Bernoulli **25**(1), 2019, 584–622
https://doi.org/10.3150/17-BEJ999

Sequential Monte Carlo as approximate sampling: bounds, adaptive resampling via ∞-ESS, and an application to particle Gibbs

JONATHAN H. HUGGINS[1] and DANIEL M. ROY[2]

[1]*Computer Science and Artificial Intelligence Laboratory (CSAIL), Massachusetts Institute of Technology, Cambridge, MA 02139, USA. E-mail: jhuggins@mit.edu*
[2]*Department of Statistical Sciences, University of Toronto, Toronto, ON M5S 3G3, Canada. E-mail: droy@utstat.toronto.edu*

Sequential Monte Carlo (SMC) algorithms were originally designed for estimating intractable conditional expectations within state-space models, but are now routinely used to generate approximate samples in the context of general-purpose Bayesian inference. In particular, SMC algorithms are often used as subroutines within larger Monte Carlo schemes, and in this context, the demands placed on SMC are different: control of mean-squared error is insufficient—one needs to control the divergence from the target distribution directly. Towards this goal, we introduce the *conditional* adaptive resampling particle filter, building on the work of Gordon, Salmond, and Smith (1993), Andrieu, Doucet, and Holenstein (2010), and Whiteley, Lee, and Heine (2016). By controlling a novel notion of effective sample size, the ∞-ESS, we establish the efficiency of the resulting SMC sampling algorithm, providing an adaptive resampling extension of the work of Andrieu, Lee, and Vihola (2018). We apply our results to arrive at new divergence bounds for SMC samplers with adaptive resampling as well as an adaptive resampling version of the Particle Gibbs algorithm with the same geometric-ergodicity guarantees as its nonadaptive counterpart.

Keywords: adaptive resampling; effective sample size; geometric ergodicity; particle Gibbs; sequential Monte Carlo; state-space models; uniform ergodicity

1. Introduction

Sequential Monte Carlo (SMC) methods are a popular class of algorithms for approximate inference [8–11,14,15]. In the context of Bayesian inference, SMC produces a particle approximation to the posterior distribution as well as an unbiased estimate of the marginal likelihood. Traditionally, particle approximations were built to estimate conditional expectations, and the analysis of SMC methods focused on this *operator* perspective, by bounding the mean squared error of the resulting estimates.

Increasingly, SMC methods are being used to produce approximate samples, usually in the inner loop of another approximate inference algorithm. A key example is the class of particle Markov chain Monte Carlo (PMCMC) methods, which aim to combine the best features of SMC and MCMC approaches by using SMC as a proposal mechanism for a Metropolis–Hastings ("particle MH") or approximate Gibbs ("particle Gibbs") sampler [1,13]. Characterizing the efficiency of PMCMC methods is an active area of investigation [2–5,16,17].

1350-7265 © 2019 ISI/BS

When SMC methods are employed for sampling, convergence guarantees from the operator perspective are insufficient. In this work, we take up the *sampling* perspective, and study the distribution of a sample drawn from the SMC particle filter approximation. Building off the work of Andrieu et al. [1,2], Gordon et al. [12] and Whiteley et al. [18], we use conditional filters to derive a minorization condition, lower bounding the density of the approximate sample's distribution in terms of that of the target distribution. The analysis extends to conditional SMC as well as to adaptive resampling versions. One of our key contributions is a novel notion of effective sample size, the ∞-ESS, which we use to establish the efficiency of the adaptive SMC sampling algorithm. Thus, our results are both a sampling analogue to the operator work of Whiteley et al. [18] and an adaptive resampling extension to the sampling work of Andrieu et al. [2]. We apply our results to arrive at new divergence bounds for SMC samplers with adaptive resampling as well as an adaptive resampling version of the Particle Gibbs algorithm with the same geometric-ergodicity guarantees as its nonadaptive counterpart.

In the remainder of this section, we provide an overview of our contributions for the special case of the the conditional adaptive resampling particle filter: we introduce the conditional adaptive resampling particle filter, present our main theoretical results characterizing its performance, and describe an application to a novel adaptive resampling Particle Gibbs algorithm.

1.1. (Conditional) adaptive resampling particle filters

We follow a similar setup and notation to Del Moral [7]. Let $(\xi_t)_{t\geq 1}$ be an inhomogeneous Markov chain on the measurable space $(E, \mathcal{E})$ with transition kernels $(M_t)_{t\geq 2}$ and initial distribution M_1. Denote expectations with respect to the Markov chain by $\mathbb{E}[\cdot]$. Let $g_t : E \to \mathbb{R}_+$, for $t \geq 1$, be a sequence of $\mathcal{E}$-measurable potential functions on E, let $g_0 \equiv 1$, and write $g_{s:t}(x_{s:t}) \triangleq \prod_{\tau=s}^{t} g_\tau(x_\tau)$. For $t = 1, 2, \ldots$, define the measure $\pi_{1:t}$ on E^t given by

$$\pi_{1:t}(\mathrm{d}x_{1:t}) \triangleq \gamma_{1:t}(\mathrm{d}x_{1:t})/Z_t,$$

where

$$\gamma_{1:t}(\mathrm{d}x_{1:t}) \triangleq \prod_{s=1}^{t} g_s(x_s) M_s(x_{s-1}, \mathrm{d}x_s) \quad \text{and} \quad Z_t \triangleq \gamma_{1:t}(1).$$

(We have written $M_1(x_0, \mathrm{d}x_1)$ for $M_1(\mathrm{d}x_1)$.) Equivalently,

$$\pi_{1:t}(\phi) \triangleq \frac{\mathbb{E}[\phi(\xi_{1:t}) g_{1:t}(\xi_{1:t})]}{Z_t}, \qquad \phi\colon E^t \to \mathbb{R} \text{ measurable},$$

where $Z_t \triangleq \mathbb{E}[g_{1:t}(\xi_{1:t})]$ is the normalization constant.[1]

[1] In the state-space setting, the potential g_t would be the conditional density (i.e., likelihood) of the observation v_t at time t as a function of unobserved state x_t: i.e., $g_t(x_t) = p_t(v_t \mid x_t)$. Then $\pi_{1:t}$ would be the posterior distribution of the unobserved state sequence given the observed sequence.

Towards the goal of efficiently approximating $\pi_{1:t}$, we introduce a novel sequential Monte Carlo algorithm: the i-times conditional adaptive resampling particle filter (c^{i}ARPF), which is a generalization of the adaptive resampling particle filter [12,15] and the conditional SMC algorithm used in particle Gibbs [1]. (In Section 3, we will introduce a further generalization.) The integer parameter $i \geq 0$ determines the number of fixed trajectories $y^1_{1:t}, \ldots, y^i_{1:t} \in E^t$ required by the algorithm, which operates by generating a collection $\tilde{\boldsymbol{X}}^{1:N}_{1:t} \triangleq \{\tilde{X}^n_{1:t}\}_{n=1}^N$ of $N > i$ particles with corresponding nonnegative weights $W^{1:N}_t \triangleq \{W^n_t\}_{n=1}^N$. When $i = 0$, we recover the standard (unconditional) adaptive resampling particle filter; when $i = 1$, we recover a generalization of the conditional SMC algorithm that includes adaptive resampling. For time $s = 1, \ldots, t$, the measure $\pi_{1:s}$ is approximated by

$$\pi^{i,N}_{1:s} \triangleq \sum_{n=1}^N \frac{W^n_s g_s(X^n_s)}{\sum_{k=1}^N W^k_s g_s(X^k_s)} \delta_{\tilde{X}^n_{1:s}}.$$

The c^{i}ARPF algorithm iteratively constructs $\tilde{\boldsymbol{X}}^{1:N}_{1:t}$ and $W^{1:N}_t$ as follows: The first i particles are deterministically set to match the fixed trajectories:

$$X^n_s = y^n_s, \qquad \tilde{X}^n_{1:s} = y^n_{1:s}, \qquad s = 1, \ldots, t \text{ and } n = 1, \ldots, i.$$

At time $s = 1$, the remaining $N - i$ particles X^n_1, for $n = i + 1, \ldots, N$, are sampled independently and identically from M_1. The corresponding (length 1) trajectories are

$$\tilde{X}^n_{1:1} = X^n_1, \qquad n = i + 1, \ldots, N.$$

Furthermore, for all $n = 1, \ldots, N$, $W^n_1 = 1$.

The remaining particle trajectories are generated as follows: First, we introduce a cutoff parameter $\eta \in [0, 1]$ and an effective sample size (ESS) function $\mathrm{ESS} : \mathbb{R}^N_+ \to [1, N]$. The ESS function measures how uniform the current weights $W^{1:N}_s$ are. Typically $\mathrm{ESS}(W^{1:N}_s) = 1$ indicates that all but one weight is zero and $\mathrm{ESS}(W^{1:N}_s) = N$ indicates all the weights are equal.

For each time $s = 2, \ldots, t$:

- If $\mathrm{ESS}(W^{1:N}_{s-1}) \leq \eta N$, a *resampling* step is introduced. For $n = 1, \ldots, N$, the weights are set to a common value

$$W^n_s = W_s \triangleq \frac{1}{N} \sum_{k=1}^N W^k_{s-1} g_{s-1}\big(X^k_{s-1}\big)$$

  and, for $n = i + 1, \ldots, N$, particle n's "ancestor" at time s, denoted A^n_s, is sampled independently, such that $A^n_s = k$, for $k = 1, \ldots, N$, with probability

$$\frac{W^k_{s-1} g_{s-1}(X^k_{s-1})}{N\, W_s}.$$

- If $\mathrm{ESS}(W_{s-1}^{1:N}) > \eta N$, then the algorithm does not resample the particles. For $n = 1, \dots, N$, the weights are copied, that is,

$$W_s^n = W_{s-1}^n g_{s-1}\big(X_{s-1}^n\big),$$

  and, for $n = i+1, \dots, N$, a record is made that particle n was its own ancestor by setting $A_s^n = n$.
- Having sampled the ancestors, the algorithm propagates the particles forward. For $n = i+1, \dots, N$, X_s^n is sampled from $M_s(X_{s-1}^{A_s^n}, \cdot)$, and the corresponding trajectories are set to

$$\tilde{X}_{1:s}^n = \big\langle \tilde{X}_{1:s-1}^{A_s^n}, X_s^n \big\rangle.$$

In the final step of the algorithm, a single particle $\tilde{X}_{1:t}^*$ is sampled from the full approximation $\pi_{1:t}^{i,N}$, and the algorithm yields an estimate $\hat{Z}_t$ of the normalization constant Z_t, where

$$\hat{Z}_t \triangleq \frac{1}{N} \sum_{n=1}^{N} W_t^n g_t\big(X_t^n\big).$$

Let $\mathbb{E}_{\boldsymbol{y}_{1:t}^{1:i}}^{i,N}[\cdot]$ denote the expectation operator with respect to the c^{i}ARPF, and write

$$P^{i,N}\big(\boldsymbol{y}_{1:t}^{1:i}, \mathrm{d}x_{1:t}\big) \triangleq P_{\boldsymbol{y}_{1:t}^{1:i}}^{i,N}(\mathrm{d}x_{1:t}) \triangleq \mathbb{E}_{\boldsymbol{y}_{1:t}^{1:i}}^{i,N}\big[\delta_{\tilde{X}_{1:t}^*}(\mathrm{d}x_{1:t})\big]$$

for the law of $\tilde{X}_{1:t}^*$ when the i fixed trajectories are $\boldsymbol{y}_{1:t}^{1:i} \in (E^t)^i$. We can now describe in more precise terms how the c^{i}ARPF kernel $P^{i,N}$ generalizes several well-known SMC kernels. When $i = 0$, $\pi_{1:t}^{0,N}$ is the standard adaptive SMC particle approximation of $\pi_{1:t}$ and $\tilde{X}_{1:t}^*$ is a single sample from the particle approximation. When $i = 1$ and resampling is done at every step by taking $\eta = 1$, $P^{1,N}$ is exactly the conditional SMC kernel used in particle Gibbs samplers [1, 2,13]. For general $\eta \in (0, 1)$, we obtain a novel adaptive resampling variant that we study in the sequel. In particular, under mild regularity conditions, $P^{1,N}$ defines a Markov kernel with invariant distribution $\pi_{1:t}$.

1.2. Controlling c^{i}ARPF efficiency with ∞-ESS

We can analyze the quality of the c^{i}ARPF kernel $P^{i,N}$ by quantifying the extent to which high-probability sets under the target distribution also have high probability under the kernel. The following theorem establishes a minorization condition for the i-times conditional filter in terms of the $(i+1)$-times conditional filter.

Theorem 1.1. *For all $t \geq 1$, $i \geq 0$, $N > i$, and $y_{1:t}^1, \dots, y_{1:t}^i \in E^t$,*

$$P^{i,N}\big(\boldsymbol{y}_{1:t}^{1:i}, S\big) \geq (1 - i/N)^t \int_S \frac{Z_t}{\mathbb{E}_{\boldsymbol{y}_{1:t}^{1:i+1}}^{i+1,N}[\hat{Z}_t]} \pi_{1:t}\big(\mathrm{d}y_{1:t}^{i+1}\big), \qquad S \subseteq E^t \text{ measurable}. \tag{1.1}$$

The integral appearing in equation (1.1) has no simple form in general, but in many settings we will be able to obtain a lower bound on the integrand that does not depend on the fixed trajectories $\boldsymbol{y}_{1:t}^{1:i+1}$. In those cases, the integral is simply replaced by this uniform lower bound. For $i = 0$, we are then immediately able to control numerous measures of divergence between $\pi_{1:t}$ and $P^{0,N}$ (i.e., the law of $\tilde{X}_{1:t}^{*}$). For example, in the case of total variation distance, we have the following corollary to Theorem 1.1.

Corollary 1.2. *If $\mathbb{E}_{y_{1:t}}^{1,N}[\hat{Z}_t/Z_t] \leq B_{t,N}$ for all $y_{1:t} \in E^t$, then*

$$d_{\mathrm{TV}}\big(\pi_{1:t}, P^{0,N}\big) \leq 1 - B_{t,N}^{-1}.$$

For $i = 1$, a uniform lower bound assumption implies a minorization condition on the kernel $P^{1,N}(y_{1:t}, \mathrm{d}x_{1:t})$, which in turn implies fast mixing of the Markov chain with kernel $P^{1,N}$.

Corollary 1.3. *If $\mathbb{E}_{\boldsymbol{y}_{1:t}^{1:2}}^{2,N}[\hat{Z}_t/Z_t] \leq B_{t,N}$ for all $y_{1:t}^1, y_{1:t}^2 \in E^t$, then the Markov chain with transition kernel $P^{1,N}(y_{1:t}, \mathrm{d}x_{1:t})$ is uniformly ergodic in total variation distance and has invariant distribution $\pi_{1:t}$. In particular, for all $y_{1:t} \in E^t$ and $k \geq 1$,*

$$d_{\mathrm{TV}}\big(\tilde{\pi}_{y_{1:t}}^{k}, \pi_{1:t}\big) \leq \big(1 - B_{t,N}^{-1}(1 - 1/N)^t\big)^k,$$

where $\tilde{\pi}_{y_{1:t}}^{k} \triangleq \delta_{y_{1:t}}[P^{1,N}]^k$ is the law of the Markov chain, with initial state $y_{1:t}$, after k transitions.

In order to apply the corollaries, it remains to bound $\mathbb{E}_{\boldsymbol{y}_{1:t}^{1:i}}^{i,N}[\hat{Z}_t/Z_t]$. Such a bound was obtained for the *nonadaptive* conditional SMC kernel in Andrieu et al. [2]. However, in our general adaptive resampling setting, one must make a careful choice of effective sample size function. To this end, we introduce a generalized notion of effective sample size, which includes several existing definitions as special cases. For $p \in (1, \infty]$, let $p_* \triangleq \frac{p}{p-1}$ be the conjugate exponent of p (so $1/p + 1/p_* = 1$). The *p-effective sample size* (p-ESS) of the weight vector $w^{1:N} \in \mathbb{R}_+^N$, $\|w^{1:N}\|_1 > 0$, is

$$\mathrm{ESS}_p\big(w^{1:N}\big) \triangleq \frac{\|w^{1:N}\|_1^{p_*}}{\|w^{1:N}\|_p^{p_*}}.$$

The following proposition highlights some elementary properties of p-ESS.

Proposition 1.4. *The p-ESS has the following properties*:

1. *For all $p \in (1, \infty]$, $1 \leq \mathrm{ESS}_p(w^{1:N}) \leq N$. The lower bound is achieved if and only if all but one of the weights is zero. The upper bound is achieved if and only if all the weights are equal.*
2. *For $1 < p < q \leq \infty$, $\mathrm{ESS}_p(w^{1:N}) \geq \mathrm{ESS}_q(w^{1:N}) \geq N^{-(1-q_*/p_*)}\,\mathrm{ESS}_p(w^{1:N})$, with equality if and only if $K \in \{1, \ldots, N\}$ weights are equal and the rest are zero.*

Part (1) demonstrates that the p-ESS satisfies basic properties one would expect of a measure of effective sample size. Part (2) places the family of p-ESS measures in a linear order: the larger the value of p, the more stringent the notion of effective sample size.

The standard definition of effective sample size is precisely the 2-ESS. Whiteley et al. [18] provided a rigorous justification for the use of 2-ESS from the operator perspective: if adaptive resampling is used to guarantee that the 2-ESS does not fall below ζN, for some fixed parameter $\zeta \in (0, 1]$, then the error bounds on the operator approximation match those of the nonadaptive sampler with ζN particles. More formally, let $\Im_t(y_{1:t}) = y_t$ be the projection onto the tth component. Under appropriate regularity conditions, for every bounded measurable $\phi : E \to \mathbb{R}$ and real $r \geq 1$,

$$\sup_{t\geq 1} \mathrm{ESS}_2\big(W_t^{1:N}\big) \geq \zeta N \quad \Longrightarrow \quad \sup_{t\geq 1} \mathbb{E}^{0,N}\big[\big|\pi_{1:t}^{0,N}(\phi \circ \Im_t) - \pi(\phi \circ \Im_t)\big|^r\big]^{1/r} \leq \frac{a(r)b(\phi)}{\sqrt{\zeta N}},$$

where $a(r)$ and $b(\phi)$ are explicit functions.

To upper bound $\mathbb{E}^{i,N}_{\boldsymbol{y}_{1:t}^{1:i}}[\hat{Z}_t/Z_t]$, however, we will require a lower bound on the ∞-ESS, which by Proposition 1.4(2) is a more stringent notion of effective sample size than 2-ESS. We suspect this additional stringency is necessary (see Conjecture 1).

