The Annals of Statistics
2020, Vol. 48, No. 5, 3066–3089
https://doi.org/10.1214/19-AOS1922

COUPLED CONDITIONAL BACKWARD SAMPLING PARTICLE FILTER

BY ANTHONY LEE[1], SUMEETPAL S. SINGH[2] AND MATTI VIHOLA[3]

[1]*School of Mathematics, University of Bristol, anthony.lee@bristol.ac.uk*
[2]*Department of Engineering, University of Cambridge*
[3]*Department of Mathematics and Statistics, University of Jyväskylä*

The conditional particle filter (CPF) is a promising algorithm for general hidden Markov model smoothing. Empirical evidence suggests that the variant of CPF with backward sampling (CBPF) performs well even with long time series. Previous theoretical results have not been able to demonstrate the improvement brought by backward sampling, whereas we provide rates showing that CBPF can remain effective with a fixed number of particles independent of the time horizon. Our result is based on analysis of a new coupling of two CBPFs, the coupled conditional backward sampling particle filter (CCBPF). We show that CCBPF has good stability properties in the sense that with fixed number of particles, the coupling time in terms of iterations increases only linearly with respect to the time horizon under a general (strong mixing) condition. The CCBPF is useful not only as a theoretical tool, but also as a practical method that allows for unbiased estimation of smoothing expectations, following the recent developments by Jacob, Lindsten and Schön (2020). Unbiased estimation has many advantages, such as enabling the construction of asymptotically exact confidence intervals and straightforward parallelisation.

1. Introduction. The conditional particle filter (CPF) introduced by Andrieu, Doucet and Holenstein (2010) is a Markov Chain Monte Carlo method that produces asymptotically unbiased samples from the posterior distribution of the states of a hidden Markov model. The CPF can be made significantly more efficient by the inclusion of backward sampling (Whiteley (2010)), or equivalently ancestor sampling (Lindsten, Jordan and Schön (2014)), steps: we refer to the resulting algorithm as the conditional backward sampling particle filter (CBPF). While there are many empirical studies reporting on the effectiveness of the CBPF for Bayesian inference and on its superiority over the CPF (see, e.g., Fearnhead and Künsch (2018), Section 7.2.2), quantitative theoretical guarantees for the CBPF are still missing. In contrast, the theoretical properties of the CPF are much better understood (Chopin and Singh (2015), Lindsten, Douc and Moulines (2015), Andrieu, Lee and Vihola (2018)).

Chopin and Singh (2015) introduced a coupling construction, called the coupled CPF (CCPF), to prove the uniform ergodicity of the CPF. Recently, Jacob, Lindsten and Schön (2020) identified the potential use of the CCPF to produce unbiased estimators by exploiting a de-biasing technique due to Glynn and Rhee (2014) (see also Jacob, O'Leary and Atchadé (2020)). This is an important algorithmic advancement to particle filtering methodology since unbiased estimation is useful for estimating confidence intervals, allows straightforward parallelisation, and when used within a stochastic approximation context, such as the stochastic approximation expectation maximisation (SAEM) scheme (Delyon, Lavielle and Moulines (1999)), unbiased estimators ensure martingale noise, which has good supporting theory.

The methodological contribution of this paper is a relatively simple yet important algorithmic modification to the CCPF by extending the CCPF to include backward sampling steps

Received December 2018; revised August 2019.
MSC2020 subject classifications. Primary 65C40; secondary 65C05, 65C35, 65C60.
Key words and phrases. Backward sampling, convergence rate, coupling, conditional particle filter, unbiased.

through an index-coupled version of Whiteley's (2010) backward sampling CPF. This approach, which we call the coupled conditional backward sampling particle filter (CCBPF), gives theoretical insight to the behaviour of the CBPF. It can also be used practically to facilitate unbiased estimation. The CCBPF appears to be far more stable than the CCPF (Chopin and Singh (2015)) (and the variant of Jacob, Lindsten and Schön (2020) that uses coupled ancestor sampling within the CCPF). Under a general (but strong) mixing condition, we prove (Theorem 6) that the coupling time of CCBPF grows at most linearly with length of the data record when a fixed number of particles are used, provided this fixed number is sufficiently large. From a computational perspective, this makes the CCBPF algorithm more appropriate than alternative coupled CPFs when the length of the data record is very large, as one only needs to have memory linear in the length of the data record.

As an important corollary of our analysis of the CCBPF, we obtain new convergence guarantees for the CBPF (Theorem 4) that verifies its superiority over the CPF. More specifically, this result differs from existing time-uniform guarantees for the CPF (Andrieu, Lee and Vihola (2018), Lindsten, Douc and Moulines (2015)) which require (super)linear growth of the number of particles N with the length of the data record T. Our result confirms the long held view, stemming from numerous empirical studies, that the CBPF remains an effective sampler with a fixed N even as T increases. An important consequence of a fixed N is that the space complexity (the amount of memory required) of the algorithm is linear, as opposed to quadratic, in T, making it feasible to run on long data records without exhausting the memory available on a computer. We remark that a variant of the the CPF that is stable with a fixed N is the blocked version of the CPF introduced by Singh, Lindsten and Moulines (2017), but that algorithm and its analysis are substantially different.

We also complement the empirical findings of Jacob, Lindsten and Schön (2020) by showing quantitative bounds on the 'one-shot' coupling probability of CCPF, that is, probability of coupling after a single iteration of the algorithm. These results are noteworthy as CCPF is applicable in some scenarios where CCBPF is not, for instance when the transition density is intractable. With the minimal assumption of bounded potentials, we prove (Theorem 8) that the coupling probability of CCPF is at least $1 - O(N^{-1})$, similar to what is shown for the CPF (Andrieu, Lee and Vihola (2018), Lindsten, Douc and Moulines (2015)). However, the constants involved grow very rapidly with T. Under strong mixing conditions, we are able to give a more precise rate of convergence as T increases (Theorem 9), which still requires N to increase exponentially with T (see Andrieu, Lee and Vihola (2018), Lindsten, Douc and Moulines (2015) for the CPF's rate of convergence under similar strong mixing assumptions). Our rates for the CCPF may be very conservative as empirical evidence (Jacob, Lindsten and Schön (2020)) suggests increasing N linearly with T may be sufficient for some models, although there is also evidence that N should grow superlinearly with T for other models.

2. Notation and preliminaries. Throughout the paper, we assume a general state space X, which is typically $\mathbb{R}^d$ equipped with the Lebesgue measure. However, our results hold for any measure space X equipped with a σ-finite dominating measure, which is denoted as '$\mathrm{d}x$'. Product spaces are equipped with the related product measures. We use the notation $a{:}b = a, \dots, b$ for any integers $a \le b$, and use similar notation in indexing $x_{a:b} = (x_a, \dots, x_b)$ and $x^{(a:b)} = (x^{(a)}, \dots, x^{(b)})$. We also use combined indexing, such that for instance $x_{1:T}^{(i_{1:T})} = (x_1^{(i_1)}, \dots, x_T^{(i_T)})$. We adopt the usual conventions concerning empty products and sums, namely $\prod_a^b(\cdot) = 1$ and $\sum_a^b(\cdot) = 0$ when $a > b$. We denote $x \wedge y := \min\{x, y\}$, $x \vee y := \max\{x, y\}$ and $(x)_+ := x \vee 0$.

We use standard notation for the k-step transition probability of a Markov kernel P by $P^k(x, A) := \int P(x, \mathrm{d}y)P^{k-1}(y, A)$ and $P^0(x, A) := \mathbb{I}\{x \in A\}$. If ν is a probability measure and f is a real-valued function, then $(\nu P)(A) := \int \nu(\mathrm{d}x)P(x, A)$, $(Pf)(x) =$

$\int P(x, \mathrm{d}y) f(y)$ and $\nu(f) := \int \nu(\mathrm{d}x) f(x)$, whenever well defined. The total variation metric between two probability measures μ, ν is defined as $\|\mu - \nu\|_{\mathrm{tv}} := \sup_{|f|\leq 1} |\mu(f) - \nu(f)|$, and $\|f\|_\infty := \sup_x |f(x)|$. If two random variables X and Y share a common law, we write $X \stackrel{d}{=} Y$. We denote the categorical distribution by $\mathrm{Categ}(\omega^{(1:N)})$ for unnormalised probabilities $\omega^{(i)}$, that is, $I \sim \mathrm{Categ}(\omega^{(1:N)})$ if $\mathbb{P}(I = i) = \omega^{(i)} / \sum_{j=1}^N \omega^{(j)}$.

We are interested in computing expectations of smoothing functionals, with respect to the probability density $\pi_T(x_{1:T}) := \gamma_T(x_{1:T})/c_T$ on a space X^T with the following unnormalised density (cf. Del Moral (2004)):

$$\gamma_T(x_{1:T}) := M_1(x_1) G_1(x_1) \prod_{t=2}^{T} M_t(x_{t-1}, x_t) G_t(x_{t-1}, x_t), \tag{1}$$

where M_1 is a probability density, M_t are Markov transition densities, $G_1 : \mathsf{X} \to [0, \infty)$ and $G_t : \mathsf{X}^2 \to [0, \infty)$ for $t \in \{2{:}T\}$ are 'potential functions', and $c_T := \int \gamma_T(x_{1:T})\, \mathrm{d}x_{1:T} \in (0, \infty)$ is an unknown normalising constant. The probability density in (1) also encompasses the posterior density of a hidden Markov model (HMM) when the pair (G_t, M_t) is defined appropriately. For example, let the HMM be defined by a Markov state process having initial density $f_1(x_1)$ and transition densities $f_t(x_t \mid x_{t-1})$, and an observed process having densities $g_t(y_t \mid x_t)$. Given the sequence of observations $y_{1:T}$, the potentials can be taken to be of the form

$$G_1(x_1) = \frac{g_1(y_1 \mid x_1) f_1(x_1)}{M_1(x_1)} \quad \text{and} \quad G_t(x_{t-1}, x_t) = \frac{g_t(y_t \mid x_t) f_t(x_t \mid x_{t-1})}{M_t(x_{t-1}, x_t)},$$

in which case $\pi_T(x_{1:T})$ corresponds to the smoothing distribution of the HMM, that is, the conditional density of the latent Markov states given the observations. In the fairly common case where $M_1 = f_1$ and $M_t(x_{t-1}, x_t) = f(x_t \mid x_{t-1})$ for $t > 1$, we write $G_t(x_t) := G_t(x_{t-1}, x_t) = g_t(y_t \mid x_t)$ to emphasise that G_t is a function only of x_t.

We will consider two different conditions for the model. Assumption 1 is generally regarded as nonrestrictive in the particle filtering literature and essentially equivalent with the uniform ergodicity of CPF (Andrieu, Lee and Vihola (2018)).

ASSUMPTION 1 (Bounded potentials). There exists $G^* < \infty$ such that $G_t(\cdot) \leq G^*$ for all $t = 1{:}T$.

Assumption 2, which subsumes Assumption 1, is a much stronger assumption introduced to prove time-uniform error bounds of particle filtering estimates (Del Moral and Guionnet (2001)). It is typically verified for models where X is compact with (i) transitions that allow movement between any two regions of the space and (ii) potentials that are lower and upper bounded away from 0 and ∞. Theoretical results using Assumption 2 are often indicative of performance in models where it does not quite hold. The assumption has been replaced by weaker but fairly involved assumptions by Whiteley (2013) in the context of time-uniform error bounds, who also briefly surveys the use of Assumption 2. Another brief survey can be found in Douc et al. (2011). At present, many theoretical papers on particle filters use Assumption 2 as weaker alternatives are very difficult to manipulate theoretically and also to verify.

ASSUMPTION 2 (Strong mixing). $G_*(1) := \inf_{x\in\mathsf{X}} G_1(x) > 0$ and $G^*(1) := \sup_{x\in\mathsf{X}} G_1(x) < \infty$, and for all $t = 2{:}T$:

(i) $M_*(t) := \inf_{x,y\in\mathsf{X}} M_t(x, y) > 0$ and $M^*(t) := \sup_{x,y\in\mathsf{X}} M_t(x, y) < \infty$,
(ii) $G_*(t) := \inf_{x,y\in\mathsf{X}} G_t(x, y) > 0$ and $G^*(t) := \sup_{x,y\in\mathsf{X}} G_t(x, y) < \infty$.

Denote $\boldsymbol{\delta} := \min_{t=1:T} \frac{G_*(t)}{G^*(t)}$ and $\boldsymbol{\epsilon} := \min_{t=1:T-1} \frac{G_*(t) M_*(t) G_*(t+1)}{G^*(t) M^*(t) G^*(t+1)}$.

REMARK 3. Note that $\delta \geq \epsilon > 0$. The expression of constant ϵ may be simplified (and improved) in two special cases, as follows:

(i) If $M_t(x, y) = M_t(y)$ for all $t = 2{:}T$, then $M_*(t)/M^*(t)$ may be omitted.
(ii) If $G_t(x, y) = G_t(y)$ for $t = 2{:}T$, then $G_*(t+1)/G^*(t+1)$ may be omitted.

In particular, if both hold, then $\epsilon = \delta$. When results are stated in the asymptotic regime where $T \to \infty$, Assumption 2 should be interpreted as holding for all $T > 1$ with a uniform lower bound $\epsilon > 0$.

3. Convergence of the conditional backward sampling particle filter. Before going to the construction of the coupled conditional particle filters, we consider the conditional particle filter (CPF) (Andrieu, Doucet and Holenstein (2010)) and the conditional backward sampling particle filter (CBPF) (Whiteley (2010)) (in short CxPF where x is a place holder), given in Algorithm 1. Both CPF and CBPF define a reversible Markov transition with respect to π_T (Chopin and Singh (2015)), with any choice of the parameter $N \geq 2$ (number of particles). The important distinction is that in the CPF, $X_{1:T}^{(J_{1:T})}$ is obtained by tracing the ancestral line of $X_T^{(J_T)}$, whereas in the CBPF ancestors are selected randomly according to the Markov transition densities and potential functions. The CPF and the CBPF have the same time complexity and space complexity, $O(TN)$. The CBPF takes a constant factor more time due to the additional computations required to sample random ancestors: if the transition densities are not expensive, this factor will typically be less than 2.

