# Approximate Smoothing and Parameter Estimation in High-Dimensional State-Space Models

Axel Finke, Sumeetpal S. Singh

***Abstract*—We present approximate algorithms for performing smoothing in a class of high-dimensional state-space models via sequential Monte Carlo methods ('particle filters'). In high dimensions, a prohibitively large number of Monte Carlo samples ('particles'), growing exponentially in the dimension of the state space, is usually required to obtain a useful smoother. Employing blocking approximations, we exploit the spatial ergodicity properties of the model to circumvent this curse of dimensionality. We thus obtain approximate smoothers that can be computed recursively in time and parallel in space. First, we show that the bias of our blocked smoother is bounded uniformly in the time horizon and in the model dimension. We then approximate the blocked smoother with particles and derive the asymptotic variance of idealised versions of our blocked particle smoother to show that variance is no longer adversely effected by the dimension of the model. Finally, we employ our method to successfully perform maximum-likelihood estimation via stochastic gradient-ascent and stochastic expectation–maximisation algorithms in a 100-dimensional state-space model.**

***Index Terms*—high dimensions, smoothing, particle filter, sequential Monte Carlo, state-space model**

## I. Introduction

ALGORITHMS known as sequential Monte Carlo (SMC) methods or particle filters (PFs) are nowadays commonly used to perform inference in state-space models [1], [2] and more generally [3]. In order to infer certain 'static' model parameters, SMC methods are often employed within other algorithms, i.e. within Markov chain Monte Carlo (MCMC) or other SMC approaches for Bayesian inference [4], [5] or within stochastic gradient-ascent and expectation–maximisation (EM) algorithms for Frequentist inference [6]. In both cases, it is imperative to control the error of SMC approximations of the filter and smoother. Unfortunately, the number of Monte Carlo samples ('particles') typically needs to scale exponentially in the dimension of the state space in order to control these errors [7]. This 'curse of dimensionality' quickly leads to a prohibitive computational cost in higher dimensions.

In order to circumvent the curse of dimensionality in the context of *filtering*, so called 'blocking approximations' have been introduced in the literature. For finite state-space dynamic Bayesian networks, blocked filter approximations (termed *Boyen–Koller* algorithm) were proposed by [8], [9]. SMC approximations of the Boyen–Koller algorithm were first considered in [10]. For models on general state spaces [11], [12] used similar algorithms. [13] termed one such algorithm blocked particle filter and showed that it permits approximations of filter marginals which are bounded uniformly in time and model dimension as long as the model is sufficiently ergodic in both time and space. Other attempts at reducing the state-dimension by introducing additional intermediate SMC steps were made in [14]–[16]. However, for problems without a very specific conditional-independence structure, these strategies induce models which are no longer Markov which may cause difficulties in an SMC context. For state-space models with Gaussian-mixture transitions, the *iterated auxiliary particle filter* introduced by [17] can make Bayesian inference via particle MCMC methods [4] feasible in high dimensions.

The problem of performing *smoothing* for high-dimensional state-space models has received much less attention thus far. For *small finite* state spaces, [18] attempted a simple approximate forward–backward recursion and more sophisticated blocked smoothing algorithms are developed by [19, personal communication] independently of the present work. However, for *general* (e.g. large discrete or even continuous) state spaces, efficient smoothing algorithms in high dimensions are still lacking.

The goal of this work is thus to exploit blocking strategies to devise approximate versions of particle smoothing algorithms known as *backward sampling* [20] and *forward smoothing* [21] for the canonical class of high-dimensional (general state-space) hidden Markov models from [13] and to provide theoretical guarantees for these methods. Our methodological contributions, mainly contained in Section IV, are as follows.

- In Subsection IV-C, we show that existing smoothing algorithms break down in high dimensions in the sense that the asymptotic variance grows exponentially in the model dimension – even if we make the favourable assumption that the filter errors are dimension-independent (Proposition 1). This unequivocally justifies the need for new dimensionally stable particle smoothers.
- In Subsection IV-D, we introduce novel blocked particle smoothing algorithms and a bias-reduction technique.
- In Subsection IV-E, we prove a uniform (in both time and model dimension) bound on the asymptotic variance of our estimator (Proposition 2) and a similarly uniform bound on the asymptotic bias (Proposition 3).

In Section V, we empirically illustrate that our algorithm induces local errors which are bounded in the model dimension. Finally, we successfully perform maximum-likelihood estimation via stochastic gradient-ascent and stochastic EM algorithms in a 100-dimensional state-space model.

This work was supported by the Engineering and Physical Sciences Research Council [grant number EP/K020153/1].

A. Finke is with the Department of Statistical Science, University College London, UK (e-mail: a.finke@ucl.ac.uk)

S. S. Singh is with the Department of Engineering, University of Cambridge, UK and the Alan Turing Institute, London, UK (e-mail: sss40@cam.ac.uk)

This article has been accepted for publication in a future issue of IEEE Transactions on Signal Processing, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSP.2017.2733504

## II. Standard Particle Filtering and Smoothing

In this section, we review standard particle methodology for performing filtering or smoothing in state-space models.

