$$\begin{aligned}
\mathbb{E}[|\widehat{\pi}_n^{M,k}(\varphi) - \breve{\pi}_n^{M,k}(\varphi)|^r]^{1/r} \leq{}& \mathbb{E}[|\widehat{\pi}_n^{M,k}(\varphi) - \breve{\pi}_n^{N}(\varphi)|^r]^{1/r} + \mathbb{E}[|\breve{\pi}_n^{N}(\overline{\varphi}_n)|^r]^{1/r} + \mathbb{E}[|\breve{\pi}_n^{M,k}(\overline{\varphi}_n^k)|^r]^{1/r} \\
&+ \mathbb{E}[|\pi_n^{M,k}(g_n\varphi)/\pi_n^{M,k}(g_n) - \pi_n^{N}(g_n\varphi)/\pi_n^{N}(g_n)|^r]^{1/r}
\end{aligned} \tag{48}$$

If we write $P_i := \mathbb{P}(\widehat{\pi}_n^{M,k} = \breve{\pi}_n^{M,i} \mid \breve{\boldsymbol{\zeta}}_n, \boldsymbol{\zeta}_n)$, where $1 \leq i \leq m$, then by the tower property of conditional expectations, Cauchy-Schwartz, Jensen's and Minkowski's inequalities we have

$$\begin{aligned}
\mathbb{E}\left[\left|\widehat{\pi}_n^{M,k}(\varphi) - \breve{\pi}_n^{N}(\varphi)\right|^r\right]^{\frac{1}{r}} &= \mathbb{E}\left[\sum_{i=1}^m P_i\left|\breve{\pi}_n^{M,i}(\varphi) - \breve{\pi}_n^{N}(\varphi)\right|^r\right]^{\frac{1}{r}} \\
&\leq \sum_{i=1}^m \mathbb{E}\left[\left|\breve{\pi}_n^{M,i}(\varphi) - \breve{\pi}_n^{N}(\varphi)\right|^{2r}\right]^{\frac{1}{2r}} \\
&\leq m\mathbb{E}\left[\left|\breve{\pi}_n^{N}(\overline{\varphi}_n)\right|^{2r}\right]^{\frac{1}{2r}} + \sum_{i=1}^m \mathbb{E}\left[\left|\breve{\pi}_n^{M,i}(\overline{\varphi}_n^i)\right|^{2r}\right]^{\frac{1}{2r}} + \sum_{i=1}^m \mathbb{E}\left[\left|\frac{\pi_n^{M,i}(g_n\varphi)}{\pi_n^{M,i}(g_n)} - \frac{\pi_n^{N}(g_n\varphi)}{\pi_n^{N}(g_n)}\right|^{2r}\right]^{\frac{1}{2r}}
\end{aligned} \tag{49}$$

In (48) and (49) all terms are of the desired form except for the terms $\mathbb{E}[|\pi_n^{M,i}(g_n\varphi)/\pi_n^{M,i}(g_n) - \pi_n^{N}(g_n\varphi)/\pi_n^{N}(g_n)|^r]^{1/r}$ where $1 \leq i \leq m$, but for these terms we have for all $1 \leq i \leq m$

$$\mathbb{E}\left[\left|\frac{\pi_n^{M,i}(g_n\varphi)}{\pi_n^{M,i}(g_n)} - \frac{\pi_n^{N}(g_n\varphi)}{\pi_n^{N}(g_n)}\right|^r\right]^{\frac{1}{r}} \leq \mathbb{E}\left[\left|\frac{\pi_n^{M,i}(g_n\varphi)}{\pi_n^{M,i}(g_n)} - \frac{\pi_n(g_n\varphi)}{\pi_n(g_n)}\right|^r\right]^{\frac{1}{r}} + \mathbb{E}\left[\left|\frac{\pi_n^{N}(g_n\varphi)}{\pi_n^{N}(g_n)} - \frac{\pi_n(g_n\varphi)}{\pi_n(g_n)}\right|^r\right]^{\frac{1}{r}} \leq \frac{4\left\|g_n\right\| C_{n,r}}{\pi_n(g_n)\sqrt{M}}$$

where the final inequality follows similarly as in (46) by using (27). By applying (47) together with Proposition 4 and (27) then claim then follows with

$$\widehat{C}_{n,r} = 2\left(\left(1 + \frac{1}{\sqrt{m}}\right)mB_{2r} + \left(2 + \frac{1}{\sqrt{m}}\right)B_r + 6C_{n,r} + 4mC_{n,2r}\right)\frac{\left\|g_n\right\|}{\pi_n(g_n)}.$$