Assumption 1.A. There exists $\zeta \in (0, 1]$ such that $\mathrm{ESS}_\infty(W_s^{1:N}) \geq \zeta N$ for all $1 \leq s \leq t$.

We can ensure that Assumption 1.A holds by choosing $\mathrm{ESS} = \mathrm{ESS}_\infty$, which will allow us to bound the estimate of the normalization constant. For $s = 1, \ldots, t$, let

$$G_{s,t}(x_s) \triangleq \mathbb{E}\big[g_{s:t-1}(\xi_{s:t-1}) \mid \xi_s = x_s\big]$$

and let $G_{0,t} \triangleq \mathbb{E}[g_{1:t-1}(\xi_{1:t-1})]$. We now arrive at our second main result.

Theorem 1.5. *If Assumption 1.A holds, then for all $i, t \geq 1$, $N > i$, $y_{1:t}^1, \ldots, y_{1:t}^i \in E^t$,*

$$\mathbb{E}^{i,N}_{\boldsymbol{y}_{1:t}^{1:i}}[\hat{Z}_t/Z_t] \leq 1 + \frac{Z_t^{-1}\sum_{s=1}^t \sum_{j=1}^i G_{0,s}G_{s,t+1}(y_{1:t}^i) - \zeta i}{\zeta N} + \Theta\big(N^{-2}\big).$$

Two possible further assumptions both lead to uniform bounds on $\mathbb{E}^{i,N}_{\boldsymbol{y}_{1:t}^{1:i}}[\hat{Z}_t]$.

Assumption 1.B. The potentials satisfy $\overline{g}_s \triangleq \sup_{x\in E} g_s(x) < \infty$ for all $1 \leq s \leq t$.

Assumption 1.C. There exists a constant $\beta > 0$ such that for any $t, s \in \mathbb{N}$,

$$\sup_{x\in E} \frac{G_{0,t}G_{t,t+s}(x)}{G_{0,t+s}} \leq \beta.$$

Corollary 1.6. *Under the same conditions as Theorem 1.5, if Assumption 1.B holds then*

$$\mathbb{E}^{i,N}_{\mathbf{y}^{1:i}_{1:t}}[\hat{Z}_t/Z_t] \le 1 + Z_t^{-1}\prod_{s=1}^{t}\overline{g}_s\left[\left(1+\frac{i}{\zeta N}\right)^t - 1\right]$$

while if Assumption 1.C holds then

$$\mathbb{E}^{i,N}_{\mathbf{y}^{1:i}_{1:t}}[\hat{Z}_t/Z_t] \le \left(1+\frac{\beta}{\zeta N}\right)^t.$$

Combining Corollaries 1.2 and 1.6 yields the following guarantees for the ARPF sampler.

Theorem 1.7. *If Assumptions 1.A and 1.B hold, then*

$$d_{\mathrm{TV}}\big(\pi_{1:t}, P^{0,N}\big) \le \frac{tZ_t^{-1}\prod_{s=1}^{t}\overline{g}_s}{\zeta N + tZ_t^{-1}\prod_{s=1}^{t}\overline{g}_s} + \Theta\big(N^{-2}\big)$$

while if Assumptions 1.A and 1.C hold then

$$d_{\mathrm{TV}}\big(\pi_{1:t}, P^{0,N}\big) \le \frac{\beta t}{\zeta N + \beta t} + \Theta\big(N^{-2}\big).$$

1.3. An application to particle Gibbs

In the language of state-space models, the setting described so far involves approximating the posterior distribution of a Markov chain given indirect stochastic observations of the chain's values. However, it is often the case that the chain and the potentials are controlled by a global parameter $\theta \in \Theta$ for which there is a prior distribution $\varpi(\mathrm{d}\theta)$. Replace M_s by M_s^θ and g_s by g_s^θ, then parameterize the other quantities defined previously in terms of M_s and g_s by θ. Let $(Y, \mathcal{Y}) \triangleq (E^t, \mathcal{B}(E^t))$. We will suppress much of the time dependence when possible to make the notation less cluttered. The target distribution on the product space $(\Theta \times Y, \mathcal{B}(\Theta \times Y))$ is

$$\pi(\mathrm{d}\theta \times \mathrm{d}y) \triangleq \gamma(\mathrm{d}\theta \times \mathrm{d}y)/Z,$$

where

$$\gamma(\mathrm{d}\theta \times \mathrm{d}y) \triangleq \varpi(\mathrm{d}\theta)\prod_{s=1}^{t} g_s^\theta(y_s)M_s^\theta(y_{s-1}, \mathrm{d}y_s) \quad \text{and} \quad Z \triangleq \gamma(1).$$

Let $\pi_\theta(\mathrm{d}y)$ and $\pi_y(\mathrm{d}\theta)$ denote the disintegrations of π along Θ and along Y, respectively.

The particle Gibbs sampler approximates the two-stage Gibbs kernel

$$\Pi(\theta, y, \mathrm{d}\vartheta \times \mathrm{d}z) \triangleq \pi_y(\mathrm{d}\vartheta)\pi_\vartheta(\mathrm{d}z).$$

In many settings, such as non-linear or non-Gaussian state-space models, it is possible to sample from $\pi_y(\mathrm{d}\vartheta)$, but difficult to sample from $\pi_\vartheta(\mathrm{d}z)$. The idea is to replace $\pi_\vartheta(\mathrm{d}z)$ with an SMC-based approximation $\Pi_\vartheta(y, \mathrm{d}z)$ that leaves $\pi_\vartheta(\mathrm{d}z)$ invariant, leading to a kernel of the form $\pi_y(\mathrm{d}\vartheta)\Pi_\vartheta(y, \mathrm{d}z)$.

We introduce the adaptive resampling particle Gibbs (ARPG) sampler, which employs the cARPF kernel $P^{1,N}_{\vartheta,y}(\mathrm{d}z)$ to approximate the conditional distribution $\pi_\vartheta(\mathrm{d}z)$ that would be used in a standard Gibbs sampler. The ARPG kernel is thus given by

$$\Pi_N(\theta, y, \mathrm{d}\vartheta \times \mathrm{d}z) \triangleq \pi_y(\mathrm{d}\vartheta) P^{1,N}_{\vartheta,y}(\mathrm{d}z).$$

Theorem 1.1 and Corollaries 1.3 and 1.6 together with the results of Andrieu et al. [2] yield guarantees on the ergodicity properties of the cARPF kernel and the ARPG sampler. Once instances of N are replaced by ζN, the guarantees essentially match those provided by Andrieu et al. [2] for the standard PG sampler.

Theorem 1.8. *If Assumption 1.A holds, then the cARPF kernel and ARPG sampler have the following properties*:

1. *If Assumption 1.B holds, then there exists $\varepsilon_{t,N} = 1 - C_t/N$ such that for any $\theta \in \Theta$, $y \in Y$, and $k \geq 1$,*

$$d_{\mathrm{TV}}\big(\delta_y\big[P^{1,N}_\theta\big]^k, \pi_\theta\big) \leq (1 - \varepsilon_{t,N})^k. \tag{1.2}$$

2. *If Assumption 1.C holds and $N \geq t/C + 1$ for any fixed $C > 0$, then for any $t \geq 1$, equation (1.2) holds with*

$$\varepsilon_{t,N} \geq \exp\left(-\frac{C}{\zeta}(2\beta + \zeta)\right).$$

3. *If either Assumption 1.B or Assumption 1.C holds, then whenever the Gibbs sampler is geometrically ergodic the ARPG sampler is geometrically ergodic as well.*

At a high level, the results we have obtained highlight the role of the expected value of $\hat{Z}_t$ in the mixing properties of conditional SMC Markov chains and particle Gibbs (PG) samplers: In order to show geometric ergodicity for adaptive resampling particle Gibbs samplers, it suffices to establish bounds on the expected value of $\hat{Z}_t$ under the twice-conditional filter, and the growth of the expectation as t increases determining how well the particle Gibbs algorithm scales. Similarly, a bound on the expected value of $\hat{Z}_t$ under the once-conditional filter implies a bound on $d_{\mathrm{TV}}(\pi_{1:t}, P^{0,N})$. Hence, as a slogan, good performance of (adaptive resampling) particle Gibbs is equivalent to good performance of (adaptive) SMC for sampling.

2. Preliminaries

In this section, we fix some additional notation, introduce a few key additional definitions, and then present αSMC [18], a generalization of the adaptive resampling particle filter described in the Introduction.

For a positive integer K, let $[K] \triangleq \{1, 2, \ldots, K\}$. If $x_i, \ldots, x_j$ are elements of a sequence, write $x_{i:j} \triangleq \langle x_i, x_{i+1}, \ldots, x_j \rangle$. We use the following conventions: $\sum_{\varnothing} = 0$, $\prod_{\varnothing} = 1$, and $0/0 = 0$.

Let $(S, \mathcal{S})$, $(S', \mathcal{S}')$ be measurable spaces. Then $K : S \times \mathcal{S}' \to \mathbb{R}$ is a kernel if $K(\cdot, B)$ is a $(S, \mathcal{S})$-measurable function for all $B \in \mathcal{S}'$ and $K(x, \cdot)$ is measure on $(S', \mathcal{S}')$ for all $x \in \mathcal{S}$. For a measure μ on $(S, \mathcal{S})$ and kernels $K, K' : S \times \mathcal{S} \to \mathbb{R}$, let $\mu K(\mathrm{d}y) \triangleq \int \mu(\mathrm{d}x) K(x, \mathrm{d}y)$ and $KK'(x, \mathrm{d}z) \triangleq \int K(x, \mathrm{d}y) K'(y, \mathrm{d}z)$. We will often use measures and kernels as operators. For a measurable function $\phi : S \to \mathbb{R}$, let $\mu(\phi) \triangleq \mathbb{E}_{\xi \sim \mu}[\phi(\xi)] = \int \phi(x) \mu(\mathrm{d}x)$ and $K(\phi)(x) \triangleq \int \phi(y) K(x, \mathrm{d}y)$. For measures μ, ν on $(S, \mathcal{S})$, we will write $\mu \ll \nu$ to denote that μ is absolutely continuous with respect to ν, in which case we will write $\mathrm{d}\mu/\mathrm{d}\nu$ for the ν-almost everywhere (ν-a.e.) unique function f satisfying $\mu(A) = \int_A f \, \mathrm{d}\nu$, for all $A \in \mathcal{S}$. When the choice is clear from context, we may write $\mathcal{B}(S)$ for the σ-algebra of the space S.

For probability measures μ and ν on $(S, \mathcal{S})$, the total variation distance between μ and ν is

$$d_{\mathrm{TV}}(\mu, \nu) \triangleq \sup_{A \in \mathcal{S}} \big| \mu(A) - \nu(A) \big|.$$

If $\mu \ll \nu$, then the Kullback–Liebler (KL) divergence is

$$d_{\mathrm{KL}}(\mu || \nu) \triangleq \mu(\log \mathrm{d}\mu/\mathrm{d}\nu)$$

and the χ^2 divergence is

$$d_{\chi^2}(\mu || \nu) \triangleq \nu\big([\mathrm{d}\mu/\mathrm{d}\nu - 1]^2\big) = \mu(\mathrm{d}\mu/\mathrm{d}\nu) - 1.$$

Finally, we note that, when there is little risk of confusion, we will ignore measure-theoretic niceties such as the distinction between equality and a.e.-equality.

Recall that $(\xi_t)_{t \geq 1}$ is an inhomogeneous Markov chain on $(E, \mathcal{E})$ with transition kernels $(M_t)_{t \geq 2}$ and initial distribution M_1, and that $\mathbb{E}[\cdot]$ denotes expectation with respect to the Markov chain. We will write $M_1(x_0, \cdot)$ for $M_1(\cdot)$ when convenient and, for all $t \geq 1$ and $x_{t-1} \in E$, we will assume that $M_t(x_{t-1}, \cdot)$ has a density with respect to some common σ-finite dominating measure (which we denote by $\mathrm{d}x$). We will abuse notation and write $M_t(x_{t-1}, x_t)$ for the density of $M_t(x_{t-1}, \cdot)$ as x_t. Recall that, for each $t \geq 1$, $g_t : E \to \mathbb{R}_+$ denotes a $\mathcal{E}$-measurable potential function, with $g_0 \equiv 1$. Finally, recall that $g_{s:t}(x_{s:t}) \triangleq \prod_{\tau=s}^{t} g_\tau(x_\tau)$.

2.1. Target distributions

We now introduce some additional target distributions. (We will also repeat the definition of $\gamma_{1:t}$ and $\pi_{1:t}$ for completeness.)

Let $\phi_{1:t} : E^t \to \mathbb{R}$ and $\phi_t : E \to \mathbb{R}$ denote generic measurable functions. For each $t \geq 1$, the unnormalized predictive and updated measures are defined, respectively, by

$$\gamma'_{1:t}(\phi_{1:t}) \triangleq \mathbb{E}\big[\phi_{1:t}(\xi_{1:t}) g_{1:t-1}(\xi_{1:t-1})\big] \quad \text{and} \quad \gamma_{1:t}(\phi_{1:t}) \triangleq \mathbb{E}\big[\phi_{1:t}(\xi_{1:t}) g_{1:t}(\xi_{1:t})\big]$$

with corresponding marginal versions

$$\gamma_t'(\phi_t) \triangleq \mathbb{E}\big[\phi_t(\xi_t) g_{1:t-1}(\xi_{1:t-1})\big] \quad \text{and} \quad \gamma_t(\phi_t) \triangleq \mathbb{E}\big[\phi_t(\xi_t) g_{1:t}(\xi_{1:t})\big].$$

Our ultimate goal is to approximate the normalized predictive and updated measures along with their marginal versions:

$$\pi_{1:t}(\phi_{1:t}) \triangleq \frac{\gamma_{1:t}(\phi_{1:t})}{Z_t}, \qquad \eta_{1:t}(\phi_{1:t}) \triangleq \frac{\gamma_{1:t}'(\phi_{1:t})}{Z_t'},$$
$$\pi_t(\phi_t) \triangleq \frac{\gamma_t(\phi_t)}{Z_t} \quad \text{and} \quad \eta_t(\phi_t) \triangleq \frac{\gamma_t'(\phi_t)}{Z_t'},$$

where $Z_t \triangleq \gamma_t(1)$ and $Z_t' \triangleq \gamma_t'(1)$ are normalization constants.

2.2. The αSMC algorithm

In the Introduction, adaptation in the particle filter was implemented via a simple multinomial resampling step, triggered when the effective sample size fell below a fixed threshold. For the remainder of the article, we will consider a more general mechanism for adaptation captured by the αSMC algorithm introduced by Whiteley et al. [18]. The αSMC algorithm can produce sequential importance sampling (SIS), sampling importance resampling (SIR, also known as the bootstrap filter), and numerous other SMC variants as special cases. Not only does the αSMC formulation aid in analyzing adaptive resampling strategies, it provides a useful framework for devising novel adaptive schemes with attractive computational properties, such as admitting parallelization even on resampling steps. In the remainder of this section, we outline the (unconditional) αSMC algorithm. In the following section, we introduce a novel i-times conditional version of αSMC, which will include the c^{i}ARPF as a special case.

The αSMC algorithm, which is given as Algorithm 1, provides a flexible resampling mechanism: at each time t, a stochastic matrix α_{t-1} is chosen from a set $\mathbb{A}_N$ of $N \times N$ matrices. We denote the value in the nth row and kth column of α_{t-1} by α_{t-1}^{nk}. The αSMC estimators are

$$\pi_{1:t}^{0,N} \triangleq \sum_{n=1}^{N} \frac{W_t^n g_t(X_t^n)}{\sum_{k=1}^N W_t^k g_t(X_t^k)} \delta_{\tilde{X}_{1:t}^n}, \qquad \pi_t^{0,N} \triangleq \sum_{n=1}^{N} \frac{W_t^n g_t(X_t^n)}{\sum_{k=1}^N W_t^k g_t(X_t^k)} \delta_{X_t^n},$$
$$\eta_{1:t}^{0,N} \triangleq \sum_{n=1}^{N} \frac{W_t^n}{\sum_{k=1}^N W_t^k} \delta_{\tilde{X}_{1:t}^n} \quad \text{and} \quad \eta_t^{0,N} \triangleq \sum_{n=1}^{N} \frac{W_t^n}{\sum_{k=1}^N W_t^k} \delta_{X_t^n},$$

and the estimators of the normalization constants Z_t and Z_t' are

$$\hat{Z}_t \triangleq \frac{1}{N} \sum_{n=1}^{N} W_t^n g_t\big(X_t^n\big) \quad \text{and} \quad \hat{Z}_t' \triangleq \frac{1}{N} \sum_{n=1}^{N} W_t^n.$$

Algorithm 1 αSMC

for $n = 1, \ldots, N$ **do**
  Sample X_1^n from M_1
  Set $\tilde{X}_{1:1}^n \leftarrow X_1^n$
  Set $W_1^n \leftarrow 1$
end for
for $t = 2, 3, \ldots$ **do**
  Select α_{t-1} from $\mathbb{A}_N$ according to some function of $\boldsymbol{X}_{1:t-1}^{1:N}$ and $\boldsymbol{A}_{1:t-2}^{1:N}$
  for $n = 1, \ldots, N$ **do**
    Set $W_t^n \leftarrow \sum_{k=1}^N \alpha_{t-1}^{nk} W_{t-1}^k g_{t-1}(X_{t-1}^k)$
    Sample A_{t-1}^n from $\mathsf{Multi}(\langle \frac{\alpha_{t-1}^{nk} W_{t-1}^k g_{t-1}(X_{t-1}^k)}{W_t^n} \rangle_{k=1}^N)$
    Sample X_t^n from $M_t(X_{t-1}^{A_{t-1}^n}, \cdot)$
    Set $\tilde{X}_{1:t}^n \leftarrow \langle X_{1:t-1}^{A_{t-1}^n}, X_t^n \rangle$
  end for
end for

Expectations with respect the law of the αSMC algorithm are written as $\mathbb{E}^{0,N}[\cdot]$.

SIS, SIR, and the standard adaptive algorithm are obtained as special cases of αSMC as follows. SIS is recovered when $\alpha_{t-1} = \boldsymbol{I}_N$, the $N \times N$ identity matrix, while SIR is recovered when $\alpha_{t-1} = \mathbf{1}_{1/N}$, the $N \times N$ matrix with all entries equal to $1/N$. The adaptive particle filter (APF) algorithm is obtained by setting α_{t-1} to $\mathbf{1}_{1/N}$ if $\mathrm{ESS}_2(W_{t-1}^{1:N}) \leq \zeta N$ and to $\boldsymbol{I}_N$ otherwise, where $\zeta \in (0, 1]$ is fixed.