### A. State-Space Models

A (homogeneous) state-space model is a stochastic process $(X_t, Y_t)_{t\in\mathbb{N}}$ on a space $\mathbb{X} \times \mathbb{Y}$ with the following properties. The process $(X_t)_{t\in\mathbb{N}}$ is a Markov chain on $\mathbb{X}$ with initial distribution $\mu(\mathrm{d}x_1) = m(x_1)\psi(\mathrm{d}x_1)$ and transitions $p(x_{t-1}, x_t)\psi(\mathrm{d}x_t)$. Here, $\psi$ denotes the reference measure on $\mathbb{X}$ with respect to which $m\colon \mathbb{X} \to (0,\infty)$ and $p\colon \mathbb{X}\times\mathbb{X} \to (0,\infty)$ are densities. The chain $(X_t)_{t\in\mathbb{N}}$ is not directly observed. At each time $t$, we instead observe the value $y_t$ of the random variate $Y_t$ whose law, conditional on the 'state' sequence $X_{1:t} := (X_1, \dots, X_t)$ taking the values $x_{1:t} \in \mathbb{X}^t$, is $g(x_t, y_t)\varphi(\mathrm{d}y_t)$. Here, $\varphi$ denotes the reference measure on $\mathbb{Y}$ with respect to which $g\colon \mathbb{X}\times\mathbb{Y} \to (0,\infty)$ is a transition density. Throughout this work, we will refer to $V := \operatorname{card}\mathbb{X}$ as the *model dimension*. Finally, all densities may depend on some model parameter $\theta$ but for simplicity, we suppress $\theta$ from the notation wherever possible.

*1) Filtering:* In many applications, e.g. when tracking objects in real time, we are interested in approximating integrals with respect to the *filter* at time $t$, $\pi_t$. This is the distribution of $X_t$ given all the observations recorded up to time $t$, i.e. $\pi_t(\mathrm{d}x_t) = \mathbb{P}(X_t \in \mathrm{d}x_t | Y_{1:t} = y_{1:t})$.

*2) Smoothing:* The *joint smoothing distribution* $\mathbb{Q}_t$ is the posterior distribution of the states $X_{1:t}$ given the data $y_{1:t}$, i.e. $\mathbb{Q}_t(\mathrm{d}x_{1:t}) := \mathbb{P}(X_{1:t} \in \mathrm{d}x_{1:t} | Y_{1:t} = y_{1:t})$. It is given by $\mathbb{Q}_t(\mathrm{d}x_{1:t}) \propto \mathit{\Gamma}_t(\mathrm{d}x_{1:t})$, where

$$\mathit{\Gamma}_t(\mathrm{d}x_{1:t}) = \mu(\mathrm{d}x_1)g(x_1, y_1)\prod_{s=2}^{t} p(x_{s-1}, x_s)g(x_s, y_s)\psi(\mathrm{d}x_s).$$

*3) Goal:* In this work, we seek to efficiently approximate integrals of certain additive functions $F_t\colon \mathbb{X}^t \to \mathbb{R}$ (see Subsection IV-A) w.r.t. the joint smoothing distribution, i.e.

$$\mathbb{F}_t := \mathbb{Q}_t(F_t) := \mathbb{E}[F_t(X_{1:t})], \quad \text{for } X_{1:t} \sim \mathbb{Q}_t,$$

for the canonical class of high-dimensional state-space models introduced in [13] (and detailed in Subsection III-A of this work). First, however, we review existing methods conventionally employed for this purpose. We will sometimes refer to these as *standard* particle filters and smoothers to distinguish them from the *blocked* particle filters and smoothers reviewed and introduced in Sections III and IV, respectively.

### B. Standard Particle Filtering

We begin by reviewing (standard) particle filters (PFs). Let $Q_t(z, \mathrm{d}x) := q_t(z, x)\psi(\mathrm{d}x)$ be a Markov transition kernel, where $q_t\colon \mathbb{X}\times\mathbb{X} \to (0,\infty)$ is some suitable transition density with respect to the reference measure $\psi$. All these quantities may also depend on $y_{1:t}$ but we suppress this dependence, for simplicity. Finally, define

$$G_1(x) := \frac{m(x)g(x, y_1)}{q_1(x)},$$
$$G_t(z, x) := \frac{p(z, x)g(x, y_t)}{q_t(z, x)}, \quad \text{for } t > 1.$$

Algorithm 1 summarises a PF. Here, $\mathrm{Cat}(p^{1:N})$ is the categorical distribution with some vector of probabilities $p^{1:N}$ and we use the convention that actions prescribed for the $n$th particle are to be performed *conditionally independently* for all $n \in \{1, \dots, N\}$, where $N$ is the number of particles.