Proof of Propostion 10 By Minkowski's inequality we have

$$\mathbb{E}\left[\left|\widehat{\pi}_n^N(\varphi)-\widehat{\pi}_n(\varphi)\right|^r\right]^{\frac{1}{r}} \leq \mathbb{E}\left[\left|\widehat{\pi}_n^N(\varphi)-\breve{\pi}_n^N(\varphi)\right|^r\right]^{\frac{1}{r}} + \mathbb{E}\left[\left|\breve{\pi}_n^N(\varphi)-\frac{\pi_n^N(g_n\varphi)}{\pi_n^N(g_n)}\right|^r\right]^{\frac{1}{r}} + \mathbb{E}\left[\left|\frac{\pi_n^N(g_n\varphi)}{\pi_n^N(g_n)}-\frac{\pi_n(g_n\varphi)}{\pi_n(g_n)}\right|^r\right]^{\frac{1}{r}}, \quad (50)$$

For the third term on the r.h.s. of (50) we have similarly as in the proof of Proposition 9

$$\mathbb{E}[|\pi_n^N(g_n\varphi)/\pi_n^N(g_n)-\pi_n(g_n\varphi)/\pi_n(g_n)|^r]^{1/r} \leq \frac{2\left\|g_n\right\| C_{n,r}}{\pi_n(g_n)}\sqrt{\frac{S}{m}}. \quad (51)$$

It remains to consider the first two terms on the r.h.s. of (50).

Let us write, analogously to Proposition 5,

$$\boldsymbol{g}_n(\boldsymbol{x}) := \frac{1}{M}\sum_{j=1}^{M} g_n(x^i),$$

for all $\boldsymbol{x} = (x^1,\dots,x^M) \in \mathbb{X}^M$. By the definition of the weights $(\widetilde{W}_n^1,\dots,\widetilde{W}_n^N)$ in (17) we see that for all $1 \leq i \leq m$, $\widetilde{W}_n^{(i-1)M+1} = \boldsymbol{g}_n(\breve{\boldsymbol{\zeta}}_n^i) = \boldsymbol{g}_n(\boldsymbol{\zeta}_n^i)$ and $\sum_{i=1}^N \widetilde{W}_n^i = \sum_{i=1}^N g_n(\zeta_n^i)$. This enables us to write

$$\breve{\pi}_n^N(\varphi) = \frac{\frac{1}{m}\sum_{i=1}^m \widetilde{W}_n^{(i-1)M+1}\frac{1}{M}\sum_{j=1}^M \varphi(\breve{\zeta}_n^{(i-1)M+j})}{\frac{1}{N}\sum_{i=1}^N g_n(\zeta_n^i)} = \frac{1}{\pi_n^N(g_n)}\frac{1}{m}\sum_{i=1}^m \boldsymbol{g}_n(\breve{\boldsymbol{\zeta}}_n^i)\boldsymbol{\varphi}(\breve{\boldsymbol{\zeta}}_n^i),$$

yielding the identity

$$\widehat{\pi}_n^N(\varphi)-\breve{\pi}_n^N(\varphi) = \frac{1}{\pi_n(g_n)}\left(\pi_n^N(g_n)\widehat{\pi}_n^N(\varphi)-\frac{1}{m}\sum_{i=1}^m \boldsymbol{g}_n(\breve{\boldsymbol{\zeta}}_n^i)\boldsymbol{\varphi}(\breve{\boldsymbol{\zeta}}_n^i)\right) + \frac{1}{\pi_n(g_n)}\left(\breve{\pi}_n^N(\varphi)-\widehat{\pi}_n^N(\varphi)\right)\left(\pi_n^N(g_n)-\pi_n(g_n)\right),$$