3. Conditional αSMC

It is useful both algorithmically and analytically to generalize αSMC in such a way that one or more particle trajectories is fixed ahead of time. The result, which we will refer to as *conditional* αSMC, is a strict generalization of the conditional adaptive particle filter given in the Introduction. We will use conditional αSMC to study the properties of (unconditional) αSMC, to design novel adaptive particle Gibbs algorithms, and to analyze their mixing properties.

For this section, fix $t \geq 1$, $i \geq 0$, and $N > i$. The *i-times conditional αSMC* (c$^i\alpha$SMC) *process* (or simply the *cαSMC process* when $i = 1$) is defined on the space $(E^N \times [N]^N \times [N]^i)^{t-1} \times E^N \times [N] \times [N]^i$, and is essentially equivalent to αSMC with the first i particle trajectories, but not their lineages, fixed a priori. If $f^{1:i} \in [N]^i$ are indices of the first i particles, let $\mathcal{D}(f^{1:i}) \triangleq \prod_{j \neq j'} \mathbb{1}(f^j \neq f^{j'})$ be the function that indicates whether the indices are distinct. As in αSMC, the matrix $\alpha_{t-1} \in \mathbb{A}_N$ is a function of $\boldsymbol{X}_{1:t-1}^{1:N}$ and $\boldsymbol{A}_{1:t-2}^{1:N}$. We have $\boldsymbol{x}_{1:t}^{1:N} \in (E^N)^t$, $\boldsymbol{f}_{1:t}^{1:i} \in ([N]^i)^t$, $\boldsymbol{a}_{1:t-1}^{1:N} \in ([N]^N)^{t-1}$, and $a_t^* \in [N]$, and use the notation

$$w_1^n \triangleq 1, \qquad w_t^n \triangleq \sum_{k=1}^N \alpha_{t-1}^{nk} w_{t-1}^k g_{t-1}\big(x_{t-1}^k\big),$$

and

$$r_n\big(k|w_{s-1}^{1:N}, \boldsymbol{x}_{1:s-1}^{1:N}\big) \triangleq \frac{\alpha_{s-1}^{nk} w_{s-1}^k g_{s-1}(x_{s-1}^k)}{w_s^n}.$$

For fixed trajectories $y_{1:t}^1, \dots, y_{1:t}^i \in E^t$, the law of the c$^i\alpha$SMC process is given by

$$\mathbb{P}_{\boldsymbol{y}_{1:t}^{1:i}}^{i,N}\big[X_1^{1:N} \in \mathrm{d}x_1^{1:N}, F_1^{1:i} = f_1^{1:i}\big] \triangleq \mathcal{C}_1^i \mathcal{D}\big(f_1^{1:i}\big) \prod_{j=1}^{i} \frac{1}{N} \delta_{y_1^j}\big(\mathrm{d}x_1^{f_1^j}\big) \prod_{n \notin f_1^{1:i}}^{N} M_1\big(\mathrm{d}x_1^n\big),$$

for $s = 2, \dots, t$,

$$\begin{aligned}
&\mathbb{P}_{\boldsymbol{y}_{1:t}^{1:i}}^{i,N}\big[X_s^{1:N} \in \mathrm{d}x_s^{1:N}, A_{s-1}^{1:N} = a_{s-1}^{1:N}, F_s^{1:i} = f_s^{1:i} \mid \\
&\qquad \boldsymbol{X}_{1:s-1}^{1:N} = \boldsymbol{x}_{1:s-1}^{1:N}, \boldsymbol{A}_{1:s-2}^{1:N} = \boldsymbol{a}_{1:s-2}^{1:N}, F_{s-1}^{1:i} = f_{s-1}^{1:i}\big] \\
&\quad \triangleq \mathcal{C}_s^i \mathcal{D}\big(f_s^{1:i}\big) \prod_{j=1}^{i} \alpha_{s-1}^{f_s^j f_{s-1}^j} \delta_{y_s^j}\big(\mathrm{d}x_s^{f_s^j}\big) \mathbb{1}\big(a_s^{f_{s-1}^j} = f_{s-1}^j\big) \\
&\qquad \times \prod_{n \notin f_s^{1:i}} r_n\big(a_{s-1}^n | w_{s-1}^{1:N}, \boldsymbol{x}_{1:s-1}^{1:N}\big) M_s\big(x_{s-1}^{a_{s-1}^n}, x_s^n\big)
\end{aligned}$$

and

$$\mathbb{P}_{\boldsymbol{y}_{1:t}^{1:i}}^{i,N}\big[A_t^* = a_t^* \mid \boldsymbol{X}_{1:t}^{1:N} = \boldsymbol{x}_{1:t}^{1:N}, \boldsymbol{A}_{1:t-1}^{1:N} = \boldsymbol{a}_{1:t-1}^{1:N}\big] \triangleq \frac{w_t^{a_t^*} g_t(x_t^{a_t^*})}{\sum_{n=1}^N w_t^n g_t(x_t^n)}.$$

The $\mathcal{C}_s^i$ terms are normalization constants that ensure the expressions are valid probabilities. Let $\tilde{X}_{1:t}^* \triangleq X_{1:t}^{A_t^*}$, let $\mathbb{E}_{\boldsymbol{y}_{1:t}^{1:i}}^{i,N}[\cdot]$ denote the expectation operator with respect to the c$^i\alpha$SMC, and write

$$P^{i,N}\big(\boldsymbol{y}_{1:t}^{1:i}, \mathrm{d}x_{1:t}\big) \triangleq P_{\boldsymbol{y}_{1:t}^{1:i}}^{i,N}(\mathrm{d}x_{1:t}) \triangleq \mathbb{E}_{\boldsymbol{y}_{1:t}^{1:i}}^{i,N}\big[\delta_{\tilde{X}_{1:t}^*}(\mathrm{d}x_{1:t})\big]$$

for the law of $\tilde{X}_{1:t}^*$.

The normalization constants $\mathcal{C}_s^i$ arise because the lineages $\boldsymbol{f}_{1:t}^{1:i}$ of the fixed trajectories $\boldsymbol{y}_{1:t}^{1:i}$ are kept distinct. The $c^i\alpha$SMC kernel enforces distinct lineages for the fixed trajectories since in general $y_{1:t}^j \neq y_{1:t}^{j'}$ for $j \neq j'$ and, from an algorithmic standpoint, allowing overlapping lineages could lead to a substantial increase in complexity, both in terms of implementation and computation. The distinct lineage requirement is enforced by the $\mathcal{D}(f_s^{1:i})$ terms. Since there is at most one fixed trajectory when $i = 0$ or 1, $\mathcal{C}_s^0 = \mathcal{C}_s^1 = 1$ for all $s \in [t]$.

Algorithm 2 provides pseudocode to sample from the law of the cαSMC process, which is a necessary part of the particle Gibbs sampler described in Section 6. Sampling from the law of the $c^i\alpha$SMC process for $i > 1$ is more delicate, but unnecessary since these are only used for analytical purposes.

Algorithm 2 Conditional αSMC

Require: Fixed trajectory $y_{1:t}$
  Sample F_1^1 uniformly from $\{1, \ldots, N\}$
  for $n = 1, \ldots, N$ **do**
    if $n = F_1^1$ **then**
      Set $X_1^n \leftarrow y_1$
    else
      Sample X_1^n from M_1
    end if
    Set $\tilde{X}_{1:1}^n \leftarrow X_1^n$
    Set $W_1^n \leftarrow 1$
  end for
  for $t = 2, 3, \ldots$ **do**
    Select α_{t-1} from $\mathbb{A}_N$ according to some function of $\boldsymbol{X}_{1:t-1}^{1:N}$ and $\boldsymbol{A}_{1:t-2}^{1:N}$
    Sample F_t^1 from $\mathsf{Multi}(\langle \alpha_{t-1}^{kF_{t-1}^1} \rangle_{k=1}^N)$
    for $n = 1, \ldots, N$ **do**
      Set $W_t^n \leftarrow \sum_{k=1}^N \alpha_{t-1}^{nk} W_{t-1}^k g_{t-1}(X_{t-1}^k)$
      if $n = F_t^1$ **then**
        Set $A_{t-1}^n \leftarrow F_{t-1}^1$
        Set $X_t^n \leftarrow y_t$
      else
        Sample A_{t-1}^n from $\mathsf{Multi}(\langle \frac{\alpha_{t-1}^{nk} W_{t-1}^k g_{t-1}(X_{t-1}^k)}{W_t^n} \rangle_{k=1}^N)$
        Sample X_t^n from $M_t(X_{t-1}^{A_{t-1}^n}, \cdot)$
      end if
      Set $\tilde{X}_{1:t}^n \leftarrow \langle X_{1:t-1}^{A_{t-1}^n}, X_t^n \rangle$
    end for
  end for
  Sample A_t^* from $\mathsf{Multi}(\langle \frac{W_t^k g_t(X_t^k)}{\sum_{n=1}^N W_t^n g_t(X_t^n)} \rangle_{k=1}^N)$

Remark 3.1. To recover the c^{i}ARPF described in the Introduction, let $\alpha_{t-1} = \mathbf{1}_{1/N}$ if $\mathrm{ESS}(W_{t-1}^{1:N}) \leq \zeta N$ and let $\alpha_{t-1} = \boldsymbol{I}_N$ otherwise. Then note that by the symmetry of $\mathbf{1}_{1/N}$ and $\boldsymbol{I}_N$, instead of sampling them, we can set the lineage for the jth fixed trajectory to j: that is, set $f_s^j = j$ for all $j \in [i]$ and $s \in [t]$.

4. Main results

We are now ready to undertake our study of the i-times conditional αSMC kernel $P^{i,N}(\boldsymbol{y}_{1:t}^{1:i}, \mathrm{d}x_{1:t})$. Specifically, our aim is to understand the conditions under which the c$^i\alpha$SMC kernel is close to $\pi_{1:t}(\mathrm{d}x_{1:t})$. Formally, we will establish a minorization condition for the i-times conditional filter in terms of the expected value of $\hat{Z}_t$ under the $(i+1)$-times conditional filter.

The remainder of the section presents conditions under which the expected value of $\hat{Z}_t$ can be bounded. One of the key assumptions is that adaptation controls the ∞-ESS.

Of particular interest are the cases $i = 0$, which corresponds to the αSMC filter, and $i = 1$, which corresponds to the conditional αSMC kernel. The former provides approximate samples from $\pi_{1:t}$. The latter can be used to define a Markov chain with invariant distribution $\pi_{1:t}$, producing an adaptive resampling particle Gibbs sampler. We consider both these applications in, respectively, Sections 5 and 6.

4.1. A minorization condition for the c$^i\alpha$SMC kernel

For the remainder of the article, we will work under the following assumption.

Assumption 4.D. For all $N \geq 1$, all $\alpha \in \mathbb{A}_N$ are doubly stochastic.

Remark 4.1. Assumption 4.D is the same as Assumption (B^{++}) in [18], although there, the condition is stated as assuming each α admits the uniform distribution as an invariant measure.

Let

$$\kappa_N \triangleq \max_{n \neq m, \alpha \in \mathbb{A}_N} \sum_{k=1}^{N} \alpha^{kn} \alpha^{km} \quad \text{and} \quad \kappa'_N \triangleq \max\{\kappa_N, 1/N\}.$$

Our first main result provides control over how much the measure $P^{i,N}_{\boldsymbol{y}^{1:i}_{1:t}}$ differs from $\pi_{1:t}$. The theorem gives a stronger result when $i = 0$ and gives a simpler result when $i = 1$, by expressing the lower bound on $P^{1,N}(y_{1:t}, S)$ in terms the more transparent quantity κ'_N instead of the normalization terms $\mathcal{C}^2_s$. For example, if $\mathbb{A}_N = \{\boldsymbol{I}_N\}$ then $\kappa_N = 0$, while if $\mathbb{A}_N = \{\mathbf{1}_{1/N}\}$, then $\kappa_N = 1/N$, so in either case $\kappa'_N = 1/N$.

Theorem 4.2. *If Assumption 4.D holds, then for all $t \geq 1$, $i \geq 0$, $N > i$, $S \subseteq E^t$ measurable, and $y^1_{1:t}, \ldots, y^i_{1:t}, x_{1:t} \in E^t$,*

$$P^{i,N}\big(\boldsymbol{y}^{1:i}_{1:t}, S\big) \geq \int_S \frac{Z_t}{\mathbb{E}^{i+1,N}_{\boldsymbol{y}^{1:i}_{1:t}, x_{1:t}}[\hat{Z}_t \prod_{s=1}^{t} \mathcal{C}^{i+1}_s / \mathcal{C}^i_s]} \pi_{1:t}(\mathrm{d}x_{1:t}). \tag{4.1}$$

In particular, in the case of $i = 0$, we have

$$\frac{\mathrm{d}P^{0,N}}{\mathrm{d}\pi_{1:t}}(x_{1:t}) = \mathbb{E}^{1,N}_{x_{1:t}}\left[\frac{Z_t}{\hat{Z}_t}\right] \geq \frac{Z_t}{\mathbb{E}^{1,N}_{x_{1:t}}[\hat{Z}_t]}$$

and in the case of $i = 1$, we have

$$P^{1,N}(y_{1:t}, S) \geq \int_S \frac{Z_t (1 - \kappa'_N)^t}{\mathbb{E}^{2,N}_{y_{1:t}, x_{1:t}}[\hat{Z}_t]} \pi_{1:t}(\mathrm{d}x_{1:t}). \tag{4.2}$$

Remark 4.3. By identical arguments, Theorem 4.2 also holds in the marginal and predictive cases. In the predictive cases, however, $\hat{Z}_t'$ and Z_t' replace, respectively, $\hat{Z}_t$ and Z_t. In the predictive case, under Assumption 4.D, $\hat{Z}_t' = \hat{Z}_{t-1}$ and $Z_t' = Z_{t-1}$, so later results pertaining to $\mathbb{E}^{i,N}_{\boldsymbol{y}_{1:t}^{1:i}}[\hat{Z}_t]$, such as Proposition 4.4 and Theorem 4.7, apply to Z_t' as well. The fact that $Z_t' = Z_{t-1}$ follows immediately from the definitions. To show that $\hat{Z}_t' = \hat{Z}_{t-1}$, apply Assumption 4.D:

$$\begin{aligned}\hat{Z}_t' &= N^{-1}\sum_n W_t^n \\ &= N^{-1}\sum_n\sum_k \alpha_{t-1}^{nk} W_{t-1}^k g_{t-1}^k \\ &= N^{-1}\sum_k W_{t-1}^k g_{t-1}^k = \hat{Z}_{t-1}.\end{aligned}$$

We will prove Theorem 4.2 in two parts: first for the case of $i = 0$, then for $i \geq 1$. For the $i = 0$ case (corresponding to vanilla αSMC), we begin by writing the joint density of the αSMC process as

$$\begin{aligned}&\psi\big(\boldsymbol{x}_{1:t}^{1:N}, \boldsymbol{a}_{1:t-1}^{1:N}\big) \\ &\quad\triangleq \left(\prod_{n=1}^N M_1\big(x_1^n\big)\right)\left(\prod_{s=2}^t\prod_{n=1}^N r_n\big(a_{s-1}^n | w_{s-1}^{1:N}, \boldsymbol{x}_{1:s-1}^{1:N}\big) M_s\big(x_{s-1}^{a_{s-1}^n}, x_s^n\big)\right).\end{aligned}$$

Under Assumption 4.D, the density of the cαSMC process with law $\mathbb{P}^{1,N}_{y_{1:t}}[\boldsymbol{X}_{1:t}^{1:N}, \boldsymbol{A}_{1:t-1}^{1:N}, F_{1:t}^1]$ can be written in the following "collapsed" form, by implicitly identifying $x_{1:t}^{f_t^1}$ with $y_{1:t}$:

$$\begin{aligned}&\tilde{\psi}\big(\boldsymbol{x}_{1:t}^{1:N}, \boldsymbol{a}_{1:t-1}^{1:N}, f_{1:t}^1\big) \\ &\quad= \frac{\psi(\boldsymbol{x}_{1:t}^{1:N}, \boldsymbol{a}_{1:t-1}^{1:N}) \prod_{s=2}^t I_s \alpha_{s-1}^{f_s^1 f_{s-1}^1}}{N M_1(x_1^{f_1^1}) \prod_{s=2}^t r_{f_s^1}(f_{s-1}^1 | w_{s-1}^{1:N}, \boldsymbol{x}_{1:s-1}^{1:N}) M_s(x_{s-1}^{f_{s-1}^1}, x_s^{f_s^1})} \\ &\quad= \frac{1}{N}\prod_{n\neq f_1^1} M_1\big(x_1^n\big)\prod_{s=2}^t\left(I_s \alpha_{s-1}^{f_s^1 f_{s-1}^1}\prod_{n\neq f_s^1} r_n\big(a_{s-1}^n | w_{s-1}^{1:N}, \boldsymbol{x}_{1:s-1}^{1:N}\big) M_s\big(x_{s-1}^{a_{s-1}^n}, x_s^n\big)\right),\end{aligned}$$

where $I_s \triangleq \mathbb{1}(a_{s-1}^{f_s^1} = f_{s-1}^1)$.