We focus first on the important implication of our result for the convergence time of the CBPF, which applies also to the ancestor sampling implementation of Lindsten, Jordan and Schön (2014) as it is probabilistically equivalent to the CBPF.

THEOREM 4. *Suppose Assumption 2 (strong mixing) holds, and denote by $P_{T,N}$ the Markov transition probability of CBPF with N particles (Algorithm 1). For any $\rho > 0$, there*

Algorithm 1 CxPF($X_{1:T}^*, N$)

$X_{1:T}^{(1)} \leftarrow X_{1:T}^*$.
$X_1^{(i)} \sim M_1(\cdot)$ for $i \in \{2{:}N\}$.
$\omega_1^{(i)} \leftarrow G_1(X_1^{(i)})$ for $i \in \{1{:}N\}$.
for $t = 2{:}T$ **do**
    $I_t^{(i)} \sim \text{Categ}(\omega_{t-1}^{(1:N)})$ for $i \in \{2{:}N\}$.
    $X_t^{(i)} \sim M_t(X_{t-1}^{(I_t^{(i)})}, \cdot)$ for $i \in \{2{:}N\}$.
    $\omega_t^{(i)} \leftarrow G_t(X_{t-1}^{(I_t^{(i)})}, X_t^{(i)})$ for $i \in \{1{:}N\}$.
end for
$J_T \sim \text{Categ}(\omega_T^{(1:N)})$
for $t = (T-1){:}1$ **do**
    if CBPF **do**
        $b_t^{(i)} \leftarrow \omega_t^{(i)} M_{t+1}(X_t^{(i)}, X_{t+1}^{(J_{t+1})}) G_{t+1}(X_t^{(i)}, X_{t+1}^{(J_{t+1})})$
        $J_t \sim \text{Categ}(b_t^{(1:N)})$
    if CPF **do**
        $J_t \leftarrow I_{t+1}^{(J_{t+1})}$ where $I_t^{(1)} = 1$.
end for
output $X_{1:T}^{(J_{1:T})}$

exists $N_0 = N_0(\epsilon, \rho) < \infty$ *such that for all* $N \geq N_0$,

$$\lim_{T\to\infty} \sup_{x\in \mathsf{X}} \| P_{T,N}^{\lceil \rho T \rceil}(x, \cdot) - \pi_T \|_{\mathrm{tv}} = 0. \tag{2}$$

More precisely, for any $\alpha, \beta \in (1, \infty)$ *there exists* $N_0 = N_0(\epsilon, \alpha, \beta) \in \mathbb{N}$ *such that for all* $N \geq N_0$:

(i) $\sup_{x\in \mathsf{X}} \| P_{T,N}^{k}(x, \cdot) - \pi_T \|_{\mathrm{tv}} \leq \alpha^T \beta^{-k}$ *for all* $k \geq 1$ *and all* $T \geq 1$.
(ii) *For any* $\rho > \log\alpha / \log\beta$, (2) *holds*.

PROOF. The upper bound (i) follows from Theorem 6 and Lemma 30, and (ii) follows directly from (i). The first statement follows because $\log\alpha/\log\beta$ can be taken to be arbitrarily small. □

Theorem 4, indicates that under the strong mixing assumption, the mixing time of CBPF increases at most linearly in the number of observations T. We remark that unlike existing results for the CPF, we do not derive a one-shot coupling bound (Chopin and Singh (2015)), or a one-step minorisation measure (Andrieu, Lee and Vihola (2018), Lindsten, Douc and Moulines (2015)), to prove the uniform ergodicity of the CBPF transition probability $P_{T,N}$. This is because the enhanced stability of CBPF's Markov kernel over the Markov kernel of CPF can only be established by considering the behaviour of the iterated kernel $P_{T,N}^k$ of Theorem 4, which has thus far proven elusive to study. Thus, in addition to the result, the proof technique is itself novel and of interest. For this reason, we dedicate Section 6 to its exposition.

REMARK 5. Intuitively, the arguments used to prove Theorem 4 demonstrate that by increasing N_0, one can take ρ to be smaller. Hence, the qualitative relationship that increasing the number of particles gives faster convergence of the CBPF is captured. However, if one was to pursue quantitative bounds on the dependence between ρ and N_0, our bounds are likely to be too conservative to be useful.

4. Coupled conditional particle filters. This section is devoted to the CCPF and CCBPF algorithms (in short CCXPF where X is a place holder). We start with Algorithm 2, where the CCXPF algorithms are given in pseudo-code. The algorithms differ only in lines 12–17, highlighting the small, but important, difference: the CCBPF incorporates index coupled backward sampling, which is central to our results.

Algorithm 3 details the index coupled resampling (Chopin and Singh (2015)), implementing maximal coupling of $\mathrm{Categ}(\omega^{(1:N)})$ and $\mathrm{Categ}(\tilde{\omega}^{(1:N)})$. Line 7 of Algorithm 2 accommodates any sampling strategy which satisfies $X_t^{(i)} \sim M_t(X_{t-1}^{(I_t^{(i)})}, \cdot)$ and $\tilde{X}_t^{(i)} \sim M_t(\tilde{X}_{t-1}^{(\tilde{I}_t^{(i)})}, \cdot)$ marginally, but may involve dependence, such as implementation using common random number generators (Jacob, Lindsten and Schön (2020)).

The CCXPF algorithms define Markov transition probabilities on $\mathsf{X}^T \times \mathsf{X}^T$. The CCXPF algorithm is a Markovian coupling of the corresponding CXPF algorithm, with the same structure: it is direct to check that CCXPF coincides marginally with CXPF in Algorithm 1, that is, if $(S, \tilde{S}) \leftarrow \mathrm{CCXPF}(s_{\mathrm{ref}}, \tilde{s}_{\mathrm{ref}}, N)$ for some $N \geq 2$ and $s_{\mathrm{ref}}, \tilde{s}_{\mathrm{ref}} \in \mathsf{X}^T$, then $S \overset{d}{=} \mathrm{CXPF}(s_{\mathrm{ref}}, N)$ and $\tilde{S} \overset{d}{=} \mathrm{CXPF}(\tilde{s}_{\mathrm{ref}}, N)$. It is also clear that if $s_{\mathrm{ref}} = \tilde{s}_{\mathrm{ref}}$, then $S = \tilde{S}$. Because CPF and CBPF are both π_T-reversible (Chopin and Singh (2015)), it is easy to see that CCXPF are $\boldsymbol{\pi}_T$-reversible, where $\boldsymbol{\pi}_T(\mathrm{d}s, \mathrm{d}\tilde{s}) = \pi_T(s)\delta_s(\mathrm{d}\tilde{s})\,\mathrm{d}s$. Just as the CPF and CBPF algorithms have time and space complexity $O(TN)$, so do the CCPF and CCBPF algorithms.

Algorithm 2 CCxPF($X^*_{1:T}, \tilde{X}^*_{1:T}, N$)

$(X^{(1)}_{1:T}, \tilde{X}^{(1)}_{1:T}) \leftarrow (X^*_{1:T}, \tilde{X}^*_{1:T})$.
$X^{(i)}_1 \leftarrow \tilde{X}^{(i)}_1 \sim M_1(\cdot)$ for $i \in \{2{:}N\}$.
$\omega^{(i)}_1 \leftarrow G_1(X^{(i)}_1)$; $\tilde{\omega}^{(i)}_1 \leftarrow G_1(\tilde{X}^{(i)}_1)$.
for $t = 2{:}T$ **do**
  $(I^{(2:N)}_t, \tilde{I}^{(2:N)}_t) \leftarrow \text{CRES}(\omega^{(1:N)}_{t-1}, \tilde{\omega}^{(1:N)}_{t-1}, N-1)$.
  $X^{(i)}_t \leftarrow \tilde{X}^{(i)}_t \sim M_t(X^{(I^{(i)}_t)}_{t-1}, \cdot)$ for $i \in \{2{:}N\}$ with $X^{(I^{(i)}_t)}_{t-1} = \tilde{X}^{(\tilde{I}^{(i)}_t)}_{t-1}$
  $(X^{(i)}_t, \tilde{X}^{(i)}_t) \sim (M_t(X^{(I^{(i)}_t)}_{t-1}, \cdot), M_t(\tilde{X}^{(\tilde{I}^{(i)}_t)}_{t-1}, \cdot))$ for $i \in \{2{:}N\}$ with $X^{(I^{(i)}_t)}_{t-1} \neq \tilde{X}^{(\tilde{I}^{(i)}_t)}_{t-1}$
  $\omega^{(i)}_t \leftarrow G_t(X^{(I^{(i)}_t)}_{t-1}, X^{(i)}_t)$; $\tilde{\omega}^{(i)}_t \leftarrow G_t(\tilde{X}^{(\tilde{I}^{(i)}_t)}_{t-1}, \tilde{X}^{(i)}_t)$.
end for
$(J_T, \tilde{J}_T) \leftarrow \text{CRES}(\omega^{(1:N)}_T, \tilde{\omega}^{(1:N)}_T, 1)$
for $t = (T-1){:}1$ **do**
  if CCBPF **do**
    $b^{(i)}_t \leftarrow \omega^{(i)}_t M_{t+1}(X^{(i)}_t, X^{(J_{t+1})}_{t+1}) G_{t+1}(X^{(i)}_t, X^{(J_{t+1})}_{t+1})$
    $\tilde{b}^{(i)}_t \leftarrow \tilde{\omega}^{(i)}_t M_{t+1}(\tilde{X}^{(i)}_t, \tilde{X}^{(\tilde{J}_{t+1})}_{t+1}) G_{t+1}(\tilde{X}^{(i)}_t, \tilde{X}^{(\tilde{J}_{t+1})}_{t+1})$
    $(J_t, \tilde{J}_t) \leftarrow \text{CRES}(b^{(1:N)}_t, \tilde{b}^{(1:N)}_t, 1)$
  if CCPF **do**
    $(J_t, \tilde{J}_t) \leftarrow (I^{(J_{t+1})}_{t+1}, \tilde{I}^{(\tilde{J}_{t+1})}_{t+1})$ where $I^{(1)}_t = \tilde{I}^{(1)}_t = 1$.
end for
output $(X^{(J_{1:T})}_{1:T}, \tilde{X}^{(\tilde{J}_{1:T})}_{1:T})$

4.1. *Convergence of the CCBPF.* In our experiments, the CCBPF had stable behaviour with a fixed and small number of particles, even for large T. Our main result for the CCBPF consolidates our empirical findings. In contrast to most results for the CPF and CCPF, the statement of the coupling behaviour for CCBPF is not one-shot in nature: instead we show that the pair of trajectories output by the repeated application of the CCBPF kernel couple themselves *progressively*, starting from their time 1 components until eventually coupling all their components until time T.

THEOREM 6. *Suppose that Assumption 2 holds. Let* $s_{\text{ref}}, \tilde{s}_{\text{ref}} \in \mathsf{X}^T$ *and let* $(S_0, \tilde{S}_0) \leftarrow (s_{\text{ref}}, \tilde{s}_{\text{ref}})$ *and* $(S_k, \tilde{S}_k) \leftarrow \text{CCBPF}(S_{k-1}, \tilde{S}_{k-1}, N)$ *for* $k \geq 1$. *Denote the coupling time* $\tau := \inf\{k \geq 1 : S_k = \tilde{S}_k\}$. *For any* $\rho > 0$, *there exists* $N_0 = N_0(\epsilon, \rho) < \infty$ *such that for all* $N \geq$

Algorithm 3 CRES($\omega^{(1:N)}, \tilde{\omega}^{(1:N)}, n$)

$w^{(1:N)} \leftarrow \frac{\omega^{(1:N)}}{\sum_{j=1}^N \omega^{(j)}}$; $\tilde{w}^{(1:N)} \leftarrow \frac{\tilde{\omega}^{(1:N)}}{\sum_{j=1}^N \tilde{\omega}^{(j)}}$; $p_c \leftarrow \sum_{j=1}^N w^{(j)} \wedge \tilde{w}^{(j)}$
$w_c^{(1:N)} \leftarrow \frac{w^{(1:N)} \wedge \tilde{w}^{(1:N)}}{p_c}$; $w_r^{(1:N)} \leftarrow \frac{w^{(1:N)} - p_c w_c^{(1:N)}}{1-p_c}$; $\tilde{w}_r^{(1:N)} \leftarrow \frac{\tilde{w}^{(1:N)} - p_c w_c^{(1:N)}}{1-p_c}$
for $i = 1{:}n$ **do**
  with probability p_c **do**
    $\tilde{I}^{(i)} \leftarrow I^{(i)} \sim w_c^{(1:N)}$
  otherwise
    $I^{(i)} \sim w_r^{(1:N)}$; $\tilde{I}^{(i)} \sim \tilde{w}_r^{(1:N)}$.
end for
output $(I^{(1:n)}, \tilde{I}^{(1:n)})$

N_0,

$$\lim_{T\to\infty} \mathbb{P}\big(\tau \geq \lceil \rho T \rceil\big) = 0. \tag{3}$$

More precisely, for any $\alpha, \beta \in (1, \infty)$ *there exists* $N_0 = N_0(\epsilon, \alpha, \beta) \in \mathbb{N}$ *such that for all* $N \geq N_0$,

$$\mathbb{P}(\tau \geq n) \leq \alpha^T \beta^{-n} \quad \textit{for all } n, T \in \mathbb{N}. \tag{4}$$

In particular, for any $\rho > \log(\alpha)/\log(\beta)$, (3) *holds*.

The proof of the bound (4) is given in Section 6, and the linear coupling time statement follows by appropriate choice of α and β. The most striking element of this statement is that the coupling time τ does not exceed ρT with greater surety as T increases.