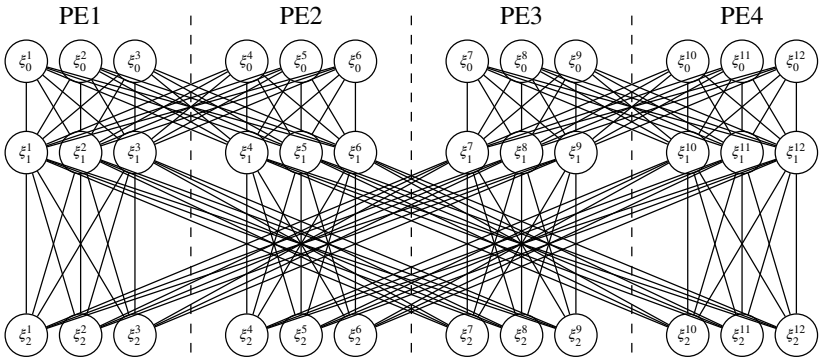
which, together with Proposition 5 and (30) yields

$$\mathbb{E}[|\widehat{\pi}_n^N(\varphi)-\breve{\pi}_n^N(\varphi)|^r]^{1/r} \leq 2\left(B_r + C_{n,r}\right)\frac{\left\|g_n\right\|}{\pi_n(g_n)}\sqrt{\frac{S}{m}}.$$

For the second term on the r.h.s. of (50) we have, similarly as in the proof of Proposition 9, by using Proposition 5

$$\mathbb{E}\left[\left|\breve{\pi}_n^N(\varphi)-\pi_n^N(g_n\varphi)/\pi_n^N(g_n)\right|^r\right]^{\frac{1}{r}} \leq \left(\frac{B_r}{\sqrt{M}}+C_{n,r}\right)\frac{2\left\|g_n\right\|}{\pi_n(g_n)}\sqrt{\frac{S}{m}}. \quad (52)$$

The claim then follows with $\widehat{C}_{n,r} = ((2+2/\sqrt{M})B_r + 6C_{n,r})\left\|g_n\right\|/\pi_n(g_n)$.



$$A_1 = \frac{1}{2}\begin{bmatrix} \mathbf{1}_{1/3} & \mathbf{1}_{1/3} & 0 & 0 \\ \mathbf{1}_{1/3} & \mathbf{1}_{1/3} & 0 & 0 \\ 0 & 0 & \mathbf{1}_{1/3} & \mathbf{1}_{1/3} \\ 0 & 0 & \mathbf{1}_{1/3} & \mathbf{1}_{1/3} \end{bmatrix}, \quad A_2 = \frac{1}{2}\begin{bmatrix} \mathbf{1}_{1/3} & 0 & \mathbf{1}_{1/3} & 0 \\ 0 & \mathbf{1}_{1/3} & 0 & \mathbf{1}_{1/3} \\ \mathbf{1}_{1/3} & 0 & \mathbf{1}_{1/3} & 0 \\ 0 & \mathbf{1}_{1/3} & 0 & \mathbf{1}_{1/3} \end{bmatrix}$$

FIGURE 1 Augmented resampling with $m = 4$ and $M = 3$. Dashed vertical lines separate the groups of particles belonging to different PEs.

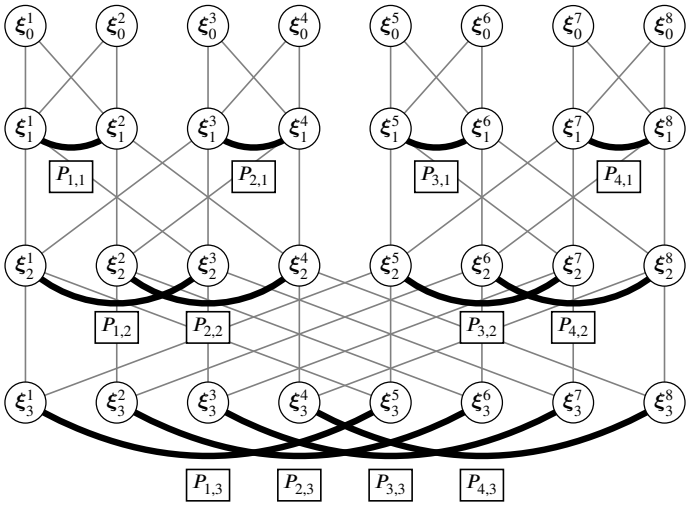


FIGURE 2 Illustration of the paired PEs at each stage of the augemented island resampling in the case $m = 8$.