Proof of Theorem 4.2, $i = 0$ case. Consider the density

$$\tilde{\pi}_{1:t}\big(\boldsymbol{x}_{1:t}^{1:N}, \boldsymbol{a}_{1:t-1}^{1:N}, f_{1:t}^1\big) \triangleq \pi_{1:t}\big(x_{1:t}^{f_t^1}\big)\tilde{\psi}\big(\boldsymbol{x}_{1:t}^{1:N}, \boldsymbol{a}_{1:t-1}^{1:N}, f_{1:t}^1\big). \tag{4.3}$$

Then

$$
\begin{aligned}
&\frac{\psi(\boldsymbol{x}_{1:t}^{1:N}, \boldsymbol{a}_{1:t-1}^{1:N}) g_t(x_t^{f_t^1}) w_t^{f_t^1}}{\tilde{\pi}_{1:t}(\boldsymbol{x}_{1:t}^{1:N}, \boldsymbol{a}_{1:t-1}^{1:N}, f_{1:t}^1) \sum_{n=1}^N g_t(x_t^n) w_t^n} \\
&\quad = \frac{M_1(x_1^{f_1^1}) \prod_{s=2}^t r_{f_s^1}(f_{s-1}^1 | w_{s-1}^{1:N}, \boldsymbol{x}_{1:s-1}^{1:N}) M_s(x_{s-1}^{f_{s-1}^1}, x_s^{f_s^1}) g_t(x_t^{f_t^1}) w_t^{f_t^1}}{\pi_{1:t}(x_{1:t}^{f_t^1}) \prod_{s=2}^t I_s \alpha_{s-1}^{f_s^1 f_{s-1}^1} N^{-1} \sum_{n=1}^N g_t(x_t^n) w_t^n} \\
&\quad = \frac{M_1(x_1^{f_1^1}) \prod_{s=2}^t \alpha_{s-1}^{f_s^1 f_{s-1}^1} w_{s-1}^{f_{s-1}^1} g_{s-1}(x_{s-1}^{f_{s-1}^1}) M_s(x_{s-1}^{f_{s-1}^1}, x_s^{f_s^1}) g_t(x_t^{f_t^1}) w_t^{f_t^1}}{\pi_{1:t}(x_{1:t}^{f_t^1}) \prod_{s=2}^t w_s^{f_s^1} \prod_{s=2}^t I_s \alpha_{s-1}^{f_s^1 f_{s-1}^1} N^{-1} \sum_{n=1}^N g_t(x_t^n) w_t^n} \\
&\quad = \frac{\prod_{s=1}^t g_s(x_s^{f_s^1}) M_s(x_{s-1}^{f_{s-1}^1}, x_s^{f_s^1})}{\pi_{1:t}(x_{1:t}^{f_t^1}) N^{-1} \sum_{n=1}^N g_t(x_t^n) w_t^n \prod_{s=2}^t I_s} \\
&\quad = \frac{Z_t}{\hat{Z}_t} \frac{1}{\prod_{s=2}^t I_s}.
\end{aligned}
$$

Using the convention that $0/0 = 0$, it follows that

$$
\begin{aligned}
P^{0,N}(\mathrm{d}x_{1:t}) &= \sum_{\boldsymbol{a}_{1:t-1}^{1:N}, a_t^*} \int \left\{ \psi\left(\boldsymbol{x}_{1:t}^{1:N}, \boldsymbol{a}_{1:t-1}^{1:N}\right) \frac{g_t(x_t^{a_t^*}) w_t^{a_t^*}}{\sum_{n=1}^N g_t(x_t^n) w_t^n} \delta_{x_{1:t}^{a_t}}(\mathrm{d}x_{1:t}) \right\} \mathrm{d}\boldsymbol{x}_{1:t}^{1:N} \\
&= \sum_{\boldsymbol{a}_{1:t-1}^{1:N}, f_{1:t}^1} \int \left\{ \frac{\psi(\boldsymbol{x}_{1:t}^{1:N}, \boldsymbol{a}_{1:t-1}^{1:N}) g_t(x_t^{f_t^1}) w_t^{f_t^1} \prod_{s=2}^t I_s}{\tilde{\pi}_{1:t}(\boldsymbol{x}_{1:t}^{1:N}, \boldsymbol{a}_{1:t-1}^{1:N}, \boldsymbol{f}_{1:t}^{1:i}) \sum_{n=1}^N g_t(x_t^n) w_t^n} \right. \\
&\qquad \left. \times \tilde{\pi}_{1:t}\left(\boldsymbol{x}_{1:t}^{1:N}, \boldsymbol{a}_{1:t-1}^{1:N}, f_{1:t}^1\right) \delta_{x_{1:t}^{f_t}}(\mathrm{d}x_{1:t}) \right\} \mathrm{d}\boldsymbol{x}_{1:t}^{1:N} \\
&= \sum_{\boldsymbol{a}_{1:t-1}^{1:N}, f_{1:t}^1} \int \left\{ \frac{Z_t}{\hat{Z}_t} \tilde{\pi}_{1:t}\left(\boldsymbol{x}_{1:t}^{1:N}, \boldsymbol{a}_{1:t-1}^{1:N}, f_{1:t}^1\right) \delta_{x_{1:t}^{f_t}}(\mathrm{d}x_{1:t}) \right\} \mathrm{d}\boldsymbol{x}_{1:t}^{1:N} \\
&= \left\{ \sum_{\boldsymbol{a}_{1:t-1}^{1:N}, f_{1:t}^1} \int \frac{Z_t}{\hat{Z}_t} \tilde{\psi}\left(\boldsymbol{x}_{1:t}^{1:N}, \boldsymbol{a}_{1:t-1}^{1:N}, f_{1:t}^1\right) \delta_{x_{1:t}}\left(\mathrm{d}x_{1:t}^{f_t}\right) \mathrm{d}\boldsymbol{x}_{1:t}^{-f_t^1} \right\} \pi_{1:t}(\mathrm{d}x_{1:t}).
\end{aligned}
$$

The result follows from Lemma A.1. □

We defer the proof of Theorem 4.2 in the $i \geq 1$ case to Appendix A.2.

4.2. Bounding $\hat{Z}_t$ under the c$^i\alpha$SMC kernel

In order to apply Theorem 4.2, we must be able to control the quantity $\mathbb{E}^{i,N}_{\boldsymbol{y}^{1:i}_{1:t}}[\hat{Z}_t]$. As an initial step toward this goal, we consider the SIR case:

Assumption SIR. For all $s \in [t-1]$, $\alpha_s = \mathbf{1}_{1/N}$.

For SIR, we can rewrite $\mathbb{E}^{i,N}_{\boldsymbol{y}^{1:i}_{1:t}}[\hat{Z}_t]$ in an equivalent but more explicit form Proposition 4.4. Our goal will then be to provide general conditions under which $\mathbb{E}^{i,N}_{\boldsymbol{y}^{1:i}_{1:t}}[\hat{Z}_t]$ can be rewritten in a similar manner (Proposition 4.7).

Recall that $G_{0,t} \triangleq \mathbb{E}[g_{1:t-1}(\xi_{1:t-1})]$ and $G_{s,t}(x_s) \triangleq \mathbb{E}[g_{s:t-1}(\xi_{s:t-1}) \mid \xi_s = x_s]$ for $s \in [t]$ and $x_s \in E$. For $t \geq 1$, $1 \leq \ell \leq s \leq t+1$, let

$$\mathcal{T}_{t,\ell,s} \triangleq \left\{\langle \tau_1, \ldots, \tau_\ell \rangle : t-s+1 < \tau_1 < \cdots < \tau_\ell = t+1\right\}$$

and, for $\boldsymbol{\tau} \in \mathcal{T}_{t,\ell,s}$, define

$$C_\ell\big(\boldsymbol{\tau}, y_{1:t}\big) \triangleq \prod_{i=1}^{\ell-1} G_{\tau_i,\tau_{i+1}}(y_{\tau_i}).$$

We will sometimes write $C^y_\ell(\boldsymbol{\tau})$ or $C^{\boldsymbol{\tau}}_\ell(y_{1:t})$ instead of $C_\ell(\boldsymbol{\tau}, y_{1:t})$. The following is a straightforward generalization of [2], Proposition 9.

Proposition 4.4. *If Assumption SIR holds, then for all $t \geq 1$, $i \geq 1$, $N \geq i$, $y^1_{1:t}, \ldots, y^i_{1:t} \in E^t$,*

$$\mathbb{E}^{i,N}_{\boldsymbol{y}^{1:i}_{1:t}}[\hat{Z}_t] = \frac{1}{N^t}\sum_{\ell=1}^{t+1}(N-i)^{t+1-\ell} \sum_{\boldsymbol{\tau}\in\mathcal{T}_{t,\ell,t+1}} G_{0,\tau_1}\prod_{m=1}^{\ell-1}\sum_{j=1}^{i} G_{\tau_m,\tau_{m+1}}\big(y^j_{\tau_m}\big). \tag{4.4}$$

In particular, in the case of $i = 1$, we have

$$\mathbb{E}^{1,N}_{y_{1:t}}[\hat{Z}_t] = \frac{1}{N^t}\sum_{\ell=1}^{t+1}(N-1)^{t+1-\ell} \sum_{\boldsymbol{\tau}\in\mathcal{T}_{t,\ell,t+1}} G_{0,\tau_1} C_\ell(\boldsymbol{\tau}, y_{1:t}).$$

In order to obtain a version of Proposition 4.4 for the general c$^i\alpha$SMC case, we will require that the algorithm enforce a lower bound on a carefully chosen notion of effective sample size called ∞-ESS. The ∞-ESS is a member of a family of effective sample size measures we call p-ESS, which also includes two commonly used definitions as special cases.

Definition 4.5. For parameter $p \in [1, \infty]$, let $p_* \triangleq \frac{p}{p-1}$ be the conjugate exponent of p (so $1/p + 1/p_* = 1$). The *p-effective sample size* (p-ESS) of the weight vector $w^{1:N} \in \mathbb{R}_+^N$ is

$$\mathrm{ESS}_p\big(w^{1:N}\big) \triangleq \begin{cases} \left(\dfrac{\|w^{1:N}\|_1}{\|w^{1:N}\|_p}\right)^{p_*}, & p > 1, \\ \dfrac{\|w^{1:N}\|_1}{\prod_{n=1}^N (w^n)^{w^n/\|w^{1:N}\|_1}}, & p = 1. \end{cases}$$

The following proposition highlights some elementary properties of p-ESS and subsumes Proposition 1.4 (see Appendix A.3 for a proof).

Proposition 4.6. *The p-ESS has the following properties*:

1. *For all $p \in [1, \infty]$, $1 \le \mathrm{ESS}_p(w^{1:N}) \le N$. The lower bound is achieved if and only if all but one of the weights is zero. The upper bound is achieved if and only if all the weights are equal.*
2. *For $1 \le p < q \le \infty$, $\mathrm{ESS}_p(w^{1:N}) \ge \mathrm{ESS}_q(w^{1:N}) \ge N^{-(1-q_*/p_*)}\,\mathrm{ESS}_p(w^{1:N})$, with equality if and only if $K \in \{1, \dots, N\}$ weights are equal and the rest are zero.*
3. *The 1-ESS satisfies*

$$\mathrm{ESS}_1\big(w^{1:N}\big) = \lim_{p \downarrow 1} \mathrm{ESS}_p\big(w^{1:N}\big) = \mathrm{ESS}_{\mathrm{ent}}\big(w^{1:N}\big) \triangleq e^{H(w^{1:N})},$$

where $H(w^{1:N}) \triangleq -\sum_n \frac{w^n}{\|w^{1:N}\|_1} \log \frac{w^n}{\|w^{1:N}\|_1}$ is the entropy.

Parts (1) and (2) generalize their counterparts in Proposition 1.4 to all $p \in [1, \infty]$, including the case $p = 1$. Part (3) shows that the 1-ESS corresponds to the entropic ESS, which is a common choice of ESS in applications [6].

In order to obtain a bound on $\mathbb{E}^{i,N}_{\boldsymbol{y}_{1:t}}[\hat{Z}_t]$, we will require a lower bound on the ∞-ESS of the weights, as formalized in Assumption 1.A. Our development follows that of [18], who used the 2-ESS lower bound guarantee to bound the L_2 norm of the weights in terms of their L_1 norm. Similarly, we will use the ∞-ESS lower bound guarantee to bound the sup-norm of the weights in terms of their L_1 norm. Specifically, under Assumption 1.A, we have

$$\zeta N \le \mathrm{ESS}_\infty\big(W_s^{1:N}\big) = \frac{\|W_s^{1:N}\|_1}{\|W_s^{1:N}\|_\infty} = \frac{\|W_s^{1:N}\|_1}{\sup_n W_s^n},$$

and so, for all $n \in [N]$ and $s \in [t]$, $W_s^n \le \frac{\|W_s^{1:N}\|_1}{\zeta N}$. We can use this upper bound on W_s^n to prove a result that is very similar to Proposition 4.4, but permits an arbitrary adaptation scheme satisfying Assumptions 1.A and 4.D. (See Appendix A.4 for a proof):

Theorem 4.7. *If Assumptions 1.A and 4.D hold, then for all $t \geq 1$, $i \geq 1$, $N \geq i$, $y^1_{1:t}, \ldots, y^i_{1:t} \in E^t$,*

$$\mathbb{E}^{i,N}_{\mathbf{y}^{1:i}_{1:t}}[\hat{Z}_t] \leq \frac{1}{N(\zeta N)^{t-1}} \sum_{\ell=1}^{t+1} \sum_{\boldsymbol{\tau} \in \mathcal{T}_{t,\ell,t+1}} (\zeta N)^{t+1-\ell} \left(\frac{N-i}{\zeta N}\right)^{\mathbb{1}(\tau_1 > 1)} G_{0,\tau_1} \prod_{m=1}^{\ell-1} \sum_{j=1}^{i} G_{\tau_m,\tau_{m+1}}\left(y^j_{\tau_m}\right). \tag{4.5}$$

In particular, in the case of $i = 1$, we have

$$\mathbb{E}^{1,N}_{y_{1:t}}[\hat{Z}_t] \leq \frac{1}{N(\zeta N)^{t-1}} \sum_{\ell=1}^{t+1} \sum_{\boldsymbol{\tau} \in \mathcal{T}_{t,\ell,t+1}} (\zeta N)^{t+1-\ell} \left(\frac{N-1}{\zeta N}\right)^{\mathbb{1}(\tau_1 > 1)} G_{0,\tau_1} C^y_\ell(\boldsymbol{\tau}).$$

The gap between Proposition 4.4 and Theorem 4.7 is that most of the factors of $N - i$ in the former are replaced by factors of N in the latter. Luckily we are interested in the $i = 1, 2$ cases, so we expect the differences between the two quantities to be fairly small. The following result, which is immediate upon expanding the left-hand sides of equations (4.4) and (4.5) and keeping only $\Omega(1/N)$ terms, formalizes this intuition:

Corollary 4.8. *If Assumption SIR holds, then for all $i, t \geq 1$, $N > i$, $y^1_{1:t}, \ldots, y^i_{1:t} \in E^t$,*

$$\mathbb{E}^{i,N}_{\mathbf{y}^{1:i}_{1:t}}[\hat{Z}_t / Z_t] = 1 + \frac{Z_t^{-1} \sum_{s=1}^{t} \sum_{j=1}^{i} G_{0,s} G_{s,t+1}(y^i_{1:t}) - ti}{N} + \Theta\left(N^{-2}\right).$$

If Assumptions 1.A and 4.D hold, then for all $i, t \geq 1$, $N > i$, $y^1_{1:t}, \ldots, y^i_{1:t} \in E^t$,

$$\mathbb{E}^{i,N}_{\mathbf{y}^{1:i}_{1:t}}[\hat{Z}_t / Z_t] \leq 1 + \frac{Z_t^{-1} \sum_{s=1}^{t} \sum_{j=1}^{i} G_{0,s} G_{s,t+1}(y^i_{1:t}) - \zeta i}{\zeta N} + \Theta\left(N^{-2}\right).$$

We suspect that using ∞-ESS is not only sufficient but necessary.

Conjecture 1. *Fix any $\varepsilon \in (0, 1)$. There exists a choice of $(E, \mathcal{E})$, $(M_t)_{t \geq 1}$, and $(g_t)_{t \geq 1}$ such that the following holds: For any sufficiently small $\zeta > 0$, under Assumption 4.D and the assumption that resampling only occurs when $\mathrm{ESS}_2(W^{1:N}_s) < \zeta N$,*

$$\mathbb{E}^{i,N}_{\mathbf{y}^{1:i}_{1:t}}[\hat{Z}_t / Z_t] > 1 + \frac{Z_t^{-1} \sum_{s=1}^{t} \sum_{j=1}^{i} G_{0,s} G_{s,t+1}(y^i_{1:t}) - ti}{\varepsilon \zeta N} + o\left(N^{-1}\right).$$

4.3. Quantitative bounds

Recall Assumptions 1.B and 1.C, either of which can be used in conjunction with Theorem 4.7 to obtain uniform, quantitative bounds on $\mathbb{E}^{i,N}_{\mathbf{y}^{1:i}_{1:t}}[\hat{Z}_t]$ by following the approach of Andrieu et al. [2]:

Assumption 1.B. The potentials satisfy $\overline{g}_s \triangleq \sup_{x\in E} g_s(x) < \infty$ for all $1 \le s \le t$.

Assumption 1.C. There exists a constant $\beta > 0$ such that for any $t, s \in \mathbb{N}$,

$$\sup_{x\in E} \frac{G_{0,t} G_{t,t+s}(x)}{G_{0,t+s}} \le \beta.$$

Assumption 1.C is implied by a standard "strong mixing" condition which is often employed in SMC analyses (e.g., [7,18]). See Andrieu et al. [2] for details.