*Algorithm 1 (particle filter):*

```
1) At time 1,
   i) sample $X_1^n \sim Q_1$,
   ii) set $w_1^n := G_1(X_1^n)$ and $W_1^n := w_1^n / \sum_{k=1}^N w_1^k$.
2) At Step $t$, $t > 1$,
   i) sample $A_{t-1}^n \sim \mathrm{Cat}(W_{t-1}^{1:N})$; write $\overline{X}_{t-1}^n := X_{t-1}^{A_{t-1}^n}$,
   ii) sample $X_t^n \sim Q_t(\overline{X}_{t-1}^n, \cdot)$;
   iii) set $w_t^n := G_t(\overline{X}_{t-1}^n, X_t^n)$ and $W_t^n := w_t^n / \sum_{k=1}^N w_t^k$.
```

If $q_t = p$ then Algorithm 1 is often termed *bootstrap* PF. For any test function $f\colon \mathbb{X} \to \mathbb{R}$, we may approximate $\pi_t(f)$ using the weighted sample $(X_t^n, W_t^n)_{n\leq N}$. That is, after the $t$th step, we may estimate $\pi_t(f)$ by

$$\pi_{t,\mathrm{PF}}^N(f) := \sum_{n=1}^N W_t^n f(X_t^n). \tag{1}$$

Throughout this work, we assume that the cost of sampling each particle $X_t^n$ and evaluating each weight $w_t^n$ grows linearly in $V$. The complexity of the PF is then $\mathcal{O}(VN)$ per time step.

### C. Standard Particle Smoothing

We now review algorithms which have been proposed to approximate expectations with respect to the joint smoothing distribution, i.e. integrals $\mathbb{F}_T = \mathbb{Q}_T(F_T)$, for some $T \in \mathbb{N}$.

*1) Backward Recursion:* It is well known [1, Corollary 3.3.8] that the joint smoothing distribution can be written as

$$\mathbb{Q}_T(\mathrm{d}x_{1:T}) = \pi_T(\mathrm{d}x_T)\prod_{t=1}^{T-1} B_{\pi_t}(x_{t+1}, \mathrm{d}x_t), \tag{2}$$

where the *backward kernels* $B_\nu\colon \mathbb{X} \times \mathrm{Borel}(\mathbb{X}) \to [0,1]$ ($\mathrm{Borel}(\mathbb{X})$ is the Borel $\sigma$-algebra and $\nu$ some probability measure on $\mathbb{X}$) are defined by

$$B_\nu(x, \mathrm{d}z) = \frac{p(z, x)\nu(\mathrm{d}z)}{\int_{\mathbb{X}} p(u, x)\nu(\mathrm{d}u)}.$$

*2) Particle Approximation:* Unfortunately, the filters $\pi_t$ and hence the backward kernels $B_{\pi_t}(x, \mathrm{d}z)$ in (2) are usually intractable unless the model is linear and Gaussian or unless the model dimension $V = \operatorname{card}\mathbb{X}$ is sufficiently small. To circumvent this intractability, standard particle smoothers [20], [21] replace the filters in (2) by Monte Carlo approximations. More precisely, let $(X_t^n, W_t^n)_{n\leq N}$ be a weighted sample (e.g. obtained from the PF in Algorithm 1) such that $\pi_t^N := \sum_{n=1}^N W_t^n \delta_{X_t^n}$, approximates $\pi_t$. Here, $\delta_x$ denotes the point mass at $x$. We then replace $B_{\pi_t}(x, \mathrm{d}z)$ in (2) by

$$B_{\pi_t^N}(x, \mathrm{d}z) = \sum_{n=1}^N \frac{W_t^n p(X_t^n, x)}{\sum_{k=1}^N W_t^k p(X_t^k, x)} \delta_{X_t^n}(\mathrm{d}z). \tag{3}$$

Conditional on $(X_t^n, W_t^n)_{n\in\{1,\dots,N\}}$, the computational complexity of evaluating $B_{\pi_t^N}(x, \cdot)$ is $\mathcal{O}(NV)$.

Replacing $B_{\pi_t}(x, \mathrm{d}z)$ by $B_{\pi_t^N}(x, \mathrm{d}z)$ in (2) then induces the following approximation of $\mathbb{Q}_T(F_T)$:

$$\begin{aligned}\mathbb{Q}_{T,\text{FS}}^N(F_T) &:= \int_{\mathbb{X}^T} F_T(x_{1:T})\pi_T^N(\mathrm{d}x_T)\prod_{t=1}^{T-1} B_{\pi_t^N}(x_{t+1}, \mathrm{d}x_t) \\ &= \sum_{n_{1:T}\in\{1,\dots,N\}^T} F_T(X_{1:T}^{n_{1:T}})W_T^{n_T}\prod_{t=1}^{T-1} B_{\pi_t^N}(X_{t+1}^{n_{t+1}}, \{X_t^{n_t}\}). \qquad (4)\end{aligned}$$

*3) Forward Smoothing:* The computational cost of summing over $N^T$ terms in (4) is normally prohibitive. However, if $F_T$ is *additive in time* in the sense that there are functions $f_1 \colon \mathbb{X} \to \mathbb{R}$ and $f_t \colon \mathbb{X}^2 \to \mathbb{R}$ (where $f_t$ may depend on $y_t$) such that

$$F_t(x_{1:t}) = f_1(x_1) + \sum_{s=2}^{t} f_s(x_{s-1}, x_s),$$

then the computational cost of evaluating this estimate can be brought down to $\mathcal{O}(N^2TV)$. The resulting algorithm is called (standard) *forward smoothing (FS)* and was introduced by [21], [22] who also derived finite-sample error bounds as well as a central limit theorem (see also [23]). Algorithm 2 outlines the idea; we use the convention that any action prescribed for *some* $n$ is to be performed conditionally independently for *all* $n \in \{1, \dots, N\}$. Note that Algorithm 2 can be implemented online, i.e. $\alpha_t^n$ can already be determined at Step $t$ of the PF.