REMARK 7. The comments in Remark 5 also apply here. Regarding tightness of Theorem 6, one may consider whether coupling might occur with high probability after a number of iterations that is sublinear in T. The simulation experiments in Section 7 suggest that the mean coupling time for the CCBPF is indeed linear in T for the models considered, suggesting that Theorem 6 is tight, and we suspect that this is the case for many other models. Similarly, in the more challenging simulation experiment, we found that values of N that were too small had mean coupling times that were superlinear in T, suggesting that there is indeed a model-dependent, minimal N_0 for linear-in-time coupling.

4.2. *Convergence of the CCPF.* We seek here to provide quantitative results to strengthen Theorem 3.1 of Jacob, Lindsten and Schön (2020), which does not quantify the dependence of the probability of coupling on N or T. Although the results are less encouraging than for the CCBPF, the dependence on T is likely to be very conservative for many models. On the other hand, results for the CCPF are interesting because it is more widely applicable than the CCBPF, specifically since implementing the CCPF does not require the ability to calculate densities M_t.

In empirical investigations, we have only seen the CCPF couple instantaneously, as opposed to the progressive coupling seen for the CCBPF. For that reason, we focus here on lower bounding the one-shot coupling probability for two arbitrary reference trajectories.

Our first result states that with T fixed, the CCPF enjoys similar strong uniform ergodicity to the CPF, with the same rate as the number of particles N is increased (cf. Andrieu, Lee and Vihola (2018), Lindsten, Douc and Moulines (2015)).

THEOREM 8. *Let* $s_{\mathrm{ref}}, \tilde{s}_{\mathrm{ref}} \in \mathsf{X}^T$, *and consider* $(S, \tilde{S}) \leftarrow \mathrm{CCPF}(s_{\mathrm{ref}}, \tilde{s}_{\mathrm{ref}}, N)$ *with* $N \geq 2$. *If* $G_t(x_{t-1}, x_t) = G_t(x_t)$ *for* $t \geq 2$ *and Assumption 1 holds, then there exists a constant* $c = c(G^*, T, c_T) \in (0, \infty)$ *such that*

$$\mathbb{P}(S = \tilde{S}) \geq 1 - \frac{c}{N + c}.$$

The proof of Theorem 8 is given in Appendix A.

Theorem 8 is stated with a fixed time horizon T, and shows that one-shot coupling occurs from any initial state $(s_{\mathrm{ref}}, \tilde{s}_{\mathrm{ref}})$ with positive probability for any $N \geq 2$. To have a reasonably large probability of one-shot coupling, it is sufficient to choose a large enough value of N.

Although Theorem 8 holds very generally, it does not provide useful quantitative bounds on the relationship between N and T; in particular, $c(G^*, T, c_T)$ may grow very quickly with T so that N would need to grow similarly quickly to control the coupling probability. In order to provide more accurate bounds relating N and T, we have had to make the strong mixing assumption.

THEOREM 9. *Under the setting of Theorem 8, but with Assumption 2,*

$$\mathbb{P}(S = \tilde{S}) \geq 1 - \frac{2^T T}{(2c_*)^{-1}(N-1)+1}.$$

Theorem 9, which follows from Lemmas 20 and 23 in Appendix B, shows that the probability of coupling does not diminish when $N = O(2^T T)$. That is, roughly doubling of particle number with every unit increase in T ensures nondiminishing coupling probability. Experiments by Jacob, Lindsten and Schön (2020) and also those in Section 7 suggest that a rate $N = O(T)$ might be enough for some models, which would be analogous to the results on the CPF (Andrieu, Lee and Vihola (2018), Lindsten, Douc and Moulines (2015)), but we have been unable to verify such a rate theoretically. Our empirical results in Section 7 also suggest that taking N superlinear in T may be necessary in other models.

5. Unbiased estimators. Let us then turn to the use of CCxPF together with the scheme of Glynn and Rhee (2014), to construct unbiased estimators, as suggested in (Jacob, Lindsten and Schön (2020)). Algorithm 4 aims to unbiasedly estimate the expectation $\mathbb{E}_{\pi_T}[h(X_{1:T})]$, where $h : \mathsf{X}^T \to \mathbb{R}$ is any integrable function of interest. We remark that the unbiased estimation procedure is not specific to the choice of h. For example, in a HMM one can approximate smoothing means and covariances by choosing several appropriate h functions. Algorithm 4 has two adjustable parameters, a 'burn-in' $b \geq 1$ and 'number of particles' $N \geq 2$ which may be tuned to maximise its efficiency. Algorithm 4 iterates either the coupled conditional particle filter CCPF or the coupled conditional backward sampling particle filter CCBPF until a perfect coupling of the trajectories S_n and $\tilde{S}_n$ is obtained.

The following result records general conditions under which the scheme above produces unbiased finite variance estimators.

THEOREM 10. *Suppose $G_t(x_{t-1}, x_t) = G_t(x_t)$ and Assumption 1 holds, $h : \mathsf{X}^T \to \mathbb{R}$ is bounded and measurable. Then Algorithm 4 with CCPF, $b \geq 1$ and $N \geq 2$ satisfies, denoting by τ the running time (iterations before producing output):*

(i) $\tau < \infty$ *almost surely.*
(ii) $\mathbb{E}[Z] = \mathbb{E}_{\pi_T}[h(X)]$ *and* $\operatorname{var}(Z) < \infty$.
(iii) *With the constant $c = c(G^*, T, c_T) \in (0, \infty)$ of Theorem 8,*

$$\mathbb{E}[\tau] \leq b + \left(\frac{c}{N+c}\right)^{b-1}\left(\frac{N+c}{N}\right),$$

$$\left|\operatorname{var}(Z) - \operatorname{var}_{\pi_T}(h(X))\right| \leq 16\|\bar{h}\|_\infty^2 \left(\frac{N+c}{N}\right)^2 \left(\frac{c}{N+c}\right)^{b/2}$$

where $\bar{h}(\cdot) = h(\cdot) - \pi_T(h)$.

Algorithm 4 UNBIASED(h, b, N)

1: Run particle filter with $(M_t, G_t)_{t=1:T}$ independently to get trajectories $\tilde{S}_0, S_{-1} \in \mathsf{X}^T$
2: Set $(S_0, -) \leftarrow$ CCXPF(S_{-1}, S_{-1}, N).
3: **for** $n = 1, 2, \ldots$ **do**
4: $\quad (S_n, \tilde{S}_n) \leftarrow$ CCXPF$(S_{n-1}, \tilde{S}_{n-1}, N)$
5: $\quad$ **if** $S_n = \tilde{S}_n$ and $n \geq b$ **then output** $Z := h(S_b) + \sum_{k=b+1}^{n}[h(S_k) - h(\tilde{S}_k)]$
6: **end for**

PROOF. Theorem 8 implies that $\mathbb{P}(\tau > k) \leq (\frac{c}{N+c})^{k-1}$ for all $k > b$, from which

$$\mathbb{E}[\tau] = \sum_{k \geq 0} \mathbb{P}(\tau > k) \leq b + \sum_{k \geq b} \mathbb{P}(\tau > k) \leq b + \left(\frac{c}{N+c}\right)^{b-1} \left(\frac{N+c}{N}\right) < \infty,$$

and the bound on $|\mathrm{var}(Z) - \mathrm{var}_{\pi_T}(h(X))|$ follows from Lemma 31. Part (ii) follows from Theorem 27 and Lemma 29. □

Theorem 10 extends the consistency result of Jacob, Lindsten and Schön (2020), by quantifying the convergence rates. Fix T: if N is large, then $\mathbb{E}[\tau] \approx b$, and if b is large, then $\mathrm{var}(Z) \approx \mathrm{var}_{\pi_T}(h(X))$. As mentioned after Theorem 6, the growth of the constant c with respect to T can be very rapid. In contrast, in case of the CCBPF, the results may be refined as follows.

THEOREM 11. *Suppose Assumption 2 holds, let $\alpha, \beta \in (1, \infty)$ and let $N_0 \in \mathbb{N}$ be from Theorem 6. Then Algorithm 4 with CCBPF and $b \geq 1$ satisfies, with $\rho := \log \alpha / \log \beta$:*

(i) $\mathbb{E}[\tau] \leq b \vee \lceil \rho T \rceil + \alpha^T \beta^{-(\lceil \rho T \rceil \vee b)} (\beta - 1)^{-1}$.
(ii) $|\mathrm{var}(Z) - \mathrm{var}_{\pi_T}(h(X))| \leq 16 \alpha^T (1 - \beta^{-1})^{-2} \beta^{-b/2} \|\bar{h}\|_\infty^2$.
In particular, if $b = \lceil \tilde{\rho} T \rceil$ with any $\tilde{\rho} > 2\rho$, $|\mathrm{var}(Z) - \mathrm{var}_{\pi_T}(h(X))| \to 0$ as $T \to \infty$.

PROOF. The results follow from Theorem 6 and Lemma 31, similarly as in the proof of Theorem 10. □

Note that the latter term in Theorem 11 (i) is at most $(\beta - 1)^{-1}$, showing that the expected coupling time is linear in T. Theorem 11(ii) may be interpreted so that the CCBPF algorithm is almost equivalent with perfect sampling from π_T, when b is increased linearly with respect to T.

We conclude the section with a number of remarks about Algorithm 4:

(i) We follow Jacob, Lindsten and Schön (2020) and suggest an initialisation based on a standard particle filter in line 1. However, this initialisation may be changed to any other scheme, which ensures that S_0 and $\tilde{S}_1$ have identical distributions. Our results above do not depend on the chosen initialisation strategy.

(ii) The estimator Z is constructed for a single function $h : \mathsf{X}^T \to \mathbb{R}$, but several estimators may be constructed simultaneously for a number of functions $h_1, \ldots, h_m$. In fact, as Glynn and Rhee (2014) note, if we let $\tau := \inf\{n \geq b : S_n = \tilde{S}_n\}$, we may regard the random signed measure

$$\hat{\mu}_b(\cdot) := \delta_{S_b}(\cdot) + \sum_{k=b+1}^{\tau} \left[\delta_{S_k}(\cdot) - \delta_{\tilde{S}_k}(\cdot)\right]$$

as the output, which will satisfy the unbiasedness $\mathbb{E}[\hat{\mu}_b(\varphi)] = \pi_T(\varphi)$ at least for all bounded measurable $\varphi : \mathsf{X} \to \mathbb{R}$.

(iii) It is also possible to construct a 'time-averaged' estimator that corresponds to a weighted average of the estimators $\hat{\mu}_b$ over a range of values for b (Jacob, O'Leary and Atchadé (2020)).

(iv) We believe that the method is valid also without Assumption 1 but may exhibit poor performance—similar to the conditional particle filter, which is sub-geometrically ergodic with unbounded potentials (Andrieu, Lee and Vihola (2018)).

6. Coupling time of CCBPF. Consider now the Markov chain $(S_k, \tilde{S}_k)_{k\geq 1}$ defined by Algorithm 4, with the stopping criterion (line 5) omitted. Define the 'perfect coupling boundary' as

$$\kappa_n := \kappa(S_n, \tilde{S}_n) := \max\{t \geq 0 : S_{n,1:t} = \tilde{S}_{n,1:t}\}.$$

We are interested in upper bounding the stopping time $\tau := \inf\{n \geq 1 : S_n = \tilde{S}_n\} = \inf\{n \geq 1 : \kappa_n = T\}$.

Since the CCBPF is complicated, in our analysis we instead focus on a simplified Markov chain that considers only the vector of numbers of identical particles at each time $t \in \{1{:}T\}$. The boundary associated with this simpler chain grows by i.i.d. increments, which are stochastically ordered with respect to the increments of the CCBPF boundary increments, ultimately allowing us to upper bound the stopping time.

We use stochastic ordering $X \leq_{\mathrm{st}} Y$ of two random variables X and Y, which holds if their distribution functions are ordered $\mathbb{P}(X \leq x) \geq \mathbb{P}(Y \leq x)$ for all $x \in \mathbb{R}$. Two random vectors X and Y are ordered $X \leq_{\mathrm{st}} Y$ if $\mathbb{E}[\phi(X)] \leq \mathbb{E}[\phi(Y)]$ for all functions $\phi : \mathbb{R}^n \to \mathbb{R}$ for which the expectations exist, and which are increasing, in the sense that $\phi(x) \leq \phi(y)$ whenever $x \leq y$, where '$\leq$' is the usual partial order $x \leq y$ if $x_i \leq y_i$ for all $i \in \{1{:}d\}$. Recall also that $X \leq_{\mathrm{st}} Y$ if and only if there exists a probability space with random variables $\tilde{X}$ and $\tilde{Y}$) such that $X \overset{d}{=} \tilde{X}$ and $Y \overset{d}{=} \tilde{Y}$ and $\tilde{X} \leq \tilde{Y}$ a.s. Shaked and Shanthikumar (2007), Theorem 6.B.1.

Our bound of τ is based on an independent random variable Δ, which satisfies $\kappa_{n+1} - \kappa_n \geq_{\mathrm{st}} \Delta \wedge (T - \kappa_n)$, under Assumption 2.