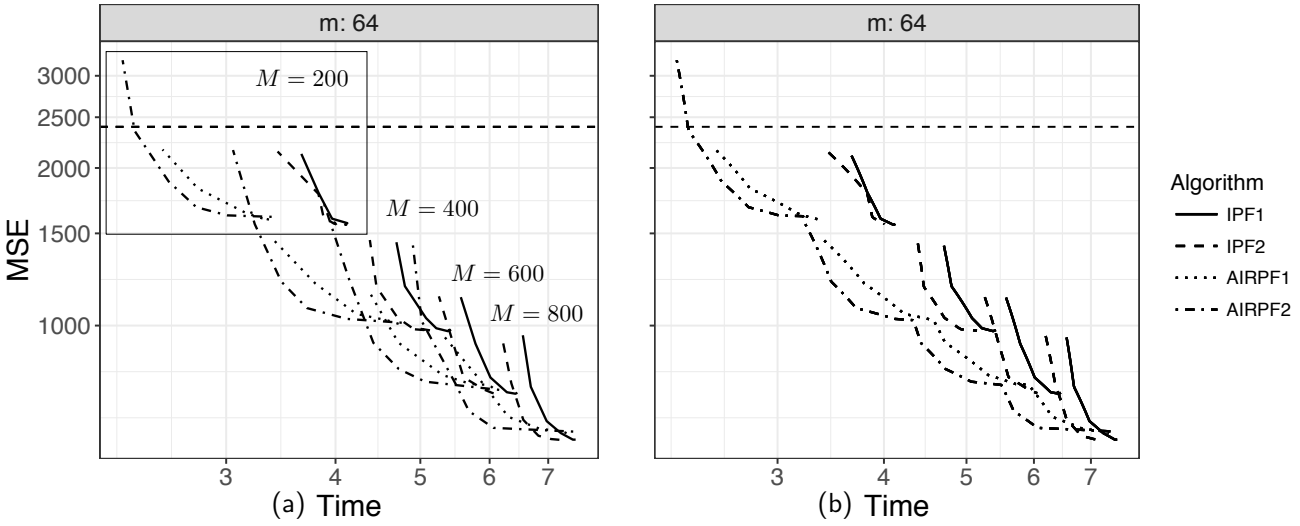


FIGURE 3 (a) MSE vs. time plots for $M \in \{200, 400, 600, 800\}$, $m = 64$, and $\theta = \{.1, .2, .4, .6, .8, 1\}$. For $M = 200$ the curves obtained with different algorithms are highlighted by a rectangle. The dashed horizontal line at 2396 represents the raw MSE obtained by using the plain observations to approximate the mean of the filtering distribution. (b) The lower envelopes of the curves in (a).