*Algorithm 2 (forward smoothing):*

```
1) Set $\alpha_1^n := f_1(X_1^n)$. For $t > 1$, set
       $\alpha_t^n := \sum_{m=1}^{N} B_{\pi_{t-1}^N}(X_t^n, \{X_{t-1}^m\})[\alpha_{t-1}^m + f_t(X_{t-1}^m, X_t^n)].$
2) Approximate $\mathbb{F}_t$ by $\mathbb{F}_t^N := \mathbb{Q}_{t,\text{FS}}^N(F_t) = \sum_{n=1}^N W_t^n \alpha_t^n$.
```

*4) Backward Sampling:* To circumvent the $\mathcal{O}(N^2)$ computational complexity of FS, we may instead estimate $\mathbb{Q}_T(F_T)$ using a simple Monte Carlo approximation based on $M < N$ sample points drawn conditionally independently from $\mathbb{Q}_{T,\text{FS}}^N$. More precisely, the algorithm samples $M$ particle paths $\widetilde{X}_{1:T}^m$ in the reverse-time direction according to the kernel from (3). This gives the (standard) *backward sampling (BS)* [20] approximation[1]

$$\mathbb{Q}_{T,\text{BS}}^N(F_T) := \frac{1}{M}\sum_{m=1}^{M} F_T(\widetilde{X}_{1:T}^m).$$

Algorithm 3 outlines the method. Here, we use the convention that any action prescribed for *some* $m$ is to be performed conditionally independently for *all* $m \in \{1, \dots, M\}$.

*Algorithm 3 (backward sampling):*

```
1) Sample $\widetilde{X}_T^m \sim \pi_T^N$ and $\widetilde{X}_t^m \sim B_{\pi_t^N}(\widetilde{X}_{t+1}^m, \cdot)$, for $t < T$.
2) Approximate $\mathbb{F}_T$ by $\mathbb{F}_T^N := \mathbb{Q}_{T,\text{BS}}^N(F_T)$.
```

The computational complexity of Algorithm 3 is $\mathcal{O}(MNTV)$. However, $\mathbb{Q}_{T,\text{BS}}^N(F_T)$ normally has a larger variance than $\mathbb{Q}_{T,\text{FS}}^N(F_T)$ since FS can be seen as a Rao–Blackwellisation of the BS approximation, i.e. since

$$\mathbb{Q}_{T,\text{FS}}^N(F_T) = \mathbb{E}\big[\mathbb{Q}_{T,\text{BS}}^N(F_T)\big|w_{1:T}^{1:N}, X_{1:T}^{1:N}\big]. \qquad (5)$$

As proposed in [23], the $\mathcal{O}(MN)$-complexity of BS can be reduced to $\mathcal{O}(N)$ using an accept-reject step which circumvents the need for evaluating the denominator in (3), and [24] developed an online implementation around this idea called particle-based rapid incremental smoother (PaRIS). However, the accept-reject step typically requires $\mathcal{O}(\mathrm{e}^V)$ *proposed* samples to obtain a single *accepted* sample. The overall complexity of PaRIS or of the backward sampler from [23] is therefore $\mathcal{O}(NT\mathrm{e}^V)$.

## III. Blocked Particle Filtering

In this section, we describe the canonical class of high-dimensional state-space models for which we will compute the smoother in the next section. The same model was used in [13] to analyse their blocked filtering algorithm which is also reviewed.

### *A. Class of High-dimensional State-Space Models*

The state-space model $(X_t, Y_t)_{t\in\mathbb{N}}$ from Subsection II-A is now developed into a high-dimensional model as follows. We assume that the state space $\mathbb{X} = \prod_{v\in\mathbb{V}} \mathbb{X}_v$ is endowed with a graph $\mathcal{G} := (\mathbb{V}, \mathcal{E})$ where vertices $v \in \mathbb{V}$ index the components of the state vector and edges $e \in \mathcal{E}$ define the spatial correlation structure. The latent states $X_t := (X_{t,v})_{v\in\mathbb{V}}$ are then $V$-dimensional, with $V = \operatorname{card}\mathbb{X} = \operatorname{card}\mathbb{V}$ again being the model dimension. For each component $X_{t,v}$, taking a value $x_{t,v} \in \mathbb{X}_v$, we obtain an observation $Y_{t,v}$, taking a value $y_{t,v} \in \mathbb{Y}_v$. Thus, $Y_t := (Y_{t,v})_{v\in\mathbb{V}}$ takes a value in $\mathbb{Y} = \prod_{v\in\mathbb{V}} \mathbb{Y}_v$. Finally, for all $v \in \mathbb{V}$, we let $\psi_v$ and $\varphi_v$ be reference measures on $\mathbb{X}_v$ and $\mathbb{Y}_v$, respectively, with $\psi = \prod_{v\in\mathbb{V}} \psi_v$ and $\varphi = \prod_{v\in\mathbb{V}} \varphi_v$. We assume that the densities satisfy the following properties.