LEMMA 12. *Suppose Assumption 2 holds, and consider the output of Algorithm 2 (CCBPF). The perfect coupling boundaries satisfy*

$$\kappa\big(X_{1:T}^{(J_{1:T})}, \tilde{X}_{1:T}^{(\tilde{J}_{1:T})}\big) - \kappa\big(X_{1:T}^{*}, \tilde{X}_{1:T}^{*}\big) \geq_{\mathrm{st}} \Delta \wedge \big(T - \kappa\big(X_{1:T}^{*}, \tilde{X}_{1:T}^{*}\big)\big),$$

where the random variable Δ is defined through the following procedure:

(i) *Let $\hat{C}_t = N$ for $t \leq 0$ and $\hat{C}_1 = N - 1$, and set $s = 1$. While $\hat{C}_s > 0$:*

- *Simulate $\hat{C}_{s+1} \sim \mathrm{Binom}(N - 1, \frac{\delta \hat{C}_s}{\delta \hat{C}_s + (N - \hat{C}_s)})$.*
- *Let $s \leftarrow s + 1$.*

(ii) *Set $\hat{C}_t = 0$ for $t > s$, $\xi_s = 0$ and $t = s - 1$. While $t \geq 0$ or $\xi_{t+1} = 0$:*

- *Simulate $\xi_t \sim \mathrm{Bernoulli}(p_t)$, where*

$$p_t := \begin{cases} p_t^{(0)} := \epsilon \dfrac{\hat{C}_t}{N} & \xi_{t+1} = 0 \\ p_t^{(1)} := \dfrac{\hat{C}_t \epsilon}{\hat{C}_t \epsilon + N - \hat{C}_t} & \xi_{t+1} = 1. \end{cases}$$

- *Let $t \leftarrow t - 1$.*

(iii) *Set $\Delta \leftarrow \min\{i \geq t : \xi_i = 0\} - 1$.*

PROOF. Denote in short $\kappa = \kappa(X_{1:T}^{*}, \tilde{X}_{1:T}^{*})$, and the indices of the coupled particles

$$C_t := \big\{j \in \{1{:}N\} : X_t^{(I_t^{(j)})} = \tilde{X}_t^{(\tilde{I}_t^{(j)})}\big\} \quad \text{for } t \in \{1{:}T\}.$$

Then the sizes of C_t satisfy the following:

$$|C_t| = N, \quad t = 1{:}\kappa,$$

$$|C_t| \,\Big|\, C_{1:t-1} \geq_{\mathrm{st}} \mathrm{Binom}\Big(N - 1, \frac{\delta |C_{t-1}|}{\delta |C_{t-1}| + N - |C_{t-1}|}\Big), \quad t = (\kappa + 1){:}T,$$

where the latter follows by Lemma 13(ii). As the function $c \mapsto \delta c(N-(1-\delta)c)^{-1}$ is increasing in c, and $\mathrm{Binom}(n,p) \geq_{\mathrm{st}} \mathrm{Binom}(n,p')$ for $p \geq p'$, it follows that $(|C_1|,\dots,|C_T|) \geq_{\mathrm{st}} \hat{C}_{(1-\kappa):(T-\kappa)}$ Shaked and Shanthikumar (2007), Theorem 6.B.3. This means that we may construct (by a suitable coupling) $\hat{C}_t$ such that $|C_t| \geq \hat{C}_{t-\kappa}$ for all $t \in \{1{:}T\}$.

By Lemma 13, the backward sampling indices satisfy:

$$\mathbb{P}(J_T = \tilde{J}_T \in C_T \mid C_{1:T}) \geq \frac{\delta|C_T|}{\delta|C_T| + N - |C_T|},$$

$$\mathbb{P}(J_t = \tilde{J}_t \in C_t \mid C_{1:T}, J_{t+1:T}) \geq \begin{cases} \dfrac{\epsilon|C_t|}{\epsilon|C_t| + N - |C_t|} & J_{t+1} = \tilde{J}_{t+1} \in C_{t+1}, \\ \epsilon\dfrac{|C_t|}{N} & \text{otherwise,} \end{cases}$$

for $t = 1{:}(T-1)$. By definition, $\delta \geq \epsilon$ and, therefore, $\frac{\delta c}{\delta c+N-c} \geq \frac{\epsilon c}{\epsilon c+N-c} \geq \frac{\epsilon c}{N}$. This, together with $|C_T| \geq_{\mathrm{st}} \hat{C}_{T-\kappa}$ implies that $\mathbb{P}(J_T = \tilde{J}_T \in C_T) \geq \mathbb{P}(\xi_{T-\kappa} = 1)$. Similarly, by Shaked and Shanthikumar (2007), Theorem 6.B.3, we deduce that

$$\big(\mathbb{I}\{J_1 = \tilde{J}_1 \in C_1\},\dots,\mathbb{I}\{J_T = \tilde{J}_T \in C_T\}\big) \geq_{\mathrm{st}} \xi_{(1-\kappa):(T-\kappa)}.$$

Because the functions $\phi_t(x_{1:T}) = \prod_{u=1}^{t} \max\{0, x_u\}$ are increasing, the claim follows. □

LEMMA 13. *Suppose* $0 < \omega_* \leq \omega^* < \infty$ *and* $\omega^{(i)}, \tilde{\omega}^{(i)} \in [\omega_*, \omega^*]$ *for* $i \in \{1{:}N\}$. *Let*

$$\varepsilon := \frac{\omega_*}{\omega^*} \quad \text{and} \quad C := \big\{j \in \{1{:}N\} : \omega^{(j)} = \tilde{\omega}^{(j)}\big\}.$$

Then $(I^{(1:n)}, \tilde{I}^{(1:n)}) \sim \mathrm{CRES}(\omega^{(1:N)}, \tilde{\omega}^{(1:N)}, n)$ *satisfy the following for all* $j \in \{1{:}n\}$:

(i) $\mathbb{P}(I^{(j)} = \tilde{I}^{(j)} = i) \geq \frac{\varepsilon}{N}$ *for all* $i = 1{:}N$,
(ii) $\mathbb{P}(I^{(j)} = \tilde{I}^{(j)} \in C) \geq \frac{|C|\varepsilon}{|C|\varepsilon+N-|C|}$.

PROOF. Note that $\mathbb{P}(I^{(j)} = \tilde{I}^{(j)} = i) = w^{(i)} \wedge \tilde{w}^{(i)}$, so the first bound is immediate. For the second, let $C^c := \{1{:}N\} \setminus C$, and observe that

$$\sum_{j\in C} w^{(j)} \wedge \tilde{w}^{(j)} = \frac{\sum_{j\in C} \omega^{(j)}}{\sum_{i\in C} \omega^{(i)} + (\sum_{i\in C^c} \omega^{(i)}) \vee (\sum_{i\in C^c} \tilde{\omega}^{(i)})}$$

$$\geq \frac{|C|\omega_*}{|C|\omega_* + |C^c|\omega^*},$$

because $x \mapsto x(x+b)^{-1}$ is increasing for $x \geq 0$ for any $b > 0$. The last bound equals (ii). □

Because $\kappa_{n+1} - \kappa_n \geq_{\mathrm{st}} \Delta \wedge (T - \kappa_n)$, we note that $\tau \leq_{\mathrm{st}} \hat{\tau}$, where

$$\hat{\tau} := \inf\left\{n \geq 0 : \sum_{k=1}^{n} \Delta_k \geq T\right\}, \tag{5}$$

and Δ_k are independent realisations of Δ in Lemma 12. The next lemma indicates that if N is large enough (given δ, ϵ), the random variables Δ are well behaved, and ensure good expectation and tail probability bounds for $\hat{\tau}$.

LEMMA 14. *Given any* $N \in \mathbb{N}$, *consider the random variable* Δ *defined in Lemma 12. For any* $\beta < \infty$ *and* $\alpha \in (1, (1-\epsilon)^{-1})$, *there exists* $N_0 < \infty$ *such that for all* $N \geq N_0$,

$$\mathbb{E}[\Delta] \geq \beta \quad \text{and} \quad \mathbb{E}\big[\alpha^{-\Delta}\big] \leq \beta^{-1}.$$

PROOF. Suppose $\varphi : \mathbb{N} \to \mathbb{R}_+$ is decreasing and $L \in \mathbb{N}$, then

$$\begin{aligned}\mathbb{E}[\varphi(\Delta) \mid \hat{C}_{1:s}] &= \sum_{t=-\infty}^{s} \mathbb{P}(\xi_{-\infty:t} = 1, \xi_{t+1} = 0 \mid \hat{C}_{1:s})\varphi(t) \\ &\le \sum_{t=-\infty}^{L\wedge s} \mathbb{P}(\xi_{-\infty:t} = 1, \xi_{t+1} = 0 \mid \hat{C}_{1:s})\varphi(t) + \varphi(L),\end{aligned}$$

and because $p_u^{(1)} = 1$ and $p_u^{(0)} = \epsilon$ for $u \le 0$, we may write

$$\begin{aligned}&\mathbb{P}(\xi_{-\infty:t} = 1, \xi_{t+1} = 0 \mid \hat{C}_{1:s}) \\ &\quad= \Bigg[\prod_{u=1}^{t-1} p_u^{(1)}\Bigg] p_t^{(0)} \Bigg[\prod_{u=t+1}^{0} (1-\epsilon)\Bigg] \mathbb{P}(\xi_{(t+1)\vee 1} = 0 \mid \hat{C}_{1:s}).\end{aligned}$$

Furthermore, for $t \in \{1{:}L\}$,

$$\begin{aligned}\mathbb{P}(\xi_t = 0 \mid \hat{C}_{1:s}) &= \mathbb{P}(\xi_{t:(L+1)} = 0 \mid \hat{C}_{1:s}) + \sum_{b\in\{0,1\}^{L-t}, b\neq 0} \mathbb{P}(\xi_t = 0, \xi_{t+1:(L+1)} = b \mid \hat{C}_{1:s}) \\ &\le \prod_{u=t}^{L}(1 - p_u^{(0)}) + \sum_{u=t}^{L}(1 - p_u^{(1)}).\end{aligned}$$

Lemma 15 implies that for $t = 1{:}L$, the terms $R_t := \hat{C}_t/N \to 1$ in probability as $N \to \infty$, and consequently also $p_t^{(1)} \to 1$ and $p_t^{(0)} \to \epsilon$. We conclude that whenever $\sum_{t<0}(1-\epsilon)^t\varphi(t) < \infty$,

$$\limsup_{N\to\infty} \mathbb{E}[\varphi(\Delta)] \le \varphi(L) + \sum_{t=-\infty}^{L} (1-\epsilon)^{L-t}\epsilon\varphi(t) = \varphi(L) + \sum_{t=0}^{\infty}(1-\epsilon)^t\epsilon\varphi(L-t).$$

We get the first bound by and applying the result above with $\varphi(t) = (L-t)_+$, because $\mathbb{E}[\Delta] \ge L - \mathbb{E}[(L-\Delta)_+]$, and $\limsup_{N\to\infty} \mathbb{E}[(L-\Delta)_+] \le 1 - \epsilon^{-1}$. The second bound follows by taking $\varphi(t) = \alpha^t$, because

$$\limsup_{N\to\infty} \mathbb{E}[\varphi(\Delta)] \le \alpha^{-L} + \sum_{t=0}^{\infty}(1-\epsilon)^t\epsilon\alpha^{t-L} = \alpha^{-L}\big[1 + \epsilon\big(1 - (1-\epsilon)\alpha\big)^{-1}\big].$$

□

LEMMA 15. *The expectation of $\hat{C}_t$ generated in Lemma 12 may be lower bounded as follows*:

$$\mathbb{E}\left[\frac{\hat{C}_t}{N}\right] \ge \frac{\delta_N^{t-1}(N-1)}{1 + \delta_N^{t-1}(N-1)} \quad \textit{where } \delta_N := \frac{N-1}{N}\boldsymbol{\delta}.$$

Therefore, for any $t \in \mathbb{N}$ and $\varepsilon > 0$, there exists N_0 such that for all $N \ge N_0$ and all $u = 1{:}t$, $\mathbb{E}[\hat{C}_u/N] \ge 1 - \varepsilon$.

PROOF. Denote $R_t := \hat{C}_t/N$, then for any $t \ge 1$, we have

$$\mathbb{E}[R_t \mid R_{t-1}] = \frac{\delta_N R_{t-1}}{1 - (1-\boldsymbol{\delta})R_{t-1}}.$$

Note that for $a, b > 0$ and $\lambda \in [0, b)$, the function $x \mapsto ax(b - \lambda x)^{-1}$ is convex on $[0, 1]$. Therefore, by Jensen's inequality,

$$\mathbb{E}\left[\frac{a_t R_t}{b_t - \lambda_t R_t} \,\middle|\, R_{t-1}\right] \ge \frac{a_t\mathbb{E}[R_t \mid R_{t-1}]}{b_t - \lambda_t\mathbb{E}[R_t \mid R_{t-1}]}$$

$$= \frac{a_t \delta_N R_{t-1}}{b_t[1-(1-\boldsymbol{\delta})R_{t-1}] - \lambda_t \delta_N R_{t-1}}$$
$$= \frac{a_{t-1} R_{t-1}}{b_{t-1} - \lambda_{t-1} R_{t-1}},$$

where $a_{t-1} = \delta_N a_t$, $b_{t-1} = b_t$ and $\lambda_{t-1} = \delta_N \lambda_t + (1-\boldsymbol{\delta}) b_t$. Starting with $a_t = 1$, $b_t = 1$ and $\lambda_t = 0$, we conclude that

$$a_1 = \delta_N^{t-1}, \qquad b_1 = 1 \quad \text{and} \quad \lambda_1 = (1-\boldsymbol{\delta}) \sum_{k=0}^{t-2} \delta_N^k = (1-\boldsymbol{\delta}) \frac{1-\delta_N^{t-1}}{1-\delta_N},$$

and consequently,

$$\mathbb{E}[R_t] \geq \frac{a_1 R_1}{b_1 + \lambda_1 R_1} = \frac{\delta_N^{t-1} \frac{N-1}{N}}{1-(1-\boldsymbol{\delta}) \frac{1-\delta_N^{t-1}}{1-\delta_N} \frac{N-1}{N}} \geq \frac{\delta_N^{t-1} \frac{N-1}{N}}{1-(1-\delta_N^{t-1}) \frac{N-1}{N}},$$

because $(1-\delta_N) > (1-\boldsymbol{\delta})$. This equals the desired bound. □

REMARK 16. In order to make the bound in Lemma 15 large, we must have $\delta_N^{t-1}(N-1) \gg 1$. Because $\delta_N^{t-1} \geq (1-t/N)\boldsymbol{\delta}^{t-1}$, it is sufficient that we take $N \gg t$ and $\log(N) \geq c + t(-\log \boldsymbol{\delta})$. Usually the latter is dominant, meaning that N of order $\boldsymbol{\delta}^{-t}$ is necessary.