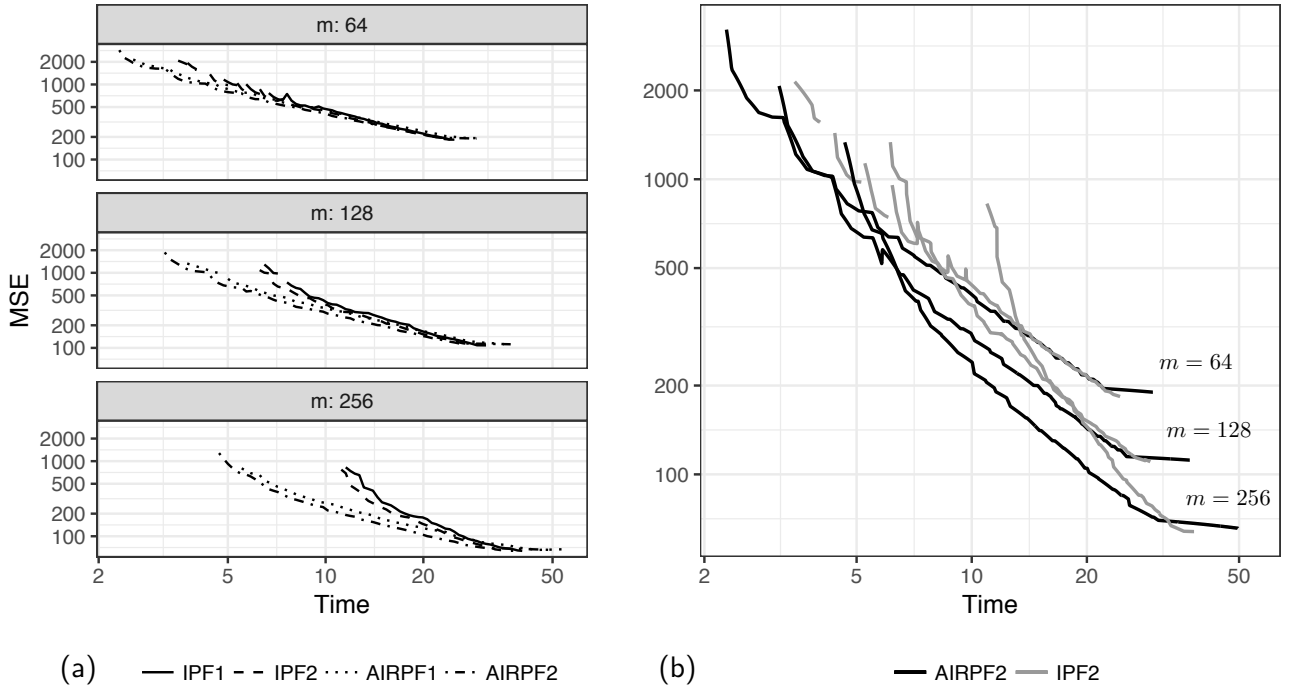


FIGURE 4 (a) The lower envelopes of MSE vs. time for all algorithms and entire ranges of M and m. (b) The lower envelopes of MSE vs. time for AIRPF2 and IPF2 and entire ranges of M and m.

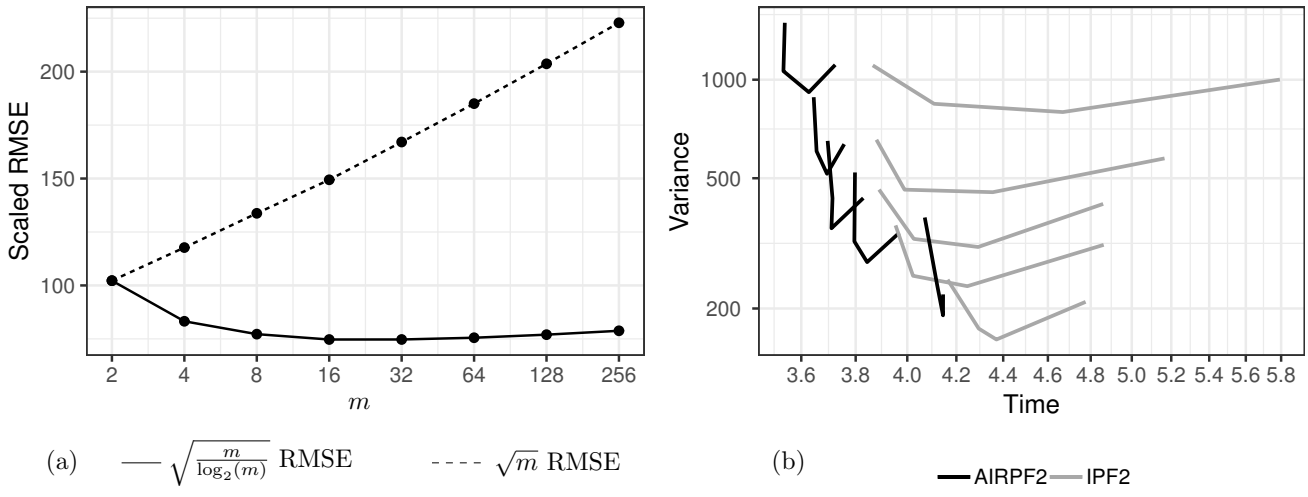


FIGURE 5 a) The scaled RMSE for AIRPF1. The nearly constant solid curve is consistent with the $\sqrt{\log_2(m)/m}$ rate of convergence. b) The integrated filter variances of AIRPF2 and IPF2 for one dimension of the prokaryotic autoregulation model.