i) The initial density factorises as $m(x) = \prod_{v\in\mathbb{V}} m_v(x_v)$, where $m_v \colon \mathbb{X}_v \to (0, \infty)$ is a density w.r.t. $\psi_v$.

ii) The transition densities $p$ and $g$ factorise as

$$p(z, x) = \prod_{v\in\mathbb{V}} p_v(z, x_v) \;\text{ and }\; g(x, y) = \prod_{v\in\mathbb{V}} g_v(x_v, y_v),$$

where $p_v \colon \mathbb{X} \times \mathbb{X}_v \to (0, \infty)$ and $g_v \colon \mathbb{X}_v \times \mathbb{Y}_v \to (0, \infty)$ are transition densities w.r.t. $\psi_v$ and $\varphi_v$, respectively.

iii) Let $\mathcal{N}_R(v)$ be the $R$-neighbourhood of the vertex $v$, i.e.

$$\mathcal{N}_R(v) := \{u \in \mathbb{V} \mid d(u, v) \le R\},$$

where $d(u, v)$ is the length of the shortest path between vertices $u$ and $v$, and $R > 0$. The parameter $R$ is fixed throughout this work and we write $\mathcal{N}(v) := \mathcal{N}_R(v)$. We assume that $R$ governs the spatial correlation of the model in the sense that $p_v(z, x_v) = p_v(z', x_v)$ for any $(z, z', x_v) \in \mathbb{X}^2 \times \mathbb{X}_v$ with $z_{\mathcal{N}(v)} = z'_{\mathcal{N}(v)}$, where $z_K := (z_v)_{v\in K}$, for any $K \subseteq \mathbb{V}$. Under the model, the $v$th component thus only depends on the components in $\mathcal{N}(v)$ at the previous time step which allows us to slightly abuse notation to write $p_v(z_{\mathcal{N}(v)}, x_v) := p_v(z, x_v)$.

The structure of the model is illustrated in Fig. 1.

For the blocked particle filters reviewed in this section, we also assume that the proposal kernel used by the PF factorises in the same way as the model transitions. That is, $q_t(z, x) := \prod_{v\in\mathbb{V}} q_{t,v}(z, x_v)$, where $q_{t,v} \colon \mathbb{X} \times \mathbb{X}_v \to (0, \infty)$ are transition densities w.r.t. $\psi_v$. We assume that $q_t$ induces the same spatial correlation structure as $p$ which allows us to

[1] Note that additivity in time of the test function is not needed for BS.

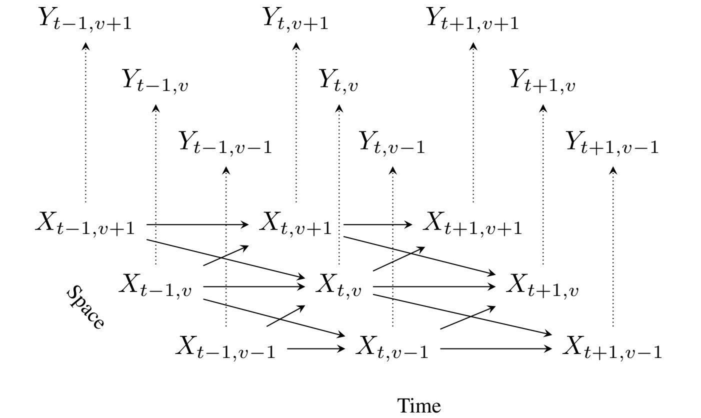


Fig. 1. Sketch of the state-space model considered in this work. In this example, $R = 1$, i.e. $\mathcal{N}(v) = \{v-1, v, v+1\}$. A similar figure can be found in [13].

write $q_{t,v}(z_{\mathcal{N}(v)}, x_v) := q_{t,v}(z, x_v)$. Again, all above-mentioned densities may depend on some model parameter $\theta$ though we suppress $\theta$ from the notation if possible.

### B. Localisation

Let $\pi_{t,K}$ denote the marginal of $\pi_t$ on $\mathbb{X}_K := \prod_{v \in K} \mathbb{X}_v$. Assume that $f\colon \mathbb{X} \to \mathbb{R}$ is *local,* i.e. it only depends on the components in some block $K \subseteq \mathbb{V}$. More formally, there exists $f_K\colon \mathbb{X}_K \to \mathbb{R}$ such that $f(z) = f(x) =: f_K(x_K)$ for any $x, z \in \mathbb{X}$ with $x_K = z_K$. In this case, $\pi_t(f) = \pi_{t,K}(f_K)$.

The number of particles $N$ must usually grow exponentially in the model dimension $V$ to control the error of the PF approximation in (1). Unfortunately, this curse of dimensionality persists even if we only want to approximate the integral of a local function, $\pi_t(f) = \pi_{t,K}(f_K)$, because the importance weights $W_t^n$ still depend on $V$.