We simulated the random variables $\hat{C}_t$, and observed a similar 'cutoff'—a $\boldsymbol{\delta}^{-1}$-fold increase in N caused the 'boundary' where $\hat{C}_t/N$ starts to drop from zero around one to zero, to extend by one step further. We believe that Lemma 15 captures the behaviour of $\hat{C}_t$ rather accurately. However, we believe that $\hat{C}_{t-\kappa}$ are often rather pessimistic compared with the actual couplings $|C_t|$.

PROOF OF THEOREM 6. The result follows from the fact that $\mathbb{P}(\tau \geq n) \leq \mathbb{P}(\hat{\tau} \geq n)$ where $\hat{\tau}$ is given in (5), and a Chernoff bound

$$\mathbb{P}(\hat{\tau} \geq n) \leq \mathbb{P}\left(\sum_{i=1}^{n} \Delta_i \leq T\right) \leq \min_{u>0} e^{uT} \prod_{i=1}^{n} \mathbb{E}\left[e^{-u\Delta_i}\right] \leq \hat{\alpha}^T \beta^{-n},$$

where the final inequality uses Lemma 14 by choosing $u = \log \hat{\alpha}$, for some $\hat{\alpha} \leq \alpha$ such that $\hat{\alpha} \in (1, (1-\boldsymbol{\epsilon})^{-1})$. □

7. Empirical comparison. We compare the CCBPF with the CCPF, as well as the CCPF with ancestor sampling proposed by Jacob, Lindsten and Schön (2020) that was found to outperform the CCPF in their experiments. Like the CCBPF, the CCPF with ancestor sampling is a coupling of the CBPF. To emphasize the main difference between the algorithms, we abbreviate the variants according to the way ancestors are sampled as declared in lines 15 and 17 of Algorithm 2. The CCPF is denoted AT (ancestor tracing), the CCBPF is denoted BS (backward sampling) and the CCPF with ancestor sampling is denoted AS (ancestor sampling).

The computational cost of Algorithm 4 is the coupling time τ multiplied by the cost of each run of the CCxPF, at least when the burn-in b is small. We refer to this as the total cost of coupling. For AT, AS and BS the cost per iteration is $O(NT)$, and AS and BS are often around twice the cost of AT. Hence the expected total cost of coupling is $\mathbb{E}[\tau]NT$, up to some constant depending on the model, with an additional factor of about 2 for AS and BS. We investigate below the dependence of τ and $\mathbb{E}[\tau]$ on N and T for two models:

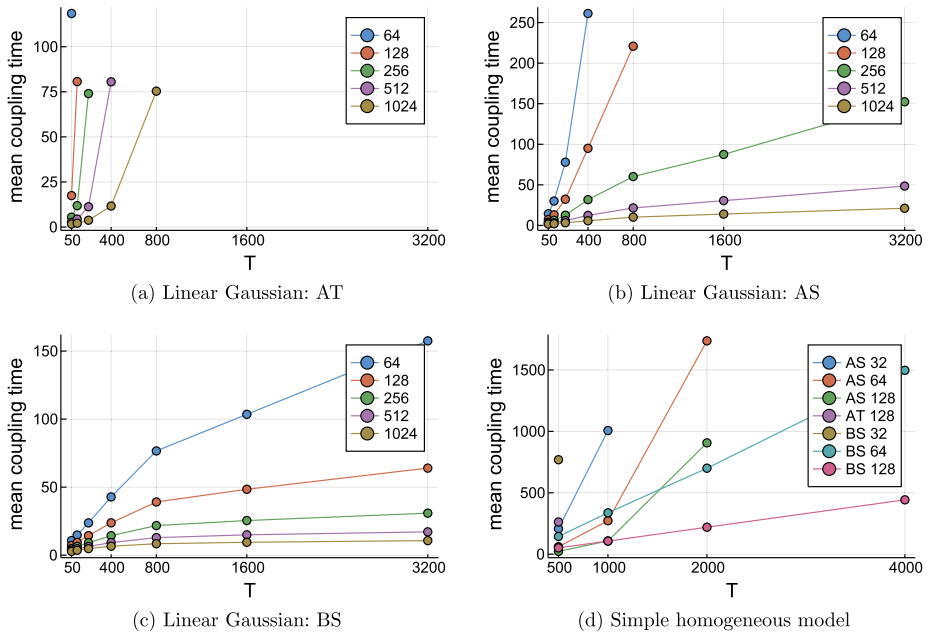
1. the linear Gaussian model used by Jacob, Lindsten and Schön (2020) with the same data;

2. a simple homogeneous model where the transitions correspond to a simple random walk model with standard normal increments, and the potential functions are $G_t(x_t) = \mathbf{1}_{[-s,s]}(x_t)$ for $t \in \{1, \ldots, T\}$, with either $s = 5$ or $s = 10$.

To give an idea of computational cost in seconds, the cost of running AT, AS and BS with $(T, N) = (1000, 1024)$ for the linear Gaussian model is respectively approximately 100 ms, 160 ms and 160 ms on both a Xeon E5-2667v3 CPU and a 2018 Macbook Air.

For fairer comparison, we report results with proposals using common random numbers to execute line 7 of Algorithm 2, as suggested in (Jacob, Lindsten and Schön (2020)). Our implementation of AT and AS differs from theirs only by minor modifications that empirically have no substantive effect on any important characteristics of the algorithm; for instance, they quantile-couple the residual indices (cf. line 7 of Algorithm 3). In Appendix D, we report the results of similar experiments run without common random numbers in which both AT and AS perform considerably worse, while BS is only slightly affected. This suggests that accurate analysis of AT or AS may require analysis of the effect of common random numbers which has thus far not been undertaken here or in Jacob, Lindsten and Schön (2020).

In Figures 1 (a)–(c), we plot the mean coupling times from 1000 replications using the linear Gaussian model, for every combination of $T \in \{50, 100, 200, 400, 800, 1600, 3200\}$ and $N \in \{64, 128, 256, 512, 1024\}$. If at least one replication resulted in a coupling time above 2000, the results for that (T, N) combination are excluded. For AT, we see that N needs to grow roughly linearly in T to maintain the same mean coupling time, and that smaller values of N do not lead to successful coupling in a reasonable amount of time: for $T \in \{1600, 3200\}$, even $N = 1024$ was insufficient to have a reasonable mean coupling time. AS and BS exhibit somewhat similar characteristics, with BS having a smaller mean coupling time for T large. For $N \in \{64, 128\}$, AS did not have a reasonable mean coupling time for large T, unlike BS. We complement Figures 1 (a)–(c) with Table 1 for combinations of (T, N) in which all algorithms are somewhat competitive. The table indicates that in this regime, BS coupling times have a smaller variance.



(a) Linear Gaussian: AT

(b) Linear Gaussian: AS

(c) Linear Gaussian: BS

(d) Simple homogeneous model

FIG. 1. *Mean coupling times associated with ancestor tracing* (*AT*), *ancestor sampling* (*AS*) *and backward sampling* (*BS*). *For* (*d*), *the lines are coloured according to the type of algorithm and the number of particles* N.

TABLE 1
Average of 1000 *coupling times* (*with standard deviations*), *with different variants of coupled ancestor tracing* (*AT*), *ancestor sampling* (*AS*) *and backward sampling* (*BS*)

T	50				100				200				400			
N	64		128		128		256		256		512		512		1024	
AT	122.3	(131.2)	17.3	(17.1)	77.3	(82.0)	12.3	(11.2)	68.2	(67.5)	10.9	(9.6)	81.5	(76.6)	11.7	(9.9)
AS	14.2	(11.0)	7.2	(5.9)	13.0	(10.4)	6.3	(4.5)	12.2	(8.8)	5.9	(4.1)	12.5	(8.2)	5.9	(3.5)
BS	11.0	(5.2)	6.9	(3.0)	9.5	(3.3)	6.3	(2.0)	9.2	(2.5)	6.4	(1.7)	9.4	(2.2)	6.6	(1.6)

In Figure 1(d), we plot the mean coupling times for all of the approaches for the simple homogeneous model with $s = 10$. We used all combinations of $T \in \{500, 1000, 2000, 4000\}$ and $N \in \{16, 32, 64, 128\}$, with results for a combination excluded if a replication resulted in a coupling time above $10T$. This is a homogeneous model, and we see that BS has a mean coupling time that is roughly linear in T. In contrast, the mean coupling time appears to grow superlinearly for AS. We observed also that for large T, BS did not result in a steady increase of the coupling boundary for $N \leq 32$. This suggests that there is indeed a threshold value of N_0 necessary for linear in T mixing: it seems one cannot take $N_0 = 2$ universally in Theorem 6 with an appropriately large ρ. While this model is relatively simple, it is challenging for coupled conditional particle filters because the potential functions are not very informative about the location of a particle in space.

The two examples show quite different behaviour for AS: in the linear Gaussian model it appears that AS may have linear-in-time convergence for the range of T considered, even if its performance appears to be worse than that of BS. However, for the simple but challenging homogeneous model AS appears not to enjoy this property. We suspect that the good behaviour of AS in the former case is due to the combination of using common random numbers (cf. Figure 5) and the fact that the potential functions are highly informative about particle locations. It is not clear that this type of behaviour will extend to more challenging scenarios, such as when the state space is higher dimensional or the smoothing distribution is multimodal.

Figure 2 shows coupling boundaries (see Section 6) by iteration of a single run of each method in the linear Gaussian model, illustrating typical progressive behaviour of the coupling boundary with BS, in contrast with AS which does not clearly display a drift towards complete coupling, and AT which makes no progress at all. The BS appears viable with much smaller number of particles, and suggests that the computationally optimal number of particles with BS may differ significantly from that of AT and AS.

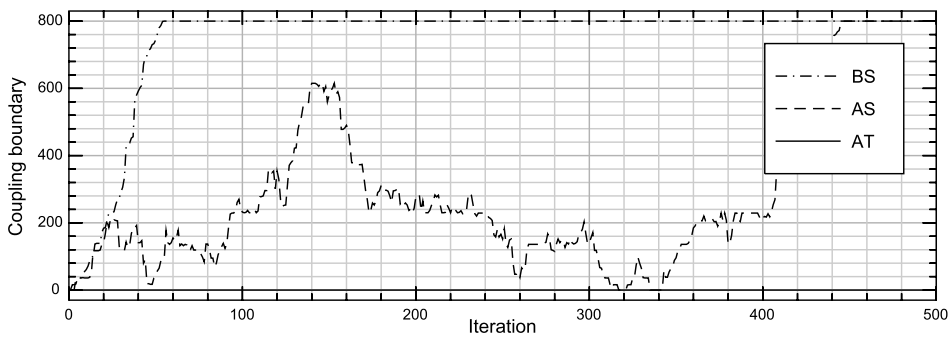


FIG. 2. *One realisation of coupling boundaries with* $T = 800$ *and* $N = 64$.

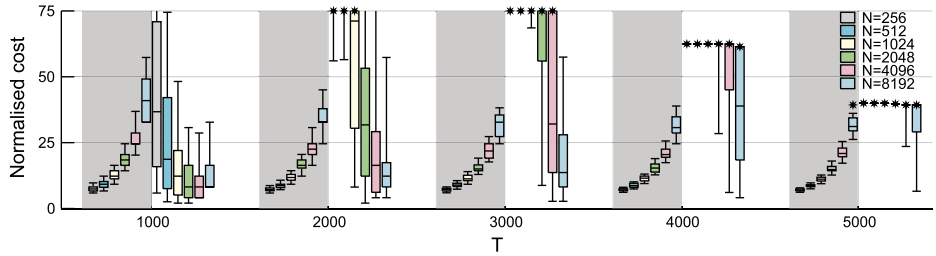
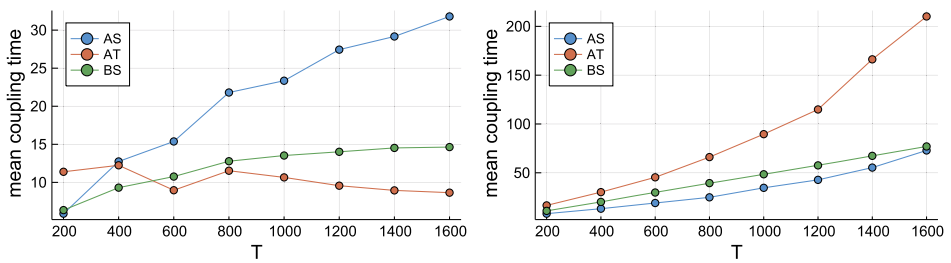


FIG. 3. *Normalised cost of coupling for the simple homogeneous model for* 100 *replications of BS and AS, the former on shaded background. The stars indicate cases where at least one replication failed to couple before hitting the maximum total cost. Since the normalised cost of coupling is plotted, the actual differences in computational cost are obtained by scaling by* T^2.

Finally, we compare the total cost of coupling in the simple homogeneous model with $s = 5$, with $N \in \{2^8, 2^9, 2^{10}, 2^{11}, 2^{12}, 2^{13}\}$ and $T \in \{1000, 2000, 3000, 4000, 5000\}$. Figure 3 shows normalised cost of coupling, defined as $N\tau/T$, over 100 replications of both algorithms. The memory consumption limited the highest number of particles to $N = 8192$ with $T = 5000$, which already exceeded 4 gigabytes in our implementation. With the shortest time horizon $T = 1000$, the AS was competitive with BS, reaching sometimes lower costs than BS. With increasing T, the AS failed to couple increasingly often before reaching the maximum number of iterations $\lfloor 10^9/(NT) \rfloor$, chosen so that the maximum time spent on a replication was approximately 2 minutes. The BS only failed to couple 3 times (out of 100) before this maximum number of iterations was reached with $N = 8192$ and $T = 5000$, and remained effective with small N. This experiment suggests that AS requires N to increase with T in order to stay effective, leading to a superlinear memory requirement that may limit its application with longer time horizons.