Proposition 4.9. *If $\alpha_s = \mathbf{1}_{1/N}$ for $s \in [t-1]$ and Assumption 1.B holds, then for all $t \ge 1$, $i \ge 1$, $N \ge i$, $y^1_{1:t}, \dots, y^i_{1:t} \in E^t$,*

$$\mathbb{E}^{i,N}_{\mathbf{y}^{1:i}_{1:t}}[\hat{Z}_t/Z_t] \le 1 + \left[Z_t^{-1} \prod_{s=1}^{t} \overline{g}_s - 1 \right] \left[1 - \left(1 - \frac{i}{N} \right)^t \right]. \tag{4.6}$$

If Assumptions 1.A, 1.B and 4.D hold, then for all $t \ge 1$, $i \ge 1$, $N \ge i$, $y^1_{1:t}, \dots, y^i_{1:t} \in E^t$,

$$\mathbb{E}^{i,N}_{\mathbf{y}^{1:i}_{1:t}}[\hat{Z}_t/Z_t] \le 1 + Z_t^{-1} \prod_{s=1}^{t} \overline{g}_s \left[\left(1 + \frac{i}{\zeta N} \right)^t - 1 \right]. \tag{4.7}$$

Proof. The proof of equation (4.6) is a straightforward generalization of that for [2], Proposition 12, with some additional bookkeeping for i (instead of 2) fixed trajectories. As for equation (4.7), we have

$$\begin{aligned}
\mathbb{E}^{i,N}_{\mathbf{y}^{1:i}_{1:t}}[\hat{Z}_t] &\le \sum_{\ell=1}^{t+1} \sum_{\boldsymbol{\tau}\in\mathcal{T}_{t,\ell,t+1}} (\zeta N)^{-\ell+1} G_{0,\tau_1} \prod_{m=1}^{\ell-1} \sum_{j=1}^{i} G_{\tau_m,\tau_{m+1}}(y^j_{\tau_m}) \\
&\le Z_t + \prod_{s=1}^{t} \overline{g}_s \sum_{\ell=2}^{t+1} \binom{t}{\ell-1} i^{\ell-1} (\zeta N)^{-\ell+1} \\
&= Z_t + \prod_{s=1}^{t} \overline{g}_s \sum_{\ell=1}^{t} \binom{t}{\ell} (\zeta N/i)^{-\ell} \\
&= Z_t + \prod_{s=1}^{t} \overline{g}_s \left[\left(1 + \frac{i}{\zeta N} \right)^t - 1 \right].
\end{aligned}$$

□

Proposition 4.10. *If $\alpha_s = \mathbf{1}_{1/N}$ for $s \in [t-1]$ and Assumption 1.C holds, then for all $t \ge 1$, $i \ge 1$, $N \ge i$, $y^1_{1:t}, \dots, y^i_{1:t} \in E^t$,*

$$\mathbb{E}^{i,N}_{\mathbf{y}^{1:i}_{1:t}}[\hat{Z}_t/Z_t] \le \left(1 + \frac{i(\beta-1)}{N} \right)^t. \tag{4.8}$$

If Assumptions 1.A, 1.C and 4.D hold, then for all $t \geq 1$, $i \geq 1$, $N \geq i$, $y^1_{1:t}, \ldots, y^i_{1:t} \in E^t$,

$$\mathbb{E}^{i,N}_{\boldsymbol{y}^{1:i}_{1:t}}[\hat{Z}_t / Z_t] \leq \left(1 + \frac{i\beta}{\zeta N}\right)^t. \tag{4.9}$$

Proof. The proof of equation (4.8) is a simple generalization of that for [2], Proposition 14. As for equation (4.9), observe that for $s \in [t+1]$, $G_{0,t+1} = G_{0,t+1}\frac{G_{0,t+1}}{G_{0,s}} = G_{0,s}\eta_s(G_{s,t+1})$, so we can write for $\ell \in [t]$, $\boldsymbol{\tau} \in \mathcal{T}_{t,\ell,t+1}$,

$$Z_t = G_{0,t+1} = G_{0,\tau_k}\prod_{i=1}^{\ell-1}\eta_{\tau_i}(G_{\tau_i,\tau_{i+1}}).$$

Combined with Assumption 1.C and writing $\bar{G}_{s,t} \triangleq \sup_{x\in E} G_{s,t}(x)$,

$$\begin{aligned}
&\sum_{\ell=1}^{t+1}(\zeta N)^{-\ell+1}\sum_{\tau\in\mathcal{T}_{t,\ell,t+1}} G_{0,\tau_1}\prod_{i=1}^{\ell-1}\sum_{j=1}^{i} G_{\tau_m,\tau_{m+1}}\left(y^j_{\tau_m}\right) \\
&\quad \leq Z_t + Z_t\sum_{\ell=2}^{t+1}(\zeta N)^{-\ell+1}\sum_{\tau\in\mathcal{T}_{t,\ell,t+1}}\frac{G_{0,\tau_1}}{G_{0,\tau_1}}\prod_{i=1}^{\ell-1}\sum_{j=1}^{i}\frac{\bar{G}_{\tau_i,\tau_{i+1}}}{\eta_{\tau_i}(G_{\tau_i,\tau_{i+1}})} \\
&\quad = Z_t\sum_{\ell=1}^{t+1}\binom{t}{\ell-1}(\zeta N)^{-\ell+1}(i\beta)^{\ell-1} \\
&\quad = Z_t\left(1+\frac{i\beta}{\zeta N}\right)^t.
\end{aligned}$$

□

To compare equations (4.6) and (4.7), consider the $\Theta(1/N)$ terms, which are, respectively,

$$\frac{ti[Z_t^{-1}\prod_{s=1}^t \overline{g}_s - 1]}{N} \quad \text{and} \quad \frac{tiZ_t^{-1}\prod_{s=1}^t \overline{g}_s}{\zeta N}.$$

Thus, up to a $-ti$ term and a factor of $1/\zeta$, the two bounds are of the same leading order in $1/N$. The $-it$ is likely an artifact of the analysis while the $1/\zeta$ term accounts for there being only ζN "effective particles." The differences between equations (4.8) and (4.9) are identical.

5. Bounding the divergence of SMC samplers

Recall that $P^{0,N}(\mathrm{d}x_{1:t})$ is the distribution of $\tilde{X}^*_{1:t} \sim \pi^{0,N}_{1:t}$, a single sample from the αSMC estimator of $\pi_{1:t}$. Equivalently, $P^{0,N}(\mathrm{d}x_{1:t}) = \mathbb{E}^{0,N}[\pi^{0,N}_{1:t}](\mathrm{d}x_{1:t}) \triangleq \mathbb{E}^{0,N}[\pi^{0,N}_{1:t}(\mathrm{d}x_{1:t})]$ is the expected value of the random measure $\pi^{0,N}_{1:t}$. As a first application of our results from Section 4,

Table 1. Divergences of the form equation (5.1). The operator $(a)^+$ gives the positive part of $a \in \mathbb{R}$

Name	Symbol	i	f
KL divergence (version 1)	d_{KL}	1	$a \mapsto \log a$
KL divergence (version 2)	d_{KL}	2	$a \mapsto -\log a$
χ^2 distance	d_{χ^2}	1	$a \mapsto a - 1$
Total variation distance (version 1)	d_{TV}	2	$a \mapsto (a-1)^+$
Total variation distance (version 2)	d_{TV}	2	$a \mapsto (1-a)^+$

we consider bounding the distance between the measures $\pi_{1:t}$ and $P^{0,N}$. That is, for some divergence $d(\mu||\nu)$ between measures, can we bound $d(\pi_{1:t}||P^{0,N})$? To the best of our knowledge, there has been minimal investigation of this question, with [7], Chapter 8, a notable exception. For example, under Assumption SIR, the bound

$$d_{\mathrm{KL}}\big(P^{0,N}||\pi_t\big) \le \frac{c}{N},$$

can be extracted as a special case of a more general propagation-of-chaos result [7], Theorem 8.3.2.

Let $\mathcal{F}_1$ be the set of functions $f : \mathbb{R}_+ \to \mathbb{R}$ that are monotonically increasing or decreasing and satisfy $f(1) = 0$. We consider the class of *monotonic divergences* of the form

$$d_{i,f}(\mu_1||\mu_2) \triangleq \mu_i\left(f \circ \frac{\mathrm{d}\mu_1}{\mathrm{d}\mu_2}\right), \qquad i \in \{1,2\},\ f \in \mathcal{F}_1. \tag{5.1}$$

Table 1 lists some common divergences that can be written this way.

The following result characterizes the divergence between $\pi_{1:t}$ and $P^{0,N}$, only assuming that f is concave.

Proposition 5.1. *Let $\mathcal{R}_t(y_{1:t}) \triangleq \mathbb{E}^{1,N}_{y_{1:t}}[\hat{Z}_t/Z_t]$ and $\mathcal{S}_t \triangleq Z_t^{-1}\sum_{s=1}^t G_{0,s}\pi_{1:t}(G_{s,t+1}) - \zeta$. If Assumption 4.D holds, then for all concave $f \in \mathcal{F}_1$,*

$$d_{1,f}\big(\pi_{1:t}||P^{0,N}\big) \le f\big(\pi_{1:t}(\mathcal{R}_t)\big).$$

In particular, if Assumption SIR holds, then

$$d_{\mathrm{KL}}\big(\pi_{1:t}||P^{0,N}\big) \le \log\left(1 + \frac{\mathcal{S}_t - t}{N} + \Theta\big(N^{-2}\big)\right),$$

$$d_{\chi^2}\big(\pi_{1:t}||P^{0,N}\big) \le \frac{\mathcal{S}_t - t}{N} + \Theta\big(N^{-2}\big)$$

while if Assumptions 1.A and 4.D hold, then

$$d_{\mathrm{KL}}\big(\pi_{1:t}||P^{0,N}\big) \le \log\left(1 + \frac{\mathcal{S}_t - \zeta}{\zeta N} + \Theta\big(N^{-2}\big)\right), \qquad d_{\chi^2}\big(\pi_{1:t}||P^{0,N}\big) \le \frac{\mathcal{S}_t - \zeta}{\zeta N} + \Theta\big(N^{-2}\big).$$

Proof. The general statement follows by applying Jensen's inequality to equation (5.1), then using Theorem 4.2. The special cases correspond to using the KL divergence (version 1) and χ^2 distance rows of Table 1 and applying Corollary 4.8. □

Similar results for (sequential) importance sampling are included in Appendix A.5.

We can also consider the divergence between $\pi_{1:t}$ and $P^{0,N}$ when $\mathbb{E}^{1,N}_{y_{1:t}}[\hat{Z}_t]$ is uniformly bounded:

Proposition 5.2. *If Assumption 4.D holds and $\mathbb{E}^{1,N}_{y_{1:t}}[\hat{Z}_t/Z_t] \leq B_{t,N}$ for all $y_{1:t} \in E^t$, then for all increasing $f \in \mathcal{F}_1$*

$$d_{i,f}\left(\pi_{1:t} || P^{0,N}\right) \leq f(B_{t,N})$$

and for all decreasing $f \in \mathcal{F}_1$,

$$d_{i,f}\left(P^{0,N} || \pi_{1:t}\right) \leq f\left(B_{t,N}^{-1}\right).$$

In particular,

$$\begin{aligned} d_{\mathrm{KL}}\left(\pi_{1:t} || P^{0,N}\right) &\leq \log B_{t,N}, \\ d_{\chi^2}\left(\pi_{1:t} || P^{0,N}\right) &\leq B_{t,N} - 1, \\ d_{\mathrm{TV}}\left(\pi_{1:t}, P^{0,N}\right) &\leq \frac{B_{t,N} - 1}{B_{t,N}} \leq B_{t,N} - 1. \end{aligned}$$

Proof. The general statements follow immediately from equation (5.1) and Theorem 4.2. The special cases correspond to using the KL divergence (version 1), χ^2 distance, and total variation distance (version 2) rows of Table 1. The second total variation inequality holds since $B_{t,N} \geq 1$. □

The bounds in Proposition 5.2 for KL divergence, χ^2 distance, and total variation distance are asymptotically equivalent if $B_{t,N} \to 1$ as $N \to \infty$. Combining Proposition 5.2 with, for example, Proposition 4.10, yields quantitative bounds for SIR and αSMC:

Corollary 5.3. *If Assumptions 1.C and SIR hold, then*

$$\begin{aligned} d_{\mathrm{KL}}\left(\pi_{1:t} || P^{0,N}\right) &\leq \frac{t(\beta - 1)}{N}, \\ d_{\chi^2}\left(\pi_{1:t} || P^{0,N}\right) &\leq \frac{t(\beta - 1)}{N} + O\left(N^{-2}\right), \\ d_{\mathrm{TV}}\left(\pi_{1:t}, P^{0,N}\right) &\leq \frac{t(\beta - 1)}{N + t(\beta - 1)} + O\left(N^{-2}\right). \end{aligned}$$

If Assumptions 1.A, 1.C *and* 4.D *hold, then*

$$d_{\mathrm{KL}}\big(\pi_{1:t}||P^{0,N}\big) \le \frac{t\beta}{\zeta N},$$

$$d_{\chi^2}\big(\pi_{1:t}||P^{0,N}\big) \le \frac{t\beta}{\zeta N} + O\big(N^{-2}\big),$$

$$d_{\mathrm{TV}}\big(\pi_{1:t}, P^{0,N}\big) \le \frac{t\beta}{\zeta N + t\beta} + O\big(N^{-2}\big).$$

6. The α-particle Gibbs sampler

As a second application of our results from Section 4, we consider the mixing properties of particle Gibbs with adaptive resampling. Recall from Section 1.3 that we introduce a global parameter $\theta \in \Theta$ with prior distribution $\varpi(\mathrm{d}\theta)$. Replace M_s by M_s^θ and g_s by g_s^θ, then parameterize the other quantities defined previously in terms of M_s and g_s by θ. The target distribution on the product space $(\Theta \times Y, \mathcal{B}(\Theta \times Y))$, $Y \triangleq E^t$, is

$$\pi(\mathrm{d}\theta \times \mathrm{d}y) \triangleq \gamma(\mathrm{d}\theta \times \mathrm{d}y)/Z,$$

where

$$\gamma(\mathrm{d}\theta \times \mathrm{d}y) \triangleq \prod_{s=1}^{t} g_s^\theta(y_s) M_s^\theta(y_{s-1}, \mathrm{d}y_s)\varpi(\mathrm{d}\theta) \quad \text{and} \quad Z \triangleq \gamma(1).$$

Particle Gibbs samplers have kernels of the form $\pi_y(\mathrm{d}\theta)\Pi_\theta(y, \mathrm{d}z)$, where $\Pi_\theta(y, \mathrm{d}z)$ is an SMC-based kernel with invariant distribution π_θ. The standard PG sampler employs the iterated conditional SMC (i-cSMC) kernel [2]: that is, $\Pi_\theta = P_{\theta,y}^{1,N}$ and require Assumption SIR to hold.

We now introduce the novel α-particle Gibbs (αPG) sampler, which employs the iterated conditional αSMC (i-cαSMC) kernel $P_{\theta,y}^{1,N}$, so $\Pi_\theta = P_{\theta,y}^{1,N}$. In Appendix A.6, we prove that the i-cαSMC kernel is reversible with respect to π_θ and hence has invariant distribution π_θ.

The first step to proving mixing results for the i-cαSMC kernel and the αPG sampler is to use Theorem 4.2 to obtain a sufficient condition for the i-cαSMC transition kernel to satisfy a minorization condition.

Proposition 6.1. *If Assumption* 4.D *holds and* $\mathbb{E}_{\theta,y^{1:2}}^{2,N}[\hat{Z}_t/Z_t] \le B_{t,N}$ *for all* $\theta \in \Theta$ *and* $y^1, y^2 \in Y$, *then*

$$P_{\theta,y}^{1,N}(y, \mathrm{d}x) \ge \varepsilon_{t,N}\pi_\theta(\mathrm{d}x), \tag{6.1}$$

where $\varepsilon_{t,N} \triangleq \frac{(1-\kappa_N')^t}{B_{t,N}}$.

The constant $\varepsilon_{t,N}$, which determines mixing speed, can be found explicitly using the quantitative bounds from Section 4.3. For example, using Assumption 1.C we obtain the following corollary.

Corollary 6.2. *If Assumptions 1.A, 1.C and 4.D and hold, then for all $y \in Y$,*

$$P^{1,N}_{\theta,y}(y, \mathrm{d}x) \geq \varepsilon_{t,N}\pi_\theta(\mathrm{d}x),$$

where

$$\varepsilon_{t,N} \triangleq \left(\frac{1-\kappa'_N}{1+\frac{2\beta}{\zeta N}}\right)^t.$$

Furthermore, if $\zeta N \geq \frac{2\beta t}{C(1-\kappa'_N)-\kappa'_N t}$ for some constant $C > 0$, then

$$\varepsilon_{t,N} \geq e^{-C}.$$

In particular, assuming $\kappa'_N \leq B/N$ for some constant $B \geq 1$, if $N \geq t/C + B$, then

$$\varepsilon_{t,N} \geq \exp\left(-\frac{C}{\zeta}(2\beta + \zeta B)\right).$$

Proof. The first part follows from Propositions 6.1 and 4.10. For the second part, we then have

$$\begin{aligned}\varepsilon_{t,N} &\geq \left(\frac{1+\frac{2\beta}{\zeta N}}{1-\kappa'_N}\right)^{-t} = \left(1 + \frac{1}{1-\kappa'_N}\left(\frac{2\beta}{\zeta N} + \kappa'_N\right)\right)^{-t} \\ &\geq \left(1 + \frac{C}{t}\right)^{-t} \geq e^{-C}.\end{aligned}$$

The final part follows after noting that if $\kappa'_N \leq B/N$, then

$$\frac{1}{1-\kappa'_N}\left(\frac{2\beta}{\zeta N} + \kappa'_N\right) \leq \frac{1}{1-B/N}\left(\frac{2\beta}{\zeta N} + B/N\right) = \frac{1}{N-B}\left(\frac{2\beta}{\zeta} + B\right). \qquad \square$$

Remark 6.3. In the case of the i-cSMC kernel, Corollary 6.2 is almost as tight as [2], Corollary 14: the former result replaces $\beta - 1$ with β.

The minorization condition (6.1) implies uniform ergodicity and a number of other types of convergence guarantees for the i-cSMC process. The following generalizes [2], Theorem 1, which applies only to the i-cSMC kernel and the PG sampler.

Theorem 6.4. *Assume that Assumptions 1.A and 4.D hold.*

I. *Let $N \geq 2$, and consider the i-cαSMC process with kernel $P = P^{1,N}_\theta$.*
   1. *P is reversible with respect to π and defines a positive operator.*
   2. *If the potentials are bounded, then there exists $\varepsilon_{t,N} = 1 - C_t/N$ such that*
      (i) *for all $y \in Y$, $P(y, \mathrm{d}z) \geq \varepsilon_{t,N}\pi_\theta(\mathrm{d}z)$,*

(ii) *for every measure* $\nu \ll \pi_\theta$ *and* $k \geq 1$,

$$d_{\chi^2}\big(\nu P^k || \pi_\theta\big) \leq d_{\chi^2}(\nu || \pi_\theta)(1 - \varepsilon_{t,N})^k,$$

(iii) *for every* $y \in Y$ *and* $k \geq 1$,

$$d_{\mathrm{TV}}\big(\delta_y P^k, \pi_\theta\big) \leq (1 - \varepsilon_{t,N})^k.$$

3. *If Assumption 1.C also holds and there is a constant* $B > 0$ *such that* $\kappa'_N \leq B/N$, *then for every* $C > 0$, *there exists* $\varepsilon_{B,C,\zeta} > 0$ *such that for* $N \geq t/C + B$ *and all* $t > 1$,

$$\varepsilon_{t,N} \geq \varepsilon_{B,C,\zeta} > 0.$$

II. *If there exists* $\beta \geq 1$ *such that, for all* $t, s \in \mathbb{N}$,

$$\pi\text{-}\operatorname*{ess\,sup}_{\theta,x} \frac{G^\theta_{0,t} G^\theta_{t,t+s}(x)}{G^\theta_{0,t+s}} \leq \beta,$$

or if

$$\pi\text{-}\operatorname*{ess\,sup}_{\theta} \frac{\prod_{s=1}^t \overline{g}^\theta_s}{\gamma_\theta(1)} < \infty,$$

then the α*PG chain is geometrically ergodic whenever the Gibbs sampler is geometrically ergodic.*

Proof. Part I.1 follows from Lemma A.4. Parts I.2–3 follow from Proposition 6.1, Corollary 6.2, and [2], Proposition 31. Part II follows from [2], Section 7. □

Remark 6.5. Part I.3 means that if Assumption 1.C holds, then scaling N linearly with t ensures a uniform convergence rate, as measured by χ^2-divergence or total variation distance.