A naïve attempt to stabilise the local filtering error, i.e. the error of the particle approximation of $\pi_{t,K}(f_K)$, as $V$ grows, would be to replace the weights in (1) by weights which only depend on the components of the particles $X_t^{1:N}$ in $K$. That is, we could approximate $\pi_{t,K}(f_K)$ by

$$\sum_{n=1}^{N} W_{t,K}^n f_K(X_{t,K}^n), \tag{6}$$

with *local* weights $W_{t,K}^n := w_{t,K}^n / \sum_{m=1}^N w_{t,K}^m$ defined through

$$w_{t,K}^n = \prod_{v \in K} G_{t,v}(X_{t-1,\mathcal{N}(v)}^{A_{t-1}^n}, X_{t,v}^n) = \prod_{v \in K} w_{t,v}^n.$$

Here, we have set $\mathcal{N}(K) := \bigcup_{v \in K} \mathcal{N}(v)$, for any $K \subseteq \mathbb{V}$, and

$$G_{t,v}(z_{\mathcal{N}(v)}, x_v) = \frac{p_v(z_{\mathcal{N}(v)}, x_v) g_v(x_v, y_{t,v})}{q_{t,v}(z_{\mathcal{N}(v)}, x_v)}.$$

The approximation in (6) is clearly independent of the dimension of $X_t$. Unfortunately, for $t > 1$, it still suffers from the curse of dimensionality because of the particles and weights indirectly depend on (the dimension of) $X_{1:t-1}$.

### C. Blocked Particle Filters

Let $\mathcal{K}$ be a partition of $\mathbb{V}$, i.e. $\bigcup_{K \in \mathcal{K}} = \mathbb{V}$ and $K \cap K' = \emptyset$ for all $K, K' \in \mathcal{K}$. To stabilise the local filtering errors, the *blocked particle filter (BPF)* from [13] seeks to make the entire evolution of the particle system depend only on local weights $W_{t,K}^n = w_{t,K}^n / \sum_{i=1}^N w_{t,K}^i$, for $K \in \mathcal{K}$ – and not just the construction of marginal filter approximations as in (6).

Algorithm 4 summarises the BPF. As before, we use the convention that any action described for *some* $v$, $K$, or $n$ is to be performed conditionally independently for *all* $v \in \mathbb{V}$, $K \in \mathcal{K}$ and $n \in \{1, \dots, N\}$.

*Algorithm 4 (blocked particle filter [13]):*

```
1) At Step 1,
    i) sample $X_1^n \sim Q_1$,
    ii) set $w_{1,v}^n := G_{1,v}(X_{1,v}^n)$ and $w_{1,K}^n := \prod_{v \in K} w_{1,v}^n$.
2) At Step $t$, $t > 1$,
    i) sample $A_{t-1,K}^n \sim \mathrm{Cat}(W_{t-1,K}^{1:N})$,
    ii) concatenate $\overline{X}_{t-1}^n := (X_{t-1,K}^{A_{t-1,K}^n})_{K \in \mathcal{K}}$,
    iii) sample $X_t^n \sim Q_t(\overline{X}_{t-1}^n, \cdot\,)$,
    iv) set $w_{t,v}^n := G_{t,v}(\overline{X}_{t-1,\mathcal{N}(v)}^n, X_{t,v}^n)$; $w_{t,K}^n := \prod_{v \in K} w_{t,v}^n$.
```

To obtain an approximation of the filter, [13] construct the blocking approximation

$$\tilde{\pi}_{t,\text{BPF}}^N(f) := \Big[\bigotimes_{K \in \mathcal{K}} \tilde{\pi}_{t,K}^N\Big](f), \tag{7}$$

where, for any $K \subseteq \mathbb{V}$ (i.e. even for $K \notin \mathcal{K}$), we have defined

$$\tilde{\pi}_{t,K}^N(f_K) := \sum_{n=1}^{N} W_{t,K}^n f_K(X_{t,K}^n). \tag{8}$$

We use the 'tilde'-symbol to stress that in contrast to standard PFs, the BPF asymptotically (as $N \to \infty$) does not target the true filter $\pi_t$ but rather an approximate, blocked filter $\tilde{\pi}_t$ which is defined in Appendix B.

The complexity of the BPF is $(N|\mathcal{K}|_\infty \operatorname{card} \mathcal{K})$ per time step where $|\mathcal{K}|_\infty := \max_{K \in \mathcal{K}} \operatorname{card} K$. Under strong mixing assumptions, [13], [25] show that local errors of $\tilde{\pi}_{t,\text{BPF}}^N$ are bounded uniformly in $t$ and $V$.