To provide finer detail, we report in Figure 4 the results of simulations testing the scaling properties AT, AS and BS for both the linear Gaussian and simple homogeneous model with $s = 5$. For the linear Gaussian model, the mean coupling time for AT appears to be stable with N proportional to T, while for both AS and BS the mean coupling time grows roughly linearly with T for N fixed. In contrast, for the simple homogeneous model the mean coupling time appears to grow superlinearly for AT even with N proportional to T, linearly for AS with N proportional to T, but linearly for BS with N fixed. This suggests that taking N proportional to T is not sufficient in general to stabilize AT even for relatively simple models. Similarly, for the homogeneous model, the relative cost of AS over BS grows with T and is already around 8 for $T = 1600$.

Our empirical results suggest that all of AT, AS and BS are competitive when T is small, and one should take N proportional to T for AT and sometimes also for AS. Since the space



(a) Linear Gaussian model. (b) Simple homogeneous model.

FIG. 4. *Mean coupling times associated with ancestor tracing* (*AT*), *ancestor sampling* (*AS*) *and backward sampling* (*BS*). *For the linear Gaussian model,* $N = 2.56T$ *for AT whereas* $N = 512$ *for AS and BS. For the simple homogeneous model,* $N = 0.64T$ *for AT and AS, whereas* $N = 128$ *for BS.*

complexity of any algorithm when taking N proportional to T is quadratic in T, this approach does not scale to large values of T. When it is no longer possible due to memory requirements to take N proportional to T, BS appears to be the only appropriate algorithm. We are not able to quantify accurately the minimal number of particles N_0 required for BS to exhibit linear-in-time convergence or the value of N that maximises its computational efficiency, so this needs to be done empirically.

APPENDIX A: ONE-SHOT COUPLING PROBABILITY OF CCPF

LEMMA 17. *Suppose $Y^{(1:n)}$ are nonnegative random numbers, $Z^{(1:n)}$ are $\mathcal{Z}$-valued random variables, $f : \mathcal{Z} \to [0, b]$ is measurable and $\mathcal{G}$ is a σ-algebra. If $Y^{(i)}$ are $\sigma(\mathcal{G}, Z^{(i)})$ measurable and $Z^{(1:n)}$ are conditionally independent given $\mathcal{G}$, then for any $\mathcal{G}$-measurable $A \geq 0$,*

$$\mathbb{E}\Big[\frac{\sum_{i=1}^n Y^{(i)}}{A + \sum_{j=1}^n f(Z^{(j)})} \,\Big|\, \mathcal{G}\Big] \geq \frac{\sum_{i=1}^n \mathbb{E}[Y^{(i)} \mid \mathcal{G}]}{A + b + \sum_{j=1}^n \mathbb{E}[f(Z^{(j)}) \mid \mathcal{G}]}.$$

PROOF. The claim is trivial whenever $\mathbb{P}(A + \sum_{j=1}^n f(Z^{(j)}) = 0 \mid \mathcal{G}) > 0$, so consider the case $A + \sum_{j=1}^n f(Z^{(j)}) > 0$. Because $x \mapsto x^{-1}$ is convex on $(0, \infty)$,

$$\begin{aligned}\mathbb{E}\Big[\frac{Y^{(i)}}{A + \sum_j f(Z^{(j)})} \,\Big|\, \mathcal{G}\Big] &\geq \mathbb{E}\Big[\frac{Y^{(i)}}{A + Z^{(i)} + \sum_{j \neq i} \mathbb{E}[f(Z^{(j)}) \mid \mathcal{G}, Z^{(i)}]} \,\Big|\, \mathcal{G}\Big] \\ &\geq \frac{\mathbb{E}[Y^{(i)} \mid \mathcal{G}]}{A + b + \sum_j \mathbb{E}[f(Z^{(j)}) \mid \mathcal{G}]},\end{aligned}$$

whence the result follows. □

LEMMA 18. *Consider an augmented state space $\bar{\mathsf{X}} = \mathsf{X} \cup \{\phi\}$, and define*:

- $\bar{G}_t(x) := G_t(x)$ *and* $\bar{G}_t(\phi) := \sup_x G_t(x)$ *for all* $t = 1{:}T$ *and* $x \in \mathsf{X}$,
- $\bar{M}_t(x, A) = M_t(x, A)$ *for all* $t = 1{:}T$, $x \in \mathsf{X}$ *and measurable* $A \subset \mathsf{X}$,
- $\bar{M}_t(\phi, \{\phi\}) = 1$ *for* $t = 2{:}T$.

Let $\overline{\mathrm{CCPF}}$ *and* CCPF *stand for the CCPF for models* $(\bar{\mathsf{X}}, \bar{M}_{1:T}, \bar{G}_{1:T})$ *and* $(\mathsf{X}, M_{1:T}, G_{1:T})$, *respectively. Then, for all* $s, \tilde{s} \in \mathsf{X}^T$,

(i) $\overline{\mathrm{CCPF}}(s, \tilde{s}, N) \stackrel{d}{=} \mathrm{CCPF}(s, \tilde{s}, N)$.

Let $C_{1:T}$ *stand for the sets generated in by* $\overline{\mathrm{CCPF}}(s, \tilde{s}, N)$ *and* $C_{1:T}^{\phi}$ *stand for those generated in* $\overline{\mathrm{CCPF}}(s, (\phi, \ldots, \phi), N)$.

(ii) *There exists a coupling such that* $C_t \supset C_t^{\phi}$ *a.s. for all* $t = 1{:}T$.

(iii) $C_t^{\phi} = \{i \in \{2{:}N\} : \tilde{X}_t^{(i)} \neq \phi\}$.

PROOF. The marginal equivalence (i) is straightforward. For the stochastic minorisation (ii), we consider running $\overline{\mathrm{CCPF}}(s, \tilde{s}, N)$ and $\overline{\mathrm{CCPF}}(s, (\phi, \ldots, \phi), N)$ simultaneously, in a coupled manner. More specifically, set

$$X_1^{(2:N)} = \tilde{X}_1^{(2:N)} = X_1^{\phi(2:N)} = \tilde{X}_1^{\phi(2:N)} \sim M_1(\cdot).$$

For $t \geq 2$, we proceed inductively, assuming that $X_{t-1}^{(i)} = \tilde{X}_{t-1}^{(i)} = X_{t-1}^{\phi(i)} = \tilde{X}_{t-1}^{\phi(i)}$ for all $i \in C_{t-1}^{\phi}$, and that $C_{t-1} \supset C_{t-1}^{\phi}$, which obviously hold for $t = 2$. Note that then

$$\begin{aligned}\omega_{t-1}^{(i)} = \tilde{\omega}_{t-1}^{(i)} = \omega_{t-1}^{\phi(i)} = \tilde{\omega}_{t-1}^{\phi(i)}, \quad i \in C_{t-1}^{\phi}, \\ \omega_{t-1}^{(i)} \vee \tilde{\omega}_{t-1}^{(i)} \leq \tilde{\omega}_{t-1}^{\phi(i)}, \quad i \notin C_{t-1}^{\phi}.\end{aligned}$$

Also, $\omega_{t-1}^{\phi(i)} \le \tilde{\omega}_{t-1}^{\phi(i)}$ for $i \notin C_{t-1}^{\phi}$, so we conclude that

$$\frac{\omega_{t-1}^{(i)}}{\sum_j \omega_{t-1}^{(j)}} \wedge \frac{\tilde{\omega}_{t-1}^{(i)}}{\sum_j \tilde{\omega}_{t-1}^{(j)}} \ge \frac{\tilde{\omega}_{t-1}^{\phi(i)}}{\sum_j \tilde{\omega}_{t-1}^{\phi(j)}} = \frac{\omega_{t-1}^{\phi(i)}}{\sum_j \omega_{t-1}^{\phi(j)}} \wedge \frac{\tilde{\omega}_{t-1}^{\phi(i)}}{\sum_j \tilde{\omega}_{t-1}^{\phi(j)}}, \quad i \in C_{t-1}^{\phi}.$$

Consequently, the outputs of CRES satisfy $\mathbb{P}(I_t^{(i)} = \tilde{I}_t^{(i)} \in C_{t-1}^{\phi}) \ge \mathbb{P}(I_t^{\phi(i)} = \tilde{I}_t^{\phi(i)} \in C_{t-1}^{\phi})$, and we may couple the outputs such that

$$\mathbb{P}\big(I_t^{(i)} = \tilde{I}_t^{(i)} = I_t^{\phi(i)} = \tilde{I}_t^{\phi(i)} \in C_{t-1}^{\phi}\big) = \mathbb{P}\big(I_t^{\phi(i)} = \tilde{I}_t^{\phi(i)} \in C_{t-1}^{\phi}\big),$$

and consequently we may also couple $X_t^{(i)}$, $\tilde{X}_t^{(i)}$, $X_t^{\phi(i)}$, $\tilde{X}_t^{\phi(i)}$ such that

$$X_t^{(i)} = \tilde{X}_t^{(i)} = X_t^{\phi(i)} = \tilde{X}_t^{\phi(i)} \sim M_t\big(X_{t-1}^{(I_t^{(i)})}, \cdot\big), \quad i \in C_t^{\phi}. \qquad \square$$

PROOF OF THEOREM 8. Consider $\overline{\mathrm{CCPF}}(s, (\phi, \dots, \phi), N)$, let $\check{C}_t := \{i \in \{2{:}N\} : \tilde{X}_t^{(i)} \ne \phi\}$, $\xi_t = \sum_{i=1}^N \delta_{\tilde{X}_t^{(i)}}$, $\xi_{\check{C}_t} = \sum_{i \in \check{C}_t} \delta_{\tilde{X}_t^{(i)}}$, then by Lemma 18

$$\mathbb{P}\big(X_{1:T}^{(J_{1:T})} = \tilde{X}_{1:T}^{(\tilde{J}_{1:T})}\big) = \mathbb{P}(J_T = \tilde{J}_T \in C_T) \ge \mathbb{E}\bigg[\frac{\xi_{\check{C}_T}(G_T)}{\xi_T(G_T)}\bigg].$$

Note that the latter quantity does not depend on $X_t^{(i)}$, but only on the marginal conditional particle filter $\tilde{X}_t^{(i)}$ with reference $(\phi, \dots, \phi)$. Setting $h_T^{(1)} := h_T^{(2)} := G_T$, we may apply Lemma 17 with $Z^{(i)} = \tilde{X}_T^{(i+1)}$ and $\mathcal{G} = \mathcal{G}_{T-1}$ where $\mathcal{G}_t = \sigma(\tilde{X}_u^{(i)} : u \le t, i = 2{:}N)$,

$$\begin{aligned}
\mathbb{E}\bigg[\frac{\xi_{\check{C}_T}(G_T)}{\xi_T(G_T)}\bigg] &= \mathbb{E}\bigg[\frac{\xi_{\check{C}_T}(h_T^{(1)})}{\xi_T(h_T^{(2)})}\bigg] \\
&\ge \mathbb{E}\bigg[\frac{\sum_{i=2}^N \mathbb{E}[\mathbb{I}\{\tilde{X}_T^{(i)} \in \check{C}_T\} h_T^{(1)}(\tilde{X}_T^{(i)}) \mid \mathcal{G}_{T-1}]}{2\|h_T^{(2)}\|_\infty + \sum_{j=2}^N \mathbb{E}[h_T^{(2)}(\tilde{X}_T^{(j)}) \mid \mathcal{G}_{T-1}]}\bigg] \\
&= \mathbb{E}\bigg[\frac{(N-1)\frac{\xi_{\check{C}_{T-1}}(G_{T-1} M_T h_T^{(1)})}{\xi_{T-1}(G_{T-1})}}{2\|h_T^{(2)}\|_\infty + (N-1)\frac{\xi_{T-1}(G_{T-1} M_T h_T^{(2)})}{\xi_{T-1}(G_{T-1})}}\bigg] \\
&= \mathbb{E}\bigg[\frac{\xi_{\check{C}_{T-1}}(h_{T-1}^{(1)})}{\xi_{T-1}(h_{T-1}^{(2)})}\bigg],
\end{aligned}$$

where $h_t^{(1)} := G_t M_{t+1} h_{t+1}^{(1)}$ and $h_t^{(2)} := G_t(2(N-1)^{-1}\|h_{t+1}^{(2)}\|_\infty + M_{t+1} h_{t+1}^{(2)})$. We have $h_t^{(1)} \le h_t^{(2)}$, so we may iterate similarly as above to obtain

$$\mathbb{E}\bigg[\frac{\xi_{\check{C}_T}(G_T)}{\xi_T(G_T)}\bigg] \ge \mathbb{E}\bigg[\frac{\xi_{\check{C}_1}(h_1^{(1)})}{\xi_1(h_1^{(2)})}\bigg] \ge \frac{h_0^{(1)}}{h_0^{(2)}},$$

by Lemma 17, where $h_0^{(1)}$, $h_0^{(2)}$ are defined as above, with convention $G_0 \equiv 1$.

Denoting $Q_t := G_t M_{t+1}$, $\bar{Q}_{t,u} := Q_t \cdots Q_u$ for $t \le u$ and $\bar{Q}_{t,t} = I$, we have $h_0^{(1)} = M_1 \bar{Q}_{1,T-1}(G_T)$, and

$$h_0^{(2)} \le 2(N-1)^{-1}\big\|h_1^{(2)}\big\|_\infty + \big\|M_1 h_1^{(2)}\big\|_\infty.$$

We may bound

$$\|h_t^{(2)}\|_\infty \le G^* \|h_{t+1}^{(2)}\|_\infty (1 + 2(N-1)^{-1}),$$
$$\|M_1 \bar{Q}_{1,t-1} h_t^{(2)}\|_\infty \le 2(N-1)^{-1} (G^*)^t \|h_{t+1}^{(2)}\|_\infty + \|M_1 \bar{Q}_{1,t} h_{t+1}^{(2)}\|_\infty,$$

and conclude that

$$h_0^{(2)} \le h_0^{(1)} + c_1 N^{-1},$$

for some $c_1 = c_1(G^*, T) \in (0, \infty)$. We conclude the claim with $c = c_1 / h_0^{(1)}$. □

APPENDIX B: ONE-SHOT COUPLING WITH RATE

Our results below hold under the following, slightly more general strong mixing assumption.