Appendix: Additional proofs

A.1. A technical lemma

Lemma A.1. *Let* X *and* Y *be random elements in Borel spaces* $(S, \mathcal{S})$ *and* $(T, \mathcal{T})$, *respectively, let* $\psi : S \times T \to \mathbb{R}_+$ *be a measurable, and let* μ *be the distribution of* X. *If*

$$\nu = \mathbb{E}\big[\psi(X, Y)\delta_X\big],$$

then $\nu \ll \mu$ *and*

$$\frac{\mathrm{d}\nu}{\mathrm{d}\mu}(X) = \mathbb{E}\big[\psi(X, Y) \mid X\big] \qquad a.s.$$

Proof. Because S is Borel, there exists an f satisfying $f(X) = \mathbb{E}[\psi(X, Y) \mid X]$ a.s. It follows from the chain rule of conditional expectation and then some elementary manipulations that, for all $A \in \mathcal{S}$,

$$\nu(A) = \mathbb{E}\big[f(X)\delta_X(A)\big] = \mathbb{E}\big[f(X)\mathbb{1}_A(X)\big] = \int_A f(x)\mu(\mathrm{d}x),$$

and so f is a version of the Radon–Nikodym derivative $\mathrm{d}\nu/\mathrm{d}\mu$. □

A.2. Proof of Theorem 4.2, $i \geq 1$ case

First, observe that we can write the c$^i\alpha$SMC kernel as

$$\begin{aligned} P^{i,N}\big(\boldsymbol{y}_{1:t}^{1:i}, \mathrm{d}x_{1:t}\big) &= \mathbb{E}^{i,N}_{\boldsymbol{y}_{1:t}^{1:i}}\big[\delta_{\tilde{X}^*_{1:t}}(\mathrm{d}x_{1:t})\big] \\ &= \mathbb{E}^{i,N}_{\boldsymbol{y}_{1:t}^{1:i}}\Bigg[\sum_{k_{1:t}\in[N]^t} \mathcal{I}_{k_{1:t}}\big(\boldsymbol{X}_{1:t}^{1:N}, \boldsymbol{A}_{1:t}^{1:N}, \mathrm{d}x_{1:t}\big)\Bigg], \end{aligned}$$

where

$$\mathcal{I}_{k_{1:t}}\big(\boldsymbol{x}_{1:t}^{1:N}, \boldsymbol{a}_{1:t}^{1:N}, \mathrm{d}x_{1:t}\big) \triangleq \delta_{\tilde{x}_{1:t}^{k_t}}(\mathrm{d}x_{1:t})\mathbb{1}\big(k_t = a_t^*\big)\prod_{s=1}^{t-1}\mathbb{1}\big(k_s = a_s^{k_{s+1}}\big).$$

Next, note that

$$\begin{aligned} &\sum_{k_1=1}^{N}\mathbb{1}\big(\tilde{x}_{1:t}^{k_t}\in S\big)\mathbb{P}^{i,N}_{\boldsymbol{y}_{1:t}^{1:i}}\big[X_1^{1:N}\in \mathrm{d}x_1^{1:N}, F_1^{1:i} = f_1^{1:i}\big] \\ &\quad= \sum_{k_1=1}^{N}\mathbb{1}\big(\tilde{x}_{1:t}^{k_t}\in S\big)\mathcal{C}_1^i\mathcal{D}\big(f_1^{1:i}\big)\prod_{j=1}^{i}\frac{1}{N}\delta_{y_1^j}\big(\mathrm{d}x_1^{f_1^j}\big)\prod_{n\notin f_1^{1:i}}^{N}M_1\big(\mathrm{d}x_1^n\big) \\ &\quad\geq \frac{N\mathcal{C}_1^i}{\mathcal{C}_1^{i+1}}\sum_{k_1=1}^{N}\int_E \mathbb{1}\big(\tilde{x}_{1:t}^{k_t}\in S\big)\mathcal{C}_1^{i+1}\mathcal{D}\big(f_1^{1:i}, k_1\big)\prod_{j=1}^{i}\frac{1}{N}\delta_{y_1^j}\big(\mathrm{d}x_1^{f_1^j}\big)\frac{1}{N}\delta_{x_1}\big(\mathrm{d}x_1^{k_1}\big) \qquad \text{(A.1)} \\ &\qquad\times\prod_{n\notin f_1^{1:i},k_1}^{N}M_1\big(\mathrm{d}x_1^n\big)M_1(\mathrm{d}x_1) \\ &\quad\geq \frac{N\mathcal{C}_1^i}{\mathcal{C}_1^{i+1}}\sum_{k_1=1}^{N}\int_E \mathbb{1}\big(\tilde{x}_{1:t}^{k_t}\in S\big)\mathbb{P}^{i+1,N}_{\boldsymbol{y}_{1:t}^{1:i},x_{1:t}}\big[X_1^{1:N}\in \mathrm{d}x_1^{1:N}, F_1^{1:i} = f_1^{1:i}, F_1^{i+1} = k_1\big]M_1(\mathrm{d}x_1). \end{aligned}$$

For the remainder of the proof, to keep notation compact when writing laws, instead of writing, for example, $X_s^{1:N} \in x_s^{1:N}$ or $F_s^{1:i} = f_s^{1:i}$, whenever a random variable is instantiated to be

(the differential) of the lowercase version of itself, we will write only the random variable: for example, $X_s^{1:N}$ or $F_s^{1:i}$. Now, for $s = 2, \dots, t$,

$$
\begin{aligned}
&\sum_{k_s=1}^{N} \mathbb{1}\big(\tilde{x}_{1:t}^{k_t} \in S\big)\mathbb{1}\big(k_{s-1} = a_{s-1}^{k_s}\big)\mathbb{P}_{\boldsymbol{y}_{1:t}^{1:i}}^{i,N}\big[X_s^{1:N}, A_{s-1}^{1:N}, F_s^{1:i} \mid \boldsymbol{X}_{1:s-1}^{1:N}, \boldsymbol{A}_{1:s-2}^{1:N}, F_{s-1}^{1:i}\big] \\
&\quad \triangleq \sum_{k_s=1}^{N} \mathbb{1}\big(\tilde{x}_{1:t}^{k_t} \in S\big)\mathbb{1}\big(k_{s-1} = a_{s-1}^{k_s}\big)\mathcal{C}_s^i \mathcal{D}\big(f_s^{1:i}\big) \prod_{j=1}^{i} \alpha_{s-1}^{f_s^j f_{s-1}^j} \delta_{y_s^j}\big(\mathrm{d}x_s^{f_s^j}\big)\mathbb{1}\big(a_s^{f_{s-1}^j} = f_{s-1}^j\big) \\
&\qquad \times \prod_{n \notin f_s^{1:i}} r_n\big(a_{s-1}^n | w_{s-1}^{1:N}, \boldsymbol{x}_{1:s-1}^{1:N}\big) M_s\big(x_{s-1}^{a_{s-1}^n}, x_s^n\big) \\
&\quad \geq \sum_{k_s=1}^{N} \frac{\mathcal{C}_s^i}{\mathcal{C}_s^{i+1}\alpha_{s-1}^{k_s k_{s-1}}} \int_E \mathbb{1}\big(\tilde{x}_{1:t}^{k_t} \in S\big)\mathcal{C}_s^{i+1}\mathcal{D}\big(f_s^{1:i}, k_s\big) \\
&\qquad \times \prod_{j=1}^{i} \alpha_{s-1}^{f_s^j f_{s-1}^j} \delta_{y_s^j}\big(\mathrm{d}x_s^{f_s^j}\big)\mathbb{1}\big(a_s^{f_{s-1}^j} = f_{s-1}^j\big) \\
&\qquad \times \alpha_{s-1}^{k_s k_{s-1}} \delta_{x_s}\big(\mathrm{d}x_s^{k_s}\big)\mathbb{1}\big(k_{s-1} = a_{s-1}^{k_s}\big) \\
&\qquad \times \prod_{n \notin f_s^{1:i}, k_s} r_n\big(a_{s-1}^n | w_{s-1}^{1:N}, \boldsymbol{x}_{1:s-1}^{1:N}\big) M_s\big(x_{s-1}^{a_{s-1}^n}, x_s^n\big) \\
&\qquad \times r_{k_s}\big(k_{s-1} | w_{s-1}^{1:N}, \boldsymbol{x}_{1:s-1}^{1:N}\big) M_s\big(x_{s-1}^{k_{s-1}}, \mathrm{d}x_s\big) \\
&\quad = \sum_{k_s=1}^{N} \frac{\mathcal{C}_s^i}{\mathcal{C}_s^{i+1}\alpha_{s-1}^{k_s k_{s-1}}} \int_E \mathbb{1}\big(\tilde{x}_{1:t}^{k_t} \in S\big) r_{k_s}\big(k_{s-1} | w_{s-1}^{1:N}, \boldsymbol{x}_{1:s-1}^{1:N}\big) M_s\big(x_{s-1}^{k_{s-1}}, \mathrm{d}x_s\big) \\
&\qquad \times \mathbb{P}_{\boldsymbol{y}_{1:t}^{1:i}, x_{1:t}}^{i+1,N}\big[X_s^{1:N}, A_{s-1}^{1:N}, F_s^{1:i}, F_s^{i+1} = k_s \mid \boldsymbol{X}_{1:s-1}^{1:N}, \boldsymbol{A}_{1:s-2}^{1:N}, F_{s-1}^{1:i}, F_{s-1}^{i+1} = k_{s-1}\big].
\end{aligned}
\tag{A.2}
$$

Using equations (A.1) and (A.2), we have (note that the terms such as those involving a_0 should be ignored)

$$
\begin{aligned}
&\sum_{k_{1:t} \in [N]^t} \mathbb{1}\big(\tilde{x}_{1:t}^{k_t} \in S\big)\mathbb{1}\big(k_t = a_t^*\big) \prod_{s=2}^{t} \mathbb{1}\big(k_{s-1} = a_{s-1}^{k_s}\big)\mathbb{P}_{\boldsymbol{y}_{1:t}^{1:i}}^{i,N}\big[\boldsymbol{X}_{1:t}^{1:N}, \boldsymbol{A}_{1:t}^{1:N}, \boldsymbol{F}_{1:t}^{1:i}\big] \\
&\quad = \sum_{k_{1:t} \in [N]^t} \mathbb{1}\big(\tilde{x}_{1:t}^{k_t} \in S\big)\mathbb{1}\big(k_t = a_t^*\big)\mathbb{P}_{\boldsymbol{y}_{1:t}^{1:i}}^{i,N}\big[A_t^* \mid \boldsymbol{X}_{1:t}^{1:N}, \boldsymbol{A}_{1:t-1}^{1:N}\big]\mathbb{P}_{\boldsymbol{y}_{1:t}^{1:i}}^{i,N}\big[X_1^{1:N}, F_1^{1:i}\big] \\
&\qquad \times \prod_{s=2}^{t} \mathbb{1}\big(k_{s-1} = a_{s-1}^{k_s}\big)\mathbb{P}_{\boldsymbol{y}_{1:t}^{1:i}}^{i,N}\big[X_s^{1:N}, A_{s-1}^{1:N}, F_s^{1:i} \mid \boldsymbol{X}_{1:s-1}^{1:N}, \boldsymbol{A}_{1:s-2}^{1:N}, F_{s-1}^{1:i}\big]
\end{aligned}
$$

$$
\begin{aligned}
&\geq \sum_{k_{1:t}\in[N]^t} \int_{E^t} \frac{N\,\mathbb{1}(\tilde{x}_{1:t}^{k_t}\in S)\mathbb{1}(k_t=a_t^*)}{(\prod_{s=1}^t \mathcal{C}_s^{i+1}/\mathcal{C}_s^i)(\prod_{s=1}^{t-1}\alpha_{s-1}^{k_s k_{s-1}})} \frac{w_t^{a_t^*} g_t(x_t^{a_t^*})}{\sum_{n=1}^N w_t^n g_t(x_t^n)} \\
&\quad \times \prod_{s=2}^t r_{k_s}\big(k_{s-1}|w_{s-1}^{1:N}, \boldsymbol{x}_{1:s-1}^{1:N}\big) M_s\big(x_{s-1}^{k_{s-1}}, \mathrm{d}x_s\big) \\
&\quad \times \prod_{s=2}^t \mathbb{P}^{i+1,N}_{\boldsymbol{y}_{1:t}^{1:i},x_{1:t}}\big[X_s^{1:N}, A_{s-1}^{1:N}, F_s^{1:i}, F_s^{i+1}=k_s \mid \boldsymbol{X}_{1:s-1}^{1:N}, \boldsymbol{A}_{1:s-2}^{1:N}, F_{s-1}^{1:i}, F_{s-1}^{i+1}=k_{s-1}\big] \\
&\quad \times \mathbb{P}^{i+1,N}_{\boldsymbol{y}_{1:t}^{1:i},x_{1:t}}\big[X_1^{1:N}\in \mathrm{d}x_1^{1:N}, F_1^{1:i}=f_1^{1:i}, F_1^{i+1}=k_1\big] M_1(\mathrm{d}x_1) \\
&= \sum_{k_{1:t}\in[N]^t} \int_{E^t} \frac{N\,\mathbb{1}(x_{1,t}\in S)}{(\prod_{s=1}^t \mathcal{C}_s^{i+1}/\mathcal{C}_s^i)\sum_{n=1}^N w_t^n g_t(x_t^n)} \\
&\quad \times \mathbb{P}^{i+1,N}_{\boldsymbol{y}_{1:t}^{1:i},x_{1:t}}\big[\boldsymbol{X}_{1:t}^{1:N}, \boldsymbol{A}_{1:t-1}^{1:N}, \boldsymbol{F}_{1:t}^{1:i}, F_{1:t}^{i+1}=k_{1:t}\big]\gamma_{1:t}(\mathrm{d}x_{1:t}) \\
&= \int_S \sum_{k_{1:t}\in[N]^t} \frac{Z_t}{\hat{Z}_t \prod_{s=1}^t \mathcal{C}_s^{i+1}/\mathcal{C}_s^i} \mathbb{P}^{i+1,N}_{\boldsymbol{y}_{1:t}^{1:i},x_{1:t}}\big[\boldsymbol{X}_{1:t}^{1:N}, \boldsymbol{A}_{1:t-1}^{1:N}, \boldsymbol{F}_{1:t}^{1:i}, F_{1:t}^{i+1}=k_{1:t}\big]\pi_{1:t}(\mathrm{d}x_{1:t}),
\end{aligned}
$$

from which equation (4.1) follows.

To prove equation (4.2), first note that under Assumption 4.D, the normalization constants for the c$^2\alpha$SMC process are given by

$$
\mathcal{C}_1^2 \triangleq \frac{N}{N-1}
$$

and, for $s=2,\dots,t$,

$$
\mathcal{C}_s^2 \triangleq \left(1-\sum_{k=1}^N \alpha_{s-1}^{kf_{s-1}^1}\alpha_{s-1}^{kf_{s-1}^2}\right)^{-1}.
$$

Thus, $\mathcal{C}_s^2 \leq \frac{1}{1-\kappa_N}$ for $s=2,\dots,t$ and hence $\mathcal{C}_s^2 \leq \frac{1}{1-\kappa_N'}$ for all $s\in[t]$.

A.3. Proof of Proposition 4.6

For (1), the fact that $\mathrm{ESS}_1(w^{1:N}) = \mathrm{ESS}_{\mathrm{ent}}(w^{1:N})$ is a straightforward algebraic manipulation. To prove the limit equality, observe that, using the Taylor series for x^p and $\log(1+x)$, we have

$$
\begin{aligned}
&\lim_{p\to 1}\left(\frac{\|w^{1:N}\|_1^p}{\sum_{n=1}^N (w^n)^p}\right)^{1/(p-1)} \\
&\quad = \lim_{p\to 1}\left(\frac{\sum_{k=0}^\infty \|w^{1:N}\|_1 (p-1)^k \log^k(\|w^{1:N}\|_1)/k!}{\sum_{n=1}^N \sum_{k=0}^\infty w^n (p-1)^k \log^k(w^n)/k!}\right)^{1/(p-1)}
\end{aligned}
$$

$$
\begin{aligned}
&= \lim_{x\to\infty}\left(\frac{\sum_{k=0}^{\infty}\|w^{1:N}\|_1 x^{-k}\log^k(\|w^{1:N}\|_1)/k!}{\sum_{n=1}^{N}\sum_{k=0}^{\infty} w^n x^{-k}\log^k(w^n)/k!}\right)^x \\
&= \lim_{x\to\infty}\left(\frac{\exp(\log(1+\sum_{k=1}^{\infty} x^{-k}\log^k(\|w^{1:N}\|_1)/k!))}{\exp(\log(1+\sum_{k=1}^{\infty}\sum_{n=1}^{N} w^n\|w^{1:N}\|_1^{-1}x^{-k}\log^k(w^n)/k!))}\right)^x \\
&= \lim_{x\to\infty}\frac{\exp(x\sum_{m=1}^{\infty}(-1)^{m+1}[\sum_{k=1}^{\infty} x^{-k}\log^k(\|w^{1:N}\|_1)/k!]^m)}{\exp(x\sum_{m=1}^{\infty}(-1)^{m+1}[\sum_{k=1}^{\infty}\sum_{n=1}^{N} w^n\|w^{1:N}\|_1^{-1}x^{-k}\log^k(w^n)/k!]^m)} \\
&= \lim_{x\to\infty}\frac{\exp(\log(\|w^{1:N}\|_1)+\Theta(x^{-1}))}{\exp(\log(\prod_{n=1}^{N}(w^n)^{w^n/\|w^{1:N}\|_1})+\Theta(x^{-1}))} \\
&= \frac{\|w^{1:N}\|_1}{\prod_{n=1}^{N}(w^n)^{w^n/\|w^{1:N}\|_1}}.
\end{aligned}
$$

To prove the remaining parts, we make repeated use of the following.