*Remark 1: Recall that the particle smoothers from Section II require approximations of the filters. Unfortunately, approximating* $\pi_t$ *by* $\tilde{\pi}_{t,\text{BPF}}^N$ *has complexity* $\mathcal{O}(N^{\operatorname{card} \mathcal{K}})$ *which is typically prohibitive. Instead, we propose to approximate* $\pi_t$ *by subsampling* $N$ *points conditionally independently from* $\tilde{\pi}_{t,\text{BPF}}^N$*. This has complexity* $\mathcal{O}(N)$*.*

*Remark 2: The blocked particle smoothers proposed in Section IV, will require approximations of filter marginals on blocks* $K' \supseteq K \in \mathcal{K}$*. More specifically,* $K'$ *will be some neighbourhood of* $K$*. Unfortunately, approximating* $\pi_{t,K'}$ *by marginalising* $\tilde{\pi}_{t,\text{BPF}}^N$ *has complexity* $\mathcal{O}(N^{\operatorname{card} \mathcal{L}})$*, where* $\mathcal{L} := \{K \in \mathcal{K} \mid K \cap K' \neq \emptyset\}$*, and this is typically prohibitive. Instead, we propose to approximate* $\pi_{t,K'}$ *by* $\tilde{\pi}_{t,K'}^N$ *(i.e. via (8) with* $K' = K$*). This has complexity* $\mathcal{O}(N)$*.*

## IV. Blocked Particle Smoothing

In this section, we formally show that standard particle smoothing breaks down in high dimensions – even under dimensionally stable filter approximations. We then propose novel blocked particle smoothers which provably circumvent this curse of dimensionality in the canonical high-dimensional model from Subsection III-A (potential extensions to other models are discussed in Subsection G of the supplementary materials).

### *A. High-dimensional Smoothing*

For the remainder of this work, we will be concerned with approximating smoothed expectations $\mathbb{Q}_t(F_t)$ for the canonical high-dimensional model in the case that $F_t \colon \mathbb{X}^t \to \mathbb{R}$ is additive in both *time and space*[2], in the sense that there exist functions $f_{1,v} \colon \mathbb{X}_v \to \mathbb{R}$ and $f_{t,v} \colon \mathbb{X}_{\mathcal{N}(v)} \times \mathbb{X}_v \to \mathbb{R}$, for $t > 1$, such that

$$F_t(x_{1:t}) = \sum_{s=1}^{t} \sum_{v \in \mathbb{V}} f_{s,v}(x_{s-1,\mathcal{N}(v)}, x_{s,v}), \tag{9}$$

with the convention that any quantity with time index $0$ is to be ignored. Each constituent function $f_{t,v}$ may also implicitly depend on $y_{t,v}$.

We now give an example of such an additive function which is important for the problem of calibrating $\theta$.

*Example 1 (score): Assume that $m_v = m_v^\theta$, $p_v = p_v^\theta$, $g_v = g_v^\theta$ and hence $\mathit{\Gamma}_T = \mathit{\Gamma}_T^\theta$ depend on some unknown parameter vector $\theta$. Approximating the (marginal) maximum-likelihood estimate (MLE) of $\theta$ via a gradient-ascent algorithm requires computing the score, i.e. the gradient of the (marginal) log-likelihood, given by $\nabla_\vartheta \log \mathit{\Gamma}_T^\vartheta(\mathbf{1})|_{\vartheta=\theta} = \mathbb{Q}_T^\theta(F_T^\theta)$, where $F_T^\theta$ is as in* (9) *with*

$$\begin{aligned} f_{1,v}^\theta(x_v) &= \nabla_\vartheta \log\bigl[m_v^\vartheta(x_v) g_v^\vartheta(x_v, y_{1,v})\bigr]\big|_{\vartheta=\theta}, \\ f_{t,v}^\theta(z_{\mathcal{N}(v)}, x_v) &= \nabla_\vartheta \log\bigl[p_v^\vartheta(z_{\mathcal{N}(v)}, x_v) g_v^\vartheta(x_v, y_{t,v})\bigr]\big|_{\vartheta=\theta}. \end{aligned}$$

### *B. Assumptions*

In this subsection, we state some assumptions under which we prove various theoretical results below.

Assumption 1 is a regularity condition routinely used in the analysis of SMC techniques (e.g. [26]). It can often be relaxed at the price of significantly complicating the analysis [27], [28]. Assumption 2 requires $F_T$ to be spatially and temporally local. Recall that by (9), the smoothing functional of interest decomposes into a sum of such local functions. For simplicity, as in [21], [22], we assume here that $f_{t,J}$ does not depend on the state at time $t-1$ but this could be relaxed. Assumption 3 lists a number of options for the Monte Carlo approximation $\pi_t^N$ of the filter $\pi_t$ needed for the standard particle smoothers.

*Assumption 1 (strong mixing condition): For any $v \in \mathbb{V}$ the dominating measure $\psi_v$ is finite, and there exist $\varepsilon > 0$ such that for all $(x, z, y) \in \mathbb{X}^2 \times \mathbb{Y}$ and all $v \in \mathbb{V}$,*

$$\varepsilon^{-1} \leq p_v(x_{\mathcal{N}(v)}, z_v) \leq \varepsilon \quad \textit{and} \quad 0 < \int_{\mathbb{X}_v} g_v(z_v, y_v)\psi_v(\mathrm{d}z_v) < \infty.$$

*Assumption 2 (local test function): There exist $J \subseteq K \in \mathcal{K}$, $r \in \{1, \dots, T\}$, and $f_{r,J} \colon \mathbb{X}_J \to \mathbb{R}$ with $\|f_{r,J}\| \leq 1$ such that $f_{r,J}$ is not $\psi$-almost everywhere constant and*

$$F_T(x_{1:T}) = f_{r,J}(x_{r,J}), \quad \textit{for all } x_{1:T} \in \mathbb{X}^T.$$

*Assumption 3 (filter approximations): For any $t \in \mathbb{N}$, approximate $\pi_t$ using*

*a) a standard PF with $N$ particles (i.e. via* (1)*),*
*b) $N$ samples drawn conditionally independently from the BPF approximation in* (7) *(see Remark 1),*
*c) $N$ IID samples from the exact filter, $\pi_t$,*
*d) $N$ IID samples from the blocked filter $\tilde{\pi}_t$ (see Appendix B).*

### *C. Breakdown of Standard Particle Smoothing*

In this subsection, we show that standard particle smoothing suffers from a curse of dimensionality, even if local filter errors are dimension-independent.