ASSUMPTION 19. For any $t = 1{:}T$, define $Q_t(x_{t:t+1}) := G_t(x_t) M_{t+1}(x_t, x_{t+1})$. There exists a constant $c_* < \infty$ such that for all $1 \le u \le t \le T$

$$\frac{\sup_{x_u} \int Q_u(x_{u:u+1}) \cdots Q_{t-1}(x_{t-1:t}) G_t(x_t) \, \mathrm{d}x_{u+1:t}}{\inf_{x_{u-1}} \int M_u(x_{u-1:u}) Q_u(x_{u:u+1}) \cdots Q_{t-1}(x_{t-1:t}) G_t(x_t) \, \mathrm{d}x_{u:t}} \le c_*.$$

LEMMA 20. *Suppose $G_t(x_{t-1}, x_t) = G_t(x_t)$ for all $t \ge 2$. Then Assumption 2 implies Assumption 19 with $c_* = \epsilon^{-1}$.*

PROOF. Suppose Assumption 2 holds. For any nonnegative, bounded test function $h : \mathsf{X} \to \mathbb{R}$, let $m(h) := \int h(y) \, \mathrm{d}y < \infty$. Assumption 2 implies that for any $x \in \mathsf{X}$ and $t \ge 2$, $M_*(t) m(h) \le \int M_t(x, y) h(y) \, \mathrm{d}y \le M^*(t) m(h)$. Let $1 \le u \le t$, and define

$$\phi_{u+1,t}(x_{u+1}) := \int Q_{u+1}(x_{u+1:u+2}) \cdots Q_{t-1}(x_{t-1:t}) G_t(x_t) \, \mathrm{d}x_{u+2:t},$$

which is nonnegative and bounded both from above and away from zero. We may calculate

$$\begin{aligned}
&\sup_{x'_u, x'_{u-1}} \frac{\int G_u(x'_u) M_{u+1}(x'_u, x_{u+1}) Q_{u+1}(x_{u+1:u+2}) \cdots Q_{t-1}(x_{t-1:t}) G_t(x_t) \, \mathrm{d}x_{u+1:t}}{\int M_u(x'_{u-1}, x_u) Q_u(x_{u:u+1}) \cdots Q_{t-1}(x_{t-1:t}) G_t(x_t) \, \mathrm{d}x_{u:t}} \\
&\quad \le \sup_{x'_u, x'_{u-1}} \frac{G_u(x'_u) M^*(u+1) m(\phi_{u+1,t})}{(\int M_u(x'_{u-1}, x_u) G_u(x_u) \, \mathrm{d}x_u) M_*(u+1) m(\phi_{u+1,t})} \\
&\quad \le \frac{G^*(u) M^*(u+1)}{G_*(u) M_*(u+1)}.
\end{aligned}$$

□

Consider CCPF in Algorithm 2, and denote $\xi_t = \sum_{i=1}^N \delta_{X_t^{(i)}}$, $\xi_{C_t} = \sum_{i \in C_t} \delta_{X_t^{(i)}}$, $\tilde{\xi}_t = \sum_{i=1}^N \delta_{\tilde{X}_t^{(i)}}$, $\tilde{\xi}_{C_t} = \sum_{i \in C_t} \delta_{\tilde{X}_t^{(i)}}$, $\rho_t := \xi_{C_t}(G_t)/\xi_t(G_t)$ and $\tilde{\rho}_t := \tilde{\xi}_{C_t}(G_t)/\tilde{\xi}_t(G_t)$.

LEMMA 21. *Let $h_2 \ge h_1 \ge 0$ be functions such that $h_2^* := \sup_x h_2(x) < \infty$, then for $t = 1{:}(T-1)$,*

$$\mathbb{E}\left[\frac{\xi_{C_{t+1}}(h_1)}{\xi_{t+1}(h_2)}\right] \ge \mathbb{E}\left[\frac{\xi_{C_t}(h'_1)}{\xi_t(h'_2)}\right] + \mathbb{E}[\tilde{\rho}_t] - 1,$$

where

$$h'_1(x) = G_t(x)(M_{t+1} h_1)(x),$$
$$h'_2(x) = G_t(x)\big[2(N-1)^{-1} h_2^* + (M_{t+1} h_2)(x)\big].$$

PROOF. It is direct to check that $I_{t+1}^{(j)} \in C_t$ and $\tilde{I}_{t+1}^{(j)} \in C_t$ implies $I_{t+1}^{(j)} = \tilde{I}_{t+1}^{(j)}$, because either $\omega_t^{(j)} \leq \tilde{\omega}_t^{(j)}$ for all $j \in C_t$ or $\omega_t^{(j)} \geq \tilde{\omega}_t^{(j)}$ for all $j \in C_t$. Therefore, we may write

$$\mathbb{E}\left[\frac{\xi_{C_{t+1}}(h_1)}{\xi_{t+1}(h_2)}\right] = \mathbb{E}\left[\frac{\sum_{i=2}^N h_1(X_{t+1}^{(i)}) 1((I_{t+1}^{(i)}, \tilde{I}_{t+1}^{(i)}) \in C_t^2)}{h_2(X_{t+1}^{(1)}) + \sum_{j=2}^N h_2(X_{t+1}^{(j)})}\right],$$

and apply Lemma 17 with $\mathcal{G} = \mathcal{G}_t := \{X_{1:t}^{(1:N)}, I_{1:t}^{(1:N)}, \tilde{X}_{1:t}^{(1:N)}, \tilde{I}_{1:t}^{(1:N)}\}$, $Y^{(i)} = h_1(X_{t+1}^{(i)}) \times 1((I_{t+1}^{(i)}, \tilde{I}_{t+1}^{(i)}) \in C_t^2)$, $Z^{(i)} = (X_{t+1}^{(i)}, \tilde{X}_{t+1}^{(i)}, I_{t+1}^{(i)}, \tilde{I}_{t+1}^{(i)})$ and $f(x, \tilde{x}, i, \tilde{i}) = h_2(x)$, yielding

$$\begin{aligned}
\mathbb{E}\left[\frac{\xi_{C_{t+1}}(h_1)}{\xi_{t+1}(h_2)}\right] &\geq \mathbb{E}\left[\frac{\sum_{i=2}^N \mathbb{E}[h_1(X_{t+1}^{(i)}) 1((I_{t+1}^{(i)}, \tilde{I}_{t+1}^{(i)}) \in C_t^2) \mid \mathcal{G}_t]}{2h_2^* + \sum_{j=2}^N \mathbb{E}[h_2(X_{t+1}^{(j)}) \mid \mathcal{G}_t]}\right] \\
&= \mathbb{E}\left[\frac{\sum_{i=2}^N \frac{\xi_{C_t}(G_t(M_{t+1}h_1))}{\xi_{C_t}(G_t)} \mathbb{E}[1((I_{t+1}^{(i)}, \tilde{I}_{t+1}^{(i)}) \in C_t^2) \mid \mathcal{G}_t]}{2h_2^* + (N-1)\frac{\xi_t(G_t(M_{t+1}h_2))}{\xi_t(G_t)}}\right] \\
&\geq \mathbb{E}\left[\frac{(N-1)\frac{\xi_{C_t}(G_t(M_{t+1}h_1))}{\xi_{C_t}(G_t)} \rho_t \tilde{\rho}_t}{2h_2^* + (N-1)\frac{\xi_t(G_t(M_{t+1}h_2))}{\xi_t(G_t)}}\right] \\
&= \mathbb{E}\left[\frac{\xi_{C_t}(G_t(M_{t+1}h_1))}{2h_2^*(N-1)^{-1}\xi_t(G_t) + \xi_t(G_t(M_{t+1}h_2))} \tilde{\rho}_t\right] \\
&= \mathbb{E}\left[\frac{\xi_{C_t}(h_1')}{\xi_t(h_2')} - (1 - \tilde{\rho}_t)\frac{\xi_{C_t}(h_1')}{\xi_t(h_2')}\right],
\end{aligned}$$

from which the claim follows because $\tilde{\rho}_t \in [0, 1]$ and $h_1' \leq h_2'$, so the latter fraction is upper bounded by one. □

LEMMA 22. *Suppose that Assumption 19 holds, then for any* $t = 1{:}T - 1$,

$$\mathbb{E}[\rho_t] \geq \beta_N^t + \sum_{u=1}^{t-1} (\mathbb{E}[\tilde{\rho}_u] - 1) \quad \textit{where } \beta_N := \left(1 + \frac{2c_*}{N-1}\right)^{-1}.$$

PROOF. We may apply Lemma 21 recursively with $h_1^{(u)} = h_2^{(u)} = \bar{G}_{u,t}$ where $\bar{G}_{u,t}(x_u) := \int Q_u(x_{u:u+1}) \bar{G}_{u+1,t}(x_{u+1}) \, dx_{u+1}$ and $\bar{G}_{t,t} = G_t$, leading to

$$\begin{aligned}
h_2'^{(u)}(x) &= G_{u-1}(x)\left[2(N-1)^{-1} \sup_{x'} \bar{G}_{u,t}(x') + (M_u \bar{G}_{u,t})(x)\right] \\
&\leq G_{u-1}(x)(M_u \bar{G}_{u,t})(x)\left(1 + \frac{2c_*}{N-1}\right) \\
&= \bar{G}_{u-1,t} \beta_N^{-1},
\end{aligned}$$

implying that

$$\mathbb{E}\left[\frac{\xi_{C_{u+1}}(\bar{G}_{u+1,t})}{\xi_{u+1}(\bar{G}_{u+1,t})}\right] \geq \mathbb{E}\left[\frac{\xi_{C_u}(\bar{G}_{u,t})}{\xi_u(\bar{G}_{u,t})}\right] \beta_N + (\mathbb{E}[\tilde{\rho}_u] - 1).$$

□

LEMMA 23. *Under Assumption 19,*

$$\mathbb{E}[\rho_T] \geq 1 - 2^T(1 - \beta_N^T) \geq 1 - \frac{2^T T}{(2c_*)^{-1}(N-1) + 1}.$$

PROOF. The first inequality follows once we prove inductively that $(1-\mathbb{E}[\rho_t]) \vee (1-\mathbb{E}[\tilde{\rho}_t]) \leq 2^t(1-\beta_N^t)$, which holds for $t=1$ by Lemma 22 (which is symmetric wrt. ρ_t and $\tilde{\rho}_t$). Then

$$\begin{aligned} 1-\mathbb{E}[\rho_t] &\leq 1-\beta_N^t + \sum_{u=1}^{t-1}(1-\mathbb{E}[\tilde{\rho}_t]) \\ &\leq (1-\beta_N^t)\left(1+\sum_{u=1}^{t-1}2^u\right) \leq 2^t(1-\beta_N^t), \end{aligned}$$

and the same bound applies to $1-\mathbb{E}[\tilde{\rho}_t]$. The latter bound follows as $1-\beta_N^T \leq T(1-\beta_N)$. □

APPENDIX C: UNBIASED ESTIMATOR BASED ON A COUPLED MARKOV KERNEL

We formalise here the construction of unbiased estimators of Markov chain equilibrium expectations due to Glynn and Rhee (2014), and complement the results in Jacob, Lindsten and Schön (2020).

DEFINITION 24 (Coupling of probability measures). Suppose μ and ν are two probability measures on S. The set of couplings $\Gamma(\mu,\nu)$ consists of all probability measures λ on $\mathsf{S}\times\mathsf{S}$ with marginals $\lambda(\cdot\times\mathsf{S})=\mu$ and $\lambda(\mathsf{S}\times\cdot)=\nu$.

DEFINITION 25 (Coupled Markov kernel). Suppose P is a Markov kernel on S, and $\boldsymbol{P}$ is a Markov kernel on $\mathsf{S}\times\mathsf{S}$. If $\boldsymbol{P}(x,\tilde{x};\cdot)\in\Gamma(P(x,\cdot),P(\tilde{x},\cdot))$ for all $x,\tilde{x}\in\mathsf{X}$, then $\boldsymbol{P}$ is a *coupled kernel* corresponding to P.

DEFINITION 26 (Coupling time). The *coupling time* of the bivariate Markov chain $(X_n,\tilde{X}_n)_{n\geq 0}$ is the random variable $\tau := \inf\{n\geq 0 : X_k=\tilde{X}_k \text{ for all } k\geq n\}$.