Fact. *For $1 \le r < s \le \infty$, and any vector $w^{1:N} \in \mathbb{R}_+^N$, $\|w^{1:N}\|_s \le \|w^{1:N}\|_r \le N^{1/r-1/s}\|w^{1:N}\|_s$, with the lower (upper) bound achieved if and only if $w^{1:N}$ has one non-zero entry ($w^{1:N}$ has all equal entries).*

For (2), apply the Fact with $r = 1$, $s = p > 1$, and note that in this case $1/r - 1/s = 1 - 1/p = 1/p_*$. We then have $1 \le \|w^{1:N}\|_1/\|w^{1:N}\|_p \le N^{1/p_*}$, proving the result for $p > 1$. For $p = 1$, the result follows from part (1) and elementary properties of the entropy.

For (3), in the case that $p > 1$, note that

$$
\begin{aligned}
\left\|w^{1:N}\right\|_1^{q_*-p_*} &\ge N^{(q_*-p_*)/q_*}\left\|w^{1:N}\right\|_q^{q_*-p_*} = N^{1-p_*/q_*}\left\|w^{1:N}\right\|_q^{q_*-p_*} \\
&= N^{-p_*(1/p-1/q)}\left\|w^{1:N}\right\|_q^{q_*-p_*},
\end{aligned}
\tag{A.3}
$$

where the final equality follows since

$$
1 - p_*/q_* = 1 - p_*(1-1/q) = 1 - p_* + p_*/q = -p_*/p + p_*/q.
$$

We conclude that

$$
\begin{aligned}
\left(\frac{\|w^{1:N}\|_1}{\|w^{1:N}\|_p}\right)^{p_*} &\ge \frac{\|w^{1:N}\|_1^{p_*}}{N^{p_*(1/p-1/q)}\|w^{1:N}\|_q^{p_*}} \\
&\ge \frac{\|w^{1:N}\|_1^{p_*}}{N^{p_*(1/p-1/q)}\|w^{1:N}\|_q^{p_*}}\,\frac{\|w^{1:N}\|_1^{q_*-p_*}}{N^{-p_*(1/p-1/q)}\|v\|_q^{q_*-p_*}} \\
&= \left(\frac{\|w^{1:N}\|_1}{\|w^{1:N}\|_q}\right)^{q_*} \\
&\ge \left(\frac{\|w^{1:N}\|_1}{\|w^{1:N}\|_p}\right)^{q_*}
\end{aligned}
$$

$$= \left(\frac{\|w^{1:N}\|_p}{\|w^{1:N}\|_1}\right)^{p_*-q_*} \left(\frac{\|w^{1:N}\|_1}{\|w^{1:N}\|_p}\right)^{p_*}$$

$$\geq N^{-(p_*-q_*)/p_*} \left(\frac{\|w^{1:N}\|_1}{\|w^{1:N}\|_p}\right)^{p_*},$$

where the first, third, and fourth inequalities follow from the Fact and the second follows from equation (A.3).

The case of $p = 1$ follows from the $p > 1$ case and part (1).

A.4. Proof of Proposition 4.7

We prove the result for $i = 1$. The general case follows from straightforward modifications.

For $t \geq 1$, let $Q_t(x_{t-1}, \mathrm{d}x_t) \triangleq g_{t-1}(x_{t-1}) M_t(x_{t-1}, \mathrm{d}x_t)$, and for $0 \leq s < t$, let

$$Q_{s,t} \triangleq Q_{s+1} Q_{s+2} \cdots Q_t,$$

so $Q_{t,t+1} = Q_t$. By convention $Q_{t,t}(x_t, \mathrm{d}y_t) = \delta_{x_t}(\mathrm{d}y_t)$ and $Q_{0,t}(\mathrm{d}x_t)$ is a measure, not a probability kernel. Notice that for $s \in [t]$, $x_s \in E$, and $\phi_t : E \to \mathbb{R}$,

$$Q_{s,t}(x_s)(\phi_t) = \mathbb{E}\big[\phi_t(\xi_t) g_{s:t-1}(\xi_{s:t-1}) \mid \xi_s = x_s\big]$$

and $Q_{0,t}(\phi_t) = M_1 Q_{1,t}(\phi_t)$. Generalizing these identities, we will abuse notation and write, for $s \in [t]$, $x_s \in E$, and $\phi_{s,t} : E^{t-s} \to \mathbb{R}$,

$$Q_{s,t}(x_s)(\phi_{s,t}) \triangleq \mathbb{E}\big[\phi_{s:t}(\xi_{s:t}) g_{s:t-1}(\xi_{s:t-1}) \mid \xi_s = x_s\big]$$

and $Q_{0,t}(\phi_{1:t}) \triangleq M_1 Q_{1,t}(\phi_{1:t})$. Note that $G_{s,t}(y) = Q_{s,t}(y)(1)$ for $s \in [t-1]$ and $G_{0,t} = Q_{0,t}(1)$.

We will use the abbreviated notation $Q^k_{s,t}(\cdot) = Q_{s,t}(\cdot)(X^k_s)$ or $Q_{s,t}(\cdot)(x^k_s)$, $G^k_{s,t} = G_{s,t}(X^k_s)$ or $G_{s,t}(x^k_s)$, $G^y_{s,t} = G_{s,t}(y_s)$, $g^k_s = g_s(X^k_s)$ or $g_s(x^k_s)$, and $g^y_s = g_s(y_s)$. The variables are X^k_s inside expectations and x^k_s outside expectations. Throughout the proof, when limits of a sum are not specified, the sum is from 1 to N.

Let $\mathcal{F}_s$ be the σ-algebra generated by $\boldsymbol{X}^{1:N}_{1:s}$, $\boldsymbol{A}^{1:N}_{1:s-1}$, and $F^1_{1:s}$, where by convention we let $\mathcal{F}_0$ be the trivial σ-algebra. The proof relies on the following lemma.

Lemma A.2. *If $y_{1:t} \in E^t$, then*

1. *for $s = 2, \ldots, t$ and any functions $\phi^n_s : E \to \mathbb{R}$, $n \in [N]$,*

$$\mathbb{E}^{1,N}_{y_{1:t}}\left[\sum_n \phi^n_s\big(X^n_s\big) \,\middle|\, \mathcal{F}_{s-1}\right]$$

$$= \sum_{f^1_s} \alpha^{f^1_s f^1_{s-1}}_{s-1} \phi^{f^1_s}_s(y_s) + \sum_{f^1_s} \sum_{n \neq f^1_s} \sum_k \alpha^{f^1_s f^1_{s-1}}_{s-1} \frac{\alpha^{nk}_{s-1} w^k_{s-1}}{w^n_s} Q^k_{s-1,s}\big(\phi^n_s\big);$$

2. *for* $\tau \in [t-s]$,

$$\mathbb{E}^{1,N}_{y_{1:t}}\left[\sum_n W^n_s G^n_{s,s+\tau} \,\middle|\, \mathcal{F}_{s-1}\right] \leq \frac{1}{\zeta N}\sum_n w^n_{s-1} g^n_{s-1} G^y_{s,s+\tau} + \sum_n w^n_{s-1} G^n_{s-1,s+\tau};$$

and

3. *for* $s = 1, \ldots, t-1$,

$$N\,\mathbb{E}^{1,N}_{y_{1:t}}[\hat{Z}_t \mid \mathcal{F}_{t-s}] \leq A_{t-s} + B_{t-s},$$

where,

$$A_{t-s} \triangleq (\zeta N)^{-s+1}\sum_n w^n_{t-s} g^n_{t-s}\left(\sum_{\ell=1}^{s}\sum_{\boldsymbol{\tau}\in\mathcal{T}_{t,\ell,s}} (\zeta N)^{s-1-\ell} G^y_{t-s+1,\tau_1} C^y_\ell(\boldsymbol{\tau})\right),$$

$$B_{t-s} \triangleq (\zeta N)^{-s+1}\sum_n w^n_{t-s}\left(\sum_{\ell=1}^{s}\sum_{\boldsymbol{\tau}\in\mathcal{T}_{t,\ell,s}} (\zeta N)^{s-\ell} G^n_{t-s,\tau_1} C^y_\ell(\boldsymbol{\tau})\right).$$

Proof. For (1),

$$\begin{aligned}
&\mathbb{E}^{1,N}_{y_{1:t}}\left[\sum_n \phi^n_s(X^n_s) \,\middle|\, \mathcal{F}_{s-1}\right] \\
&= \sum_{f^1_s}\sum_{a_s^{-f^1_{s-1}}} \alpha^{f^1_s f^1_{s-1}}_{s-1} \prod_{k\neq f^1_s} r_k\big(a^k_{s-1} | w^{1:N}_{s-1}, \boldsymbol{x}^{1:N}_{1:s-1}\big) \\
&\quad \times \mathbb{E}^{1,N}_{y_{1:t}}\left[\sum_n \phi^n_s(X^n_s) \,\middle|\, \mathcal{F}_{s-1}, A^{1:N}_{s-1} = a^{1:N}_{s-1}, F^1_s = f^1_s\right] \\
&= \sum_{f^1_s}\sum_{a_s^{-f^1_{s-1}}} \alpha^{f^1_s f^1_{s-1}}_{s-1} \prod_{k\neq f^1_s} \frac{\alpha^{ka^k_{s-1}}_{s-1} w^{a^k_{s-1}}_{s-1} g_{s-1}(x^{a^k_{s-1}}_{s-1})}{w^k_s}\Bigg(\phi^{f^1_s}_s(y_s) \\
&\quad + \sum_{n\neq f^1_s} \mathbb{E}\big[\phi^n_s(\xi_s) \mid \xi_{s-1} = x^{a^n_{s-1}}_{s-1}\big]\Bigg) \\
&= \sum_{f^1_s} \alpha^{f^1_s f^1_{s-1}}_{s-1} \phi^{f^1_s}_s(y_s)
\end{aligned}$$

$$+\sum_{f_s^1}\sum_{n\neq f_s^1}\sum_k \alpha_{s-1}^{f_s^1 f_{s-1}^1}\frac{\alpha_{s-1}^{nk}w_{s-1}^k g_{s-1}(x_{s-1}^k)}{w_s^n}\mathbb{E}\big[\phi_s^n(\xi_s)\mid \xi_{s-1}=x_{s-1}^k\big]$$

$$=\sum_{f_s^1}\alpha_{s-1}^{f_s^1 f_{s-1}^1}\phi_s^{f_s^1}(y_s)+\sum_{f_s^1}\sum_{n\neq f_s^1}\sum_k \alpha_{s-1}^{f_s^1 f_{s-1}^1}\frac{\alpha_{s-1}^{nk}w_{s-1}^k}{w_s^n}Q_{s-1,s}^k\big(\phi_s^n\big).$$

For (2), choosing $\phi_s^n(x)=w_s^n G_{s,s+\tau}(x)$, we have

$$\begin{aligned}
&\mathbb{E}_{y_{1:t}}^{1,N}\Big[\sum_n W_s^n G_{s,s+\tau}^n \,\Big|\, \mathcal{F}_{s-1}\Big]\\
&=\sum_{f_s^1}\alpha_{s-1}^{f_s^1 f_{s-1}^1}w_s^{f_s^1}G_{s,s+\tau}^y+\sum_{f_s^1}\sum_{n\neq f_s^1}\sum_k \alpha_{s-1}^{f_s^1 f_{s-1}^1}\frac{\alpha_{s-1}^{nk}w_{s-1}^k}{w_s^n}Q_{s-1,s}^k\big(w_s^n G_{s,s+\tau}\big)\\
&=G_{s,s+\tau}^y\sum_{f_s^1}\alpha_{s-1}^{f_s^1 f_{s-1}^1}w_s^{f_s^1}+\sum_{f_s^1}\sum_{n\neq f_s^1}\sum_k \alpha_{s-1}^{f_s^1 f_{s-1}^1}\alpha_{s-1}^{nk}w_{s-1}^k G_{s-1,s+\tau}^k\\
&\le G_{s,s+\tau}^y\sum_{f_s^1}\alpha_{s-1}^{f_s^1 f_{s-1}^1}\frac{\|w_s^{1:N}\|_1}{\zeta N}+\sum_{f_s^1}\sum_n\sum_k \alpha_{s-1}^{f_s^1 f_{s-1}^1}\alpha_{s-1}^{nk}w_{s-1}^k G_{s-1,s+\tau}^k\\
&=\frac{G_{s,s+\tau}^y}{\zeta N}\sum_n w_s^n+\sum_n\sum_k \alpha_{s-1}^{nk}w_{s-1}^k G_{s-1,s+\tau}^k\\
&=\frac{G_{s,s+\tau}^y}{\zeta N}\sum_n\sum_k \alpha_{s-1}^{nk}w_{s-1}^k g_{s-1}^k+\sum_k w_{s-1}^k G_{s-1,s+\tau}^k\\
&=\frac{1}{\zeta N}\sum_k w_{s-1}^k g_{s-1}^k G_{s,s+\tau}^y+\sum_k w_{s-1}^k G_{s-1,s+\tau}^k,
\end{aligned}$$

where the inequality follows from Assumption 1.A, and we have repeatedly used Assumption 4.D.

To show (3), we start by using (2) with $s=t$ and $\tau=1$:

$$\begin{aligned}
\mathbb{E}_{y_{1:t}}^{1,N}\Big[\sum_n W_t^n g_t^n \,\Big|\, \mathcal{F}_{t-1}\Big]&=\frac{1}{\zeta N}\sum_k w_{t-1}^k g_{t-1}^k g_t^y+\frac{\zeta N}{\zeta N}\sum_m w_{t-1}^m G_{t-1,t+1}^m\\
&=A_{t-1}+B_{t-1}.
\end{aligned}$$

Hence, (3) holds for $s=1$. We now assume that the bound holds for some $s\in\{1,\dots,t-2\}$ and establish that it also holds for $s+1$. Using the inductive hypothesis,

$$\begin{aligned}
N\,\mathbb{E}_{y_{1:t}}^{1,N}[\hat{Z}_t\mid\mathcal{F}_{t-s-1}]&=\mathbb{E}_{y_{1:t}}^{1,N}\big[N\,\mathbb{E}_{y_{1:t}}^{1,N}[\hat{Z}_t\mid\mathcal{F}_{t-s}]\mid\mathcal{F}_{t-s-1}\big]\\
&\le\mathbb{E}_{y_{1:t}}^{1,N}[A_{t-s}+B_{t-s}\mid\mathcal{F}_{t-s-1}].
\end{aligned}$$

Using (2), we have

$$
\begin{aligned}
A &\triangleq \mathbb{E}^{1,N}_{y_{1:t}}[A_{t-s} \mid \mathcal{F}_{t-s-1}] \\
&= (\zeta N)^{-s+1}\left(\sum_{\ell=1}^{s}\sum_{\boldsymbol{\tau}\in\mathcal{T}_{t,\ell,s}} (\zeta N)^{s-1-\ell} G^{y}_{t-s+1,\tau_1} C^{y}_{\ell}(\boldsymbol{\tau})\right)\mathbb{E}^{1,N}_{y_{1:t}}\left[\sum_n W^n_{t-s} g^n_{t-s}\right] \\
&\le (\zeta N)^{-s}\left(\sum_{\ell=1}^{s}\sum_{\boldsymbol{\tau}\in\mathcal{T}_{t,\ell,s}} (\zeta N)^{s-1-\ell} G^{y}_{t-s+1,\tau_1} C^{y}_{\ell}(\boldsymbol{\tau})\right) \\
&\quad \times \left(\sum_n w^n_{t-s-1} g^n_{t-s-1} g^y_{t-s} + \zeta N \sum_n w^n_{t-s-1} G^n_{t-s-1,t-s+1}\right)
\end{aligned}
$$

and

$$
\begin{aligned}
B &\triangleq \mathbb{E}^{1,N}_{y_{1:t}}[B_{t-s} \mid \mathcal{F}_{t-s-1}] \\
&= (\zeta N)^{-s+1}\left(\sum_{\ell=1}^{s}\sum_{\boldsymbol{\tau}\in\mathcal{T}_{t,\ell,s}} (\zeta N)^{s-\ell} \mathbb{E}^{1,N}_{y_{1:t}}\left[\sum_n W^n_{t-s} G^n_{t-s,\tau_1} \,\middle|\, \mathcal{F}_{t-s-1}\right] C^{y}_{\ell}(\boldsymbol{\tau})\right) \\
&\le (\zeta N)^{-s}\left(\sum_{\ell=1}^{s}\sum_{\boldsymbol{\tau}\in\mathcal{T}_{t,\ell,s}} (\zeta N)^{s-\ell}\left(\sum_n w^n_{t-s-1} g^n_{t-s-1} G^y_{t-s,\tau_1}\right.\right. \\
&\quad \left.\left. + \zeta N \sum_n w^n_{t-s-1} G^n_{t-s-1,\tau_1}\right) C^{y}_{\ell}(\boldsymbol{\tau})\right).
\end{aligned}
$$

Hence,

$$
\begin{aligned}
A + B \le{}& (\zeta N)^{-s} \sum_n w^n_{t-s-1} g^n_{t-s-1}\left(\sum_{\ell=1}^{s}\sum_{\boldsymbol{\tau}\in\mathcal{T}_{t,\ell,s}} (\zeta N)^{s-1-\ell} G^{y}_{t-s,t-s+1} G^{y}_{t-s+1,\tau_1} C^{y}_{\ell}(\boldsymbol{\tau})\right) \\
&+ (\zeta N)^{-s} \sum_n w^n_{t-s-1} g^n_{t-s-1}\left(\sum_{\ell=1}^{s}\sum_{\boldsymbol{\tau}\in\mathcal{T}_{t,\ell,s}} (\zeta N)^{s-\ell} G^{y}_{t-s,\tau_1} C^{y}_{\ell}(\boldsymbol{\tau})\right) \\
&+ (\zeta N)^{-s} \sum_n w^n_{t-s-1}\left(\sum_{\ell=1}^{s}\sum_{\boldsymbol{\tau}\in\mathcal{T}_{t,\ell,s}} (\zeta N)^{s-\ell} G^{n}_{t-s-1,t-s+1} G^{y}_{t-s+1,\tau_1} C^{y}_{\ell}(\boldsymbol{\tau})\right) \\
&+ (\zeta N)^{-s} \sum_n w^n_{t-s-1}\left(\sum_{\ell=1}^{s}\sum_{\boldsymbol{\tau}\in\mathcal{T}_{t,\ell,s}} (\zeta N)^{s-\ell+1} G^{n}_{t-s-1,\tau_1} C^{y}_{\ell}(\boldsymbol{\tau})\right).
\end{aligned}
$$