THEOREM 27. *Let P be an ergodic Markov kernel on X with invariant distribution π (i.e., for all $x\in\mathsf{X}$, $\|P^n(x,\cdot)-\pi\|_{\mathrm{tv}} \xrightarrow{n\to\infty} 0$), and suppose $\boldsymbol{P}$ is a corresponding coupled Markov kernel. Let ν be any probability distribution on X, and suppose that $\lambda\in\Gamma(\nu P,\nu)$.*

Consider a Markov chain $(X_n,\tilde{X}_n)_{n\geq 0}$ with initial distribution λ and transition probability $\boldsymbol{P}$, and $h\in L^2(\pi)$. Let the coupling time τ of $(X_n,\tilde{X}_n)_{n\geq 0}$ be a.s. finite, $\sup_{n\geq 0}\mathbb{E}[h^2(X_n)]<\infty$ for all $n\geq 0$ and

$$\sup_{\{m,L:L\geq m\}} \mathbb{E}\big[Z_{m,L}^2\big] < \infty \quad \text{where } Z_{m,L} := \sum_{n=m}^{L}\big[h(X_n)-h(\tilde{X}_n)\big]\mathbb{I}\{n<\tau\}. \tag{6}$$

Then for all $b\geq 0$, $\mathbb{E}[Z_b]=\pi(h)$ and $\mathrm{var}(Z_b)<\infty$, where

$$Z_b := h(X_b) + \sum_{n=b+1}^{\infty}\big[h(X_n)-h(\tilde{X}_n)\big]\mathbb{I}\{n<\tau\}.$$

PROOF. Note that $X_n \overset{d}{=} \tilde{X}_{n+1}$ for all $n\geq 0$, and so for any $m>b$,

$$\mathbb{E}\big[h(X_m)\big] = \mathbb{E}\Bigg[\overbrace{h(X_b)+\sum_{n=b+1}^{m}\big[h(X_n)-h(\tilde{X}_n)\big]\mathbb{I}\{n<\tau\}}^{=:\zeta_m}\Bigg].$$

Fix $b > 0$. By (6), $\{\zeta_m\}$ is a bounded sequence in L^2. The almost sure finiteness of the stopping time ensures Z_b is well defined and $\zeta_m \to Z_b$ almost surely. Thus Z_b is also square integrable and $\zeta_m \to Z_b$ in L^1. Since νP^n converges to π in total variation, the assumptions $\sup_n \nu P^n(h^2) < \infty$ and $\pi(h^2) < \infty$ imply $\nu(P^n h) \to \pi(h)$. Finally, $\mathbb{E}[Z_b] = \pi(h)$ follows from $\mathbb{E}[\zeta_m] = \mathbb{E}[h(X_m)] = \nu(P^{m+1}h) \to \pi(h)$. $\square$

LEMMA 28. *Letting $\bar{h} := h - \pi(h)$, the variance of the estimator satisfies*

$$\bigl|\mathrm{var}(Z_b) - \mathrm{var}_\pi\bigl(h(X)\bigr)\bigr| \le \|\bar{h}\|_\infty^2 \bigl\|\nu P^{b+1} - \pi\bigr\|_{\mathrm{tv}} + 2\|\bar{h}\|_\infty \bigl(\mathbb{E}Z_{b+1,\infty}^2\bigr)^{1/2} + \mathbb{E}Z_{b+1,\infty}^2.$$

PROOF. We may write

$$\mathrm{var}(Z_b) = \mathbb{E}\bigl(\bar{h}(X_b) + Z_{b+1,\infty}\bigr)^2 = \mathbb{E}\bar{h}^2(X_b) + 2\mathbb{E}\bigl[\bar{h}(X_b)Z_{b+1,\infty}\bigr] + \mathbb{E}Z_{b+1,\infty}^2.$$

Let $X \sim \pi$, then

$$\bigl|\mathrm{var}(Z_b) - \mathrm{var}\bigl(h(X)\bigr)\bigr| \le \bigl|\mathbb{E}\bar{h}^2(X_b) - \mathbb{E}\bar{h}^2(X)\bigr| + 2\|\bar{h}\|_\infty \mathbb{E}|Z_{b+1,\infty}| + \mathbb{E}Z_{b+1,\infty}^2.$$

The first term is upper bounded by $\|\bar{h}\|_\infty^2 \|\nu P^{b+1} - \pi\|_{\mathrm{tv}}$, and $(\mathbb{E}|Z_{b+1,\infty}|)^2 \le \mathbb{E}Z_{b+1,\infty}^2$. $\square$

Below, we use $\|h\|_{\mathrm{osc}} := \sup_{x,y\in\mathsf{X}} |h(x) - h(y)|$, and $\|h\|_\infty = \sup_{x\in\mathsf{X}} |h(x)| \ge \|h\|_{\mathrm{osc}}/2$.

Under the following assumed distribution on the coupling time, not only is the sequence $Z_{m,L}$ defined in (6) uniformly square integrable, the corresponding sequence $\{\zeta_m\}$ is a L^2 Cauchy sequence.

LEMMA 29. *Suppose that there exist $C < \infty$ and $\lambda \in [0, 1)$ such that for all $n \in \mathbb{N}$, $\mathbb{P}(\tau > n) \le C\lambda^n$, then $\mathbb{E}[Z_{m,L}^2] \le 2C\lambda^m(1-\lambda)^{-2}\|h\|_{\mathrm{osc}}^2$ for all $L \ge m \ge 1$.*

PROOF. Let $\Delta h_n := h(X_n) - h(\tilde{X}_n)$, then $|\Delta h_n| \le \|h\|_{\mathrm{osc}}$, and so

$$\mathbb{E}\bigl[Z_{m,L}^2\bigr] = \mathbb{E}\Biggl[\sum_{n,\ell=m}^{L} \Delta h_n \Delta h_\ell \mathbb{I}\{\tau > n \vee \ell\}\Biggr] \le \|h\|_{\mathrm{osc}}^2 \sum_{n,\ell=m}^{L} \mathbb{P}(\tau > n \vee \ell). \tag{7}$$

The latter sum may be upper bounded by

$$C\sum_{n,\ell=m}^{L} \lambda^{n\vee\ell} \le C\sum_{i=0}^{\infty}(2i+1)\lambda^{i+m} = C\lambda^m\biggl(2\frac{\lambda}{(1-\lambda)^2} + \frac{1}{(1-\lambda)}\biggr).$$

Simple calculation yields the desired bound. $\square$

LEMMA 30. *Suppose $\mathbf{P}$ is a coupled kernel corresponding to a π-ergodic Markov kernel P. Let $\tau_{x,\tilde{x}}$ stand for the coupling time of the Markov chain $(X_n, \tilde{X}_n)_{n\ge0}$ with transition probability $\mathbf{P}$ and with $(X_0, \tilde{X}_0) \equiv (x, \tilde{x})$. Then*

$$\bigl\|P(x,\cdot) - \pi\bigr\|_{\mathrm{tv}} \le 2\sup_{\tilde{x}\in\mathsf{X}} \mathbb{P}(\tau_{x,\tilde{x}} > n).$$

PROOF. Let τ_x stand for the coupling time of $(X_n, \tilde{X}_n)_{n\ge0}$ with $X_0 \equiv x$ and $\tilde{X}_0 \sim \pi$. By the standard coupling inequality, $\|P(x,\cdot) - \pi\|_{\mathrm{tv}} \le 2\mathbb{P}(\tau_x > n)$, and

$$\mathbb{P}(\tau_x > n) = \int \mathbb{P}(\tau_x > n \mid \tilde{X}_0 = \tilde{x})\pi(\mathrm{d}\tilde{x}) = \int \mathbb{P}(\tau_{x,\tilde{x}} > n)\pi(\mathrm{d}\tilde{x}).$$

$\square$

LEMMA 31. *Let $\tau_{x,\tilde{x}}$ be as in Lemma 30, and assume that there exist $C \in [1, \infty)$ and $\lambda \in (0, 1)$ such that $\sup_{\tilde{x},x\in\mathsf{X}} \mathbb{P}(\tau_{x,\tilde{x}} > n) \leq C\lambda^n$. Then*

$$\left|\mathrm{var}(Z_b) - \mathrm{var}_\pi\big(h(X)\big)\right| \leq \frac{16C}{(1-\lambda)^2}\lambda^{(b+1)/2}\|\bar{h}\|_\infty^2.$$

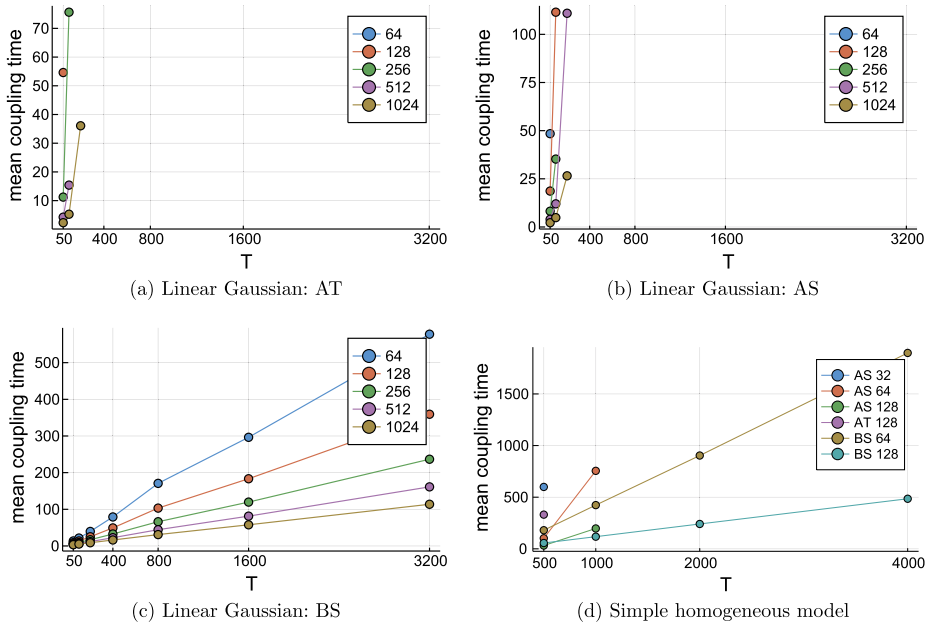
PROOF. Using Lemma 28 together with Lemma 30 and Lemma 29 yields

$$\left|\mathrm{var}(Z_b) - \mathrm{var}_\pi\big(h(X)\big)\right| \leq 2C\|\bar{h}\|_\infty^2\lambda^{b+1} + 2\|\bar{h}\|_\infty\left(\frac{2C\lambda^{b+1}}{(1-\lambda)^2}\|h\|_{\mathrm{osc}}^2\right)^{\frac{1}{2}} + \frac{2C\lambda^{b+1}}{(1-\lambda)^2}\|h\|_{\mathrm{osc}}^2.$$

The claim follows easily, because $\|h\|_{\mathrm{osc}} \leq 2\|\bar{h}\|_\infty$. □

APPENDIX D: SUPPLEMENTARY SIMULATION RESULTS

Figure 5 corresponds to Figure 1 but where common random numbers are not used in Line 7 of Algorithm 2. We observe that AT and AS show much worse performance, whereas BS is hardly affected.



(a) Linear Gaussian: AT (b) Linear Gaussian: AS

(c) Linear Gaussian: BS (d) Simple homogeneous model

FIG. 5. *Mean coupling times associated with ancestor tracing (AT), ancestor sampling (AS) and backward sampling (BS). For (d), the lines are coloured according to the type of algorithm and the number of particles N.*

Acknowledgments. The authors would like to thank the Isaac Newton Institute for Mathematical Sciences and the Institute for Mathematical Sciences at the National University of Singapore for support and hospitality during the programmes "Scalable inference; statistical, algorithmic, computational aspects" and "Bayesian Computation for High-Dimensional Statistical Models," respectively, when work on this paper was undertaken. This work was supported by EPSRC grant numbers EP/K032208/1, EP/R014604/1 and EP/R034710/1, and

by the Alan Turing Institute under the EPSRC grant EP/N510129/1. MV was supported by Academy of Finland grants 274740, 284513, 312605 and 315619. The authors wish to acknowledge CSC, IT Center for Science, Finland, for computational resources.

REFERENCES

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 https://doi.org/10.1111/j.1467-9868.2009.00736.x

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 https://doi.org/10.3150/15-BEJ785

CHOPIN, N. and SINGH, S. S. (2015). On particle Gibbs sampling. *Bernoulli* **21** 1855–1883. MR3352064 https://doi.org/10.3150/14-BEJ629

DEL MORAL, P. (2004). *Feynman–Kac Formulae*: *Genealogical and Interacting Particle Systems with Applications*. *Probability and Its Applications* (*New York*). Springer, New York. MR2044973 https://doi.org/10.1007/978-1-4684-9393-1

DEL MORAL, P. and GUIONNET, A. (2001). On the stability of interacting processes with applications to filtering and genetic algorithms. *Ann. Inst. Henri Poincaré Probab. Stat.* **37** 155–194. MR1819122 https://doi.org/10.1016/S0246-0203(00)01064-5

DELYON, B., LAVIELLE, M. and MOULINES, E. (1999). Convergence of a stochastic approximation version of the EM algorithm. *Ann. Statist.* **27** 94–128. MR1701103 https://doi.org/10.1214/aos/1018031103

DOUC, R., GARIVIER, A., MOULINES, E. and OLSSON, J. (2011). Sequential Monte Carlo smoothing for general state space hidden Markov models. *Ann. Appl. Probab.* **21** 2109–2145. MR2895411 https://doi.org/10.1214/10-AAP735

FEARNHEAD, P. and KÜNSCH, H. R. (2018). Particle filters and data assimilation. *Annu. Rev. Stat. Appl.* **5** 421–452. MR3774754 https://doi.org/10.1146/annurev-statistics-031017-100232

GLYNN, P. W. and RHEE, C.-H. (2014). Exact estimation for Markov chain equilibrium expectations. *J. Appl. Probab.* **51A** 377–389. MR3317370 https://doi.org/10.1239/jap/1417528487

JACOB, P. E., LINDSTEN, F. and SCHÖN, T. B. Smoothing with couplings of conditional particle filters. *J. Amer. Statist. Assoc.* To appear.

JACOB, P. E., O'LEARY, J. and ATCHADÉ, Y. F. Unbiased Markov chain Monte Carlo with couplings. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* To appear.

LINDSTEN, F., DOUC, R. and MOULINES, E. (2015). Uniform ergodicity of the particle Gibbs sampler. *Scand. J. Stat.* **42** 775–797. MR3391692 https://doi.org/10.1111/sjos.12136

LINDSTEN, F., JORDAN, M. I. and SCHÖN, T. B. (2014). Particle Gibbs with ancestor sampling. *J. Mach. Learn. Res.* **15** 2145–2184. MR3231604

SHAKED, M. and SHANTHIKUMAR, J. G. (2007). *Stochastic Orders*. *Springer Series in Statistics*. Springer, New York. MR2265633 https://doi.org/10.1007/978-0-387-34675-5

SINGH, S. S., LINDSTEN, F. and MOULINES, E. (2017). Blocking strategies and stability of particle Gibbs samplers. *Biometrika* **104** 953–969. MR3737314 https://doi.org/10.1093/biomet/asx051

WHITELEY, N. (2010). Discussion on Particle Markov chain Monte Carlo methods. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **72** 306–307.

WHITELEY, N. (2013). Stability properties of some particle filters. *Ann. Appl. Probab.* **23** 2500–2537. MR3127943 https://doi.org/10.1214/12-AAP909