Summing the parenthesized double sums of the first two terms yields

$$\begin{aligned}
&\sum_{\ell=1}^{s}\sum_{\boldsymbol{\tau}\in\mathcal{T}_{t,\ell,s}}(\zeta N)^{s-1-\ell}G^{y}_{t-s,t-s+1}G^{y}_{t-s+1,\tau_1}C^{y}_{\ell}(\boldsymbol{\tau})+\sum_{\ell=1}^{s}\sum_{\boldsymbol{\tau}\in\mathcal{T}_{t,\ell,s}}(\zeta N)^{s-\ell}G^{y}_{t-s,\tau_1}C^{y}_{\ell}(\boldsymbol{\tau})\\
&\quad=\sum_{\ell=1}^{s+1}\sum_{\substack{\boldsymbol{\tau}\in\mathcal{T}_{t,\ell,s+1}\\ \tau_1=t-s+1}}(\zeta N)^{s-\ell}G^{y}_{t-s,\tau_1}C^{y}_{\ell}(\boldsymbol{\tau})+\sum_{\ell=1}^{s+1}\sum_{\substack{\boldsymbol{\tau}\in\mathcal{T}_{t,\ell,s+1}\\ \tau_1>t-s+1}}(\zeta N)^{s-\ell}G^{y}_{t-s,\tau_1}C^{y}_{\ell}(\boldsymbol{\tau})\\
&\quad=\sum_{\ell=1}^{s+1}\sum_{\boldsymbol{\tau}\in\mathcal{T}_{t,\ell,s+1}}(\zeta N)^{s-\ell}G^{y}_{t-s,\tau_1}C^{y}_{\ell}(\boldsymbol{\tau}),
\end{aligned}$$

so the first two terms are equal to $A_{t-(s+1)}$. Summing the parenthesized double sums of the last two terms yields

$$\begin{aligned}
&\sum_{\ell=1}^{s}\sum_{\boldsymbol{\tau}\in\mathcal{T}_{t,\ell,s}}(\zeta N)^{s-\ell}G^{n}_{t-s-1,t-s+1}G^{y}_{t-s+1,\tau_1}C^{y}_{\ell}(\boldsymbol{\tau})\\
&\qquad+\sum_{\ell=1}^{s}\sum_{\boldsymbol{\tau}\in\mathcal{T}_{t,\ell,s}}(\zeta N)^{s-\ell+1}G^{n}_{t-s-1,\tau_1}C^{y}_{\ell}(\boldsymbol{\tau})\\
&\quad=\sum_{\ell=1}^{s+1}\sum_{\substack{\boldsymbol{\tau}\in\mathcal{T}_{t,\ell,s+1}\\ \tau_1=t-s+1}}(\zeta N)^{s-\ell+1}G^{n}_{t-s-1,\tau_1}C^{y}_{\ell}(\boldsymbol{\tau})+\sum_{\ell=1}^{s+1}\sum_{\substack{\boldsymbol{\tau}\in\mathcal{T}_{t,\ell,s+1}\\ \tau_1>t-s+1}}(\zeta N)^{s-\ell+1}G^{n}_{t-s-1,\tau_1}C^{y}_{\ell}(\boldsymbol{\tau})\\
&\quad=\sum_{\ell=1}^{s+1}\sum_{\boldsymbol{\tau}\in\mathcal{T}_{t,\ell,s+1}}(\zeta N)^{s-\ell+1}G^{n}_{t-s-1,\tau_1}C^{y}_{\ell}(\boldsymbol{\tau}),
\end{aligned}$$

so the last two terms are equal to $B_{t-(s+1)}$. □

Using part (3) of Lemma A.2 with $s=t-1$, we have

$$N\,\mathbb{E}^{1,N}_{y_{1:t}}[\hat{Z}_t]\le\mathbb{E}^{1,N}_{y_{1:t}}[A_1+B_1].$$

Therefore,

$$\begin{aligned}
\mathbb{E}^{1,N}_{y_{1:t}}[A_1]&=(\zeta N)^{-t+2}\mathbb{E}^{1,N}_{y_{1:t}}\Big[\sum_{n}g^{n}_{1}\Big]\Big(\sum_{\ell=1}^{t-1}\sum_{\boldsymbol{\tau}\in\mathcal{T}_{t,\ell,t-1}}(\zeta N)^{t-2-\ell}G^{y}_{2,\tau_1}C^{y}_{\ell}(\boldsymbol{\tau})\Big)\\
&=(\zeta N)^{-t+2}\big(G^{y}_{1,2}+(N-1)G_{0,2}\big)\Big(\sum_{\ell=1}^{t-1}\sum_{\boldsymbol{\tau}\in\mathcal{T}'_{\ell,t-1}}(\zeta N)^{t-2-\ell}G^{y}_{2,\tau_1}C^{y}_{\ell}(\boldsymbol{\tau})\Big)
\end{aligned}$$

and

$$\mathbb{E}^{1,N}_{y_{1:t}}[B_1] = (\zeta N)^{-t+2}\left(\sum_{\ell=1}^{t-1}\sum_{\boldsymbol{\tau}\in\mathcal{T}_{t,\ell,t-1}}(\zeta N)^{t-1-\ell}\mathbb{E}^{1,N}_{y_{1:t}}\left[\sum_n G^n_{1,\tau_1}\right]C^y_\ell(\boldsymbol{\tau})\right)$$

$$= (\zeta N)^{-t+2}\left(\sum_{\ell=1}^{t-1}\sum_{\boldsymbol{\tau}\in\mathcal{T}_{t,\ell,t-1}}(\zeta N)^{t-1-\ell}\big(G^y_{1,\tau_1} + (N-1)G_{0,\tau_1}\big)C^y_\ell(\boldsymbol{\tau})\right).$$

Hence, using arguments analogous to those from the proof of Lemma A.2 and the fact that $G_{0,1} = 1$ yields

$$\begin{aligned}
\mathbb{E}^{1,N}_{y_{1:t}}[A_1 + B_1] &= (\zeta N)^{-t+2}\sum_{\ell=1}^{t}\sum_{\boldsymbol{\tau}\in\mathcal{T}_{t,\ell,t}}(\zeta N)^{t-1-\ell}G^y_{1,\tau_1}C^y_\ell(\boldsymbol{\tau}) \\
&\quad + \frac{N-1}{\zeta N}(\zeta N)^{-t+2}\sum_{\ell=1}^{t}\sum_{\boldsymbol{\tau}\in\mathcal{T}_{t,\ell,t}}(\zeta N)^{t-\ell}G_{0,\tau_1}C^y_\ell(\boldsymbol{\tau}) \\
&= (\zeta N)^{-t+2}\sum_{\ell=1}^{t+1}\sum_{\substack{\boldsymbol{\tau}\in\mathcal{T}_{t,\ell,t+1}\\ \tau_1=1}}(\zeta N)^{t-\ell}G_{0,\tau_1}C^y_\ell(\boldsymbol{\tau}) \\
&\quad + \frac{N-1}{\zeta N}(\zeta N)^{-t+2}\sum_{\ell=1}^{t+1}\sum_{\substack{\boldsymbol{\tau}\in\mathcal{T}_{t,\ell,t+1}\\ \tau_1>1}}(\zeta N)^{t-\ell}G_{0,\tau_1}C^y_\ell(\boldsymbol{\tau}) \\
&= (\zeta N)^{-t+2}\sum_{\ell=1}^{t+1}\sum_{\boldsymbol{\tau}\in\mathcal{T}_{t,\ell,t+1}}(\zeta N)^{t-\ell}\left(\frac{N-1}{\zeta N}\right)^{\mathbb{1}(\tau_1>1)}G_{0,\tau_1}C^y_\ell(\boldsymbol{\tau}).
\end{aligned}$$

A.5. Divergence of importance samplers

The key quantity in this section is the variance of the potentials:

$$\mathcal{V}_t \triangleq \mathrm{Var}\big[Z_t^{-1}g_{1:t}(\xi_{1:t})\big] = \mathbb{E}\big[\big(Z_t^{-1}g_{1:t}(\xi_{1:t}) - 1\big)^2\big].$$

Theorem A.3. *If $\alpha_s = \boldsymbol{I}_N$ for all $s \in [t-1]$, then*

$$d_{\mathrm{KL}}\big(\pi_{1:t}||P^{0,N}\big) \le \log\left(1 + \frac{\mathcal{V}_t}{N}\right),$$

$$d_{\chi^2}\big(\pi_{1:t}||P^{0,N}\big) \le \frac{\mathcal{V}_t}{N}.$$

Proof. By Theorem 4.2 and Jensen's inequality

$$\begin{aligned}\frac{\mathrm{d}P^{0,N}}{\mathrm{d}\pi_{1:t}}(x_{1:t}) &= \mathbb{E}^{1,N}_{x_{1:t}}\left[\frac{Z_t}{\hat{Z}_t}\right] \\ &\geq \frac{N}{\mathbb{E}^{1,N}_{x_{1:t}}[\sum_{k=1}^{N} Z_t^{-1} g_{1:t}(\tilde{X}^k_{1:t})]} \\ &= \frac{N}{N-1+Z_t^{-1}g_{1:t}(x_{1:t})}.\end{aligned}$$

By definition of the χ^2 divergence,

$$\begin{aligned}d_{\chi^2}\big(\pi_{1:t}||\bar{\pi}^{\mathrm{S},N}_{1,t}\big) &= \pi_{1:t}\left(\frac{\mathrm{d}\pi_{1:t}}{\mathrm{d}P^{0,N}}\right) - 1 \\ &= M_{1:t}\left(\frac{\mathrm{d}\pi_{1:t}}{\mathrm{d}P^{0,N}}\frac{\mathrm{d}\pi_{1:t}}{\mathrm{d}M_{1:t}}\right) - 1 \\ &\leq M_{1:t}\left(\frac{N-1+Z_t^{-1}g_{1:t}}{N}Z_t^{-1}g_{1:t}\right) - 1 \\ &= \frac{M_{1:t}(Z_t^{-1}g_{1:t})^2 - 1}{N} \\ &= \frac{\mathrm{Var}[Z_t^{-1}g_{1:t}(\xi_{1:t})]}{N}.\end{aligned}$$

The bound of the KL divergence follows from the elementary inequality $d_{\mathrm{KL}}(\mu||\nu) \leq \log(1 + d_{\chi^2}(\mu||\nu))$. □

A.6. Invariant distribution of the i-cαSMC kernel

Lemma A.4. *$P^{1,N}_{\theta,y}(\mathrm{d}z)$ is reversible with respect to $\pi_\theta(\mathrm{d}z)$.*

Proof. We mostly suppress dependence on θ since θ is fixed. We will show that the cαSMC kernel is Gibbs sampler for the artificial joint density given in equation (4.3), which we recall is

$$\tilde{\pi}_{1:t}\big(\boldsymbol{x}^{1:N}_{1:t}, \boldsymbol{a}^{1:N}_{1:t-1}, f^1_{1:t}\big) \triangleq \pi_{1:t}\big(x^{f^1_t}_{1:t}\big)\tilde{\psi}\big(\boldsymbol{x}^{1:N}_{1:t}, \boldsymbol{a}^{1:N}_{1:t-1}, f^1_{1:t}\big).$$

In particular, letting $\boldsymbol{\omega}_{1:t} \triangleq (\boldsymbol{x}^{1:N}_{1:t}, \boldsymbol{a}^{1:N}_{1:t-1}, f^1_{1:t})$, by definition,

$$\mathbb{P}^{i,N}_{\boldsymbol{y}^{1:i}_{1:t}}[\mathrm{d}\boldsymbol{\omega}_{1:t}] = \tilde{\pi}_{1:t}\big(\mathrm{d}\boldsymbol{\omega}_{1:t} \mid \tilde{x}^{f^1_t}_{1:t} = y_{1:t}\big).$$

Furthermore,

$$\begin{aligned}
\tilde{\pi}_{1:t}\big(\tilde{x}_{1:t}^{f_t^1} = \tilde{x}_{1:t}^{a_t^*} \mid \boldsymbol{\omega}_{1:t}\big) &= \tilde{\pi}_{1:t}\big(f_t^1 = a_t^* \mid \boldsymbol{\omega}_{1:t}\big) \\
&\propto \frac{M_1(x_1^{a_1^*}) g_1(x_1^{a_1^*}) \prod_{s=2}^t I_s M_s(x_{s-1}^{a_{s-1}^*}, x_s^{a_{s-1}^*} s) g_s(x_s^{a_s^*}) \alpha_{s-1}^{a_s^* a_{s-1}^*}}{M_1(x_1^{a_1^*}) \prod_{s=2}^t r_{a_s^*}(a_{s-1}^* | w_{s-1}^{1:N}, \boldsymbol{x}_{1:s-1}^{1:N}) M_s(x_{s-1}^{a_{s-1}^*}, x_s^{a_{s-1}^*})} \\
&= w_t^{a_t^*} g_t\big(x_t^{a_t^*}\big) \\
&\propto \mathbb{P}_{\boldsymbol{y}_{1:t}^{1:i}}^{i,N}\big[a_t^* \mid \boldsymbol{x}_{1:t}^{1:N}, \boldsymbol{a}_{1:t-1}^{1:N}, f_{1:t}^1\big].
\end{aligned}$$

Reversibility now follows easily:

$$\begin{aligned}
P_y^{1,N}(\mathrm{d}z)\pi(\mathrm{d}y) &= \int \tilde{\pi}_{1:t}(\mathrm{d}z \mid \boldsymbol{\omega}_{1:t}) \tilde{\pi}_{1:t}(\mathrm{d}\boldsymbol{\omega}_{1:t} \mid y)\pi(\mathrm{d}y) \\
&= \int \tilde{\pi}_{1:t}(\mathrm{d}z \mid \boldsymbol{\omega}_{1:t}) \tilde{\pi}_{1:t}(\mathrm{d}y \mid \boldsymbol{\omega}_{1:t}) \tilde{\pi}_{1:t}(\mathrm{d}\boldsymbol{\omega}_{1:t}) \\
&= \int \tilde{\pi}_{1:t}(\mathrm{d}y \mid \boldsymbol{\omega}_{1:t}) \tilde{\pi}_{1:t}(\mathrm{d}\boldsymbol{\omega}_{1:t} \mid z)\pi(\mathrm{d}z) \\
&= P_z^{1,N}(\mathrm{d}y)\pi(\mathrm{d}z). \qquad \square
\end{aligned}$$

Acknowledgements

The authors would like to thank Arnaud Doucet for critical feedback and numerous helpful suggestions, Cameron Freer for feedback on early versions of this work, Josh Tenenbaum for discussions that helped to inspire this work, Vikash Mansinghka for suggesting we investigate the expected value of SMC estimators, and an anonymous referee for suggestions on improving the presentation of the results. JHH was supported by the U.S. Government under FA9550-11-C-0028 and awarded by the DoD, Air Force Office of Scientific Research, National Defense Science and Engineering Graduate (NDSEG) Fellowship, 32 CFR 168a. This research was carried out in part while DMR held a Research Fellowship at Emmanuel College, Cambridge, with funding also from a Newton International Fellowship through the Royal Society, an NSERC Discovery Grant, Connaught Award, and U.S. Air Force Office of Scientific Research grant #FA9550-15-1-0074.

References

[1] Andrieu, C., Doucet, A. and Holenstein, R. (2010). Particle Markov chain Monte Carlo methods. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **72** 269–342. MR2758115

[2] Andrieu, C., Lee, A. and Vihola, M. (2018). Uniform ergodicity of the iterated conditional SMC and geometric ergodicity of particle Gibbs samplers. *Bernoulli* **24** 842–872. MR3706778

[3] Andrieu, C. and Roberts, G.O. (2009). The pseudo-marginal approach for efficient Monte Carlo computations. *Ann. Statist.* **37** 697–725. MR2502648
[4] Andrieu, C. and Vihola, M. (2015). Convergence properties of pseudo-marginal Markov chain Monte Carlo algorithms. *Ann. Appl. Probab.* **25** 1030–1077. MR3313762
[5] Chopin, N. and Singh, S.S. (2015). On particle Gibbs sampling. *Bernoulli* **21** 1855–1883. MR3352064
[6] Cornebise, J., Moulines, É. and Olsson, J. (2008). Adaptive methods for sequential importance sampling with application to state space models. *Stat. Comput.* **18** 461–480. MR2461889
[7] Del Moral, P. (2004). *Feynman–Kac Formulae: Genealogical and Interacting Particle Systems with Applications. Probability and Its Applications*. New York: Springer. MR2044973
[8] Del Moral, P., Doucet, A. and Jasra, A. (2006). Sequential Monte Carlo samplers. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **68** 411–436. MR2278333
[9] Doucet, A., de Freitas, N. and Gordon, N. (2001). *Sequential Monte Carlo in Practice*. New York: Springer.
[10] Doucet, A., Godsill, S.J. and Andrieu, C. (2000). On sequential Monte Carlo sampling methods for Bayesian filtering. *Stat. Comput.* **10** 197–208.
[11] Doucet, A. and Johansen, A.M. (2010). A tutorial on particle filtering and smoothing: Fifteen years later. In *Handbook of Nonlinear Filtering* (D. Crisan and B. Rozovsky, eds.). Cambridge: Cambridge Univ. Press.
[12] Gordon, N.J., Salmond, D.J. and Smith, A.F. (1993). Novel approach to nonlinear/non-Gaussian Bayesian state estimation. *IEE Proc., F, Radar Signal Process.* **140** 107–113.
[13] Holenstein, R. (2009). Particle Markov Chain Monte Carlo. Ph.D. thesis, Univ. British Columbia, Vancouver.
[14] Kantas, N., Doucet, A., Singh, S.S. and Maciejowski, J.M. (2009). An overview of sequential Monte Carlo methods for parameter estimation in general state-space models. In 15*th IFAC Symposium on System Identification* 774–785.
[15] Künsch, H.R. (2013). Particle filters. *Bernoulli* **19** 1391–1403. MR3102556
[16] Lee, A. and Latuszynski, K. (2014). Variance bounding and geometric ergodicity of Markov chain Monte Carlo kernels for approximate Bayesian computation. *Biometrika* **101** 655–671.
[17] Lindsten, F., Douc, R. and Moulines, E. (2015). Uniform ergodicity of the particle Gibbs sampler. *Scand. J. Stat.* **42** 775–797. MR3391692
[18] Whiteley, N., Lee, A. and Heine, K. (2016). On the role of interaction in sequential Monte Carlo algorithms. *Bernoulli* **22** 494–529.

Received March 2015 and revised April 2017