The efficiency of standard particle smoothing relies strongly on the mixing properties of the transitions $p$. Unfortunately, the mixing of $p(x, z) = \prod_{v \in \mathbb{V}} p_v(x, z_v)$ degrades exponentially in the model dimension $V$, unless the local transitions are perfectly mixing in the following sense.

*Definition 1 (perfect mixing): The local transitions $p_v$ are called* perfectly mixing *if $\int_A p_v(x_{\mathcal{N}(v)}, z_v)\psi_v(\mathrm{d}z_v)$ is $\prod_{u \in \mathcal{N}(v)} \psi_u$-almost everywhere constant for all $A \subseteq \mathbb{X}_v$.*

Note that if the transitions are perfectly mixing, the latent states are independent over time in which case particle filtering/smoothing methodology is not needed.

Our main result in this subsection is Proposition 1, proved in Appendix A. It shows that the asymptotic variance associated with the standard FS approximation $\mathbb{Q}_{T,\mathrm{FS}}^N(F_T)$, defined in (4) (see also Algorithm 2), grows exponentially in the model dimension unless the model transitions are perfectly mixing. The result holds even though we assume highly favourable conditions, summarised in Assumption 4, and even if local filter errors are dimension-independent.

*Assumption 4 (spatially IID model): We have $R = 0$, i.e. $p(x, z) = \prod_{v \in \mathbb{V}} p_v(x_v, z_v)$, for any $(x, z) \in \mathbb{X}^2$. In addition, for any $t \leq T$ and any $u, v \in \mathbb{V}$, we have $\mathbb{X}_u = \mathbb{X}_v =: \mathring{\mathbb{X}}$, $\mathbb{Y}_u = \mathbb{Y}_v =: \mathring{\mathbb{Y}}$, $p_u(x, z) = p_v(x, z) =: \mathring{p}(x, z)$, $g_u(x, y_{t,u}) = g_v(x, y_{t,v}) =: \mathring{g}_t(x)$, $\psi_u = \psi_v =: \mathring{\psi}$ and $\varphi_u = \varphi_v =: \mathring{\varphi}$.*

We stress that we only use Assumption 4 (which implies a complete absence of spatial interactions) to highlight the fact that standard particle smoothers break down in high dimensions even under such highly favourable conditions. The novel smoothing methodology proposed below does not rely on this assumption.

*Proposition 1:*

1) *Under Assumptions 1, 2, 4 and if the filter is approximated according to Assumption 3a (using a bootstrap PF), or according to Assumption 3c,*

$$\sqrt{N}\bigl[\mathbb{Q}_{T,\mathrm{FS}}^N(F_T) - \mathbb{Q}_T(F_T)\bigr] \Rightarrow \mathrm{N}(0, \sigma_T^2(F_T)),$$

*as $N \to \infty$, where*

$$\sigma_T^2(F_T) \geq \sum_{t=1}^{T} a_{t,T}(f_{r,J})(c_{t,T})^V. \tag{10}$$

*Here, for each $t \leq T$, $a_{t,T}(f_{r,J}) > 0$ and $c_{t,T} \geq 1$ do not depend on the model dimension, $V$.*

2) *For all $t \leq T$, $c_{t,T} > 1$, unless the model transitions are perfectly mixing.*

Note that $c_{t,T} > 1$ implies that the asymptotic variance in Equation 10 grows exponentially in $V$.

[2] Strictly speaking, as with standard particle smoothers, additivity in *time* in only needed for the FS but not the BS variant of the algorithm.

Although Proposition 1 has been established for FS (which computes the sum in (4) exactly), it extends immediately to BS and to the PaRIS algorithm from [24] as these construct sampling approximations of (4). Due to (5), they cannot attain a smaller variance than FS (for equal numbers of particles $N$).

### D. Proposed Algorithms

In this subsection, we propose novel 'blocked' particle smoothers aimed at circumventing the curse of dimensionality analysed in Proposition 1.

*1) Blocked Backward Kernels:* The curse of dimensionality suffered by standard particle smoothing methods is due to the dependence of the backward kernel in (3) on the mixing properties of the full model transitions $p$ (as outlined above, these deteriorate as the model dimension increases). Thus, we need to design approximate versions of the backward kernels which only operate on some 'fixed' block $K \subseteq \mathbb{V}$ of components and which therefore only rely on the mixing properties of the transitions associated with $K$:

$$p_K(x_{\mathcal{N}(K)}, z_K) := \prod_{v \in K} p_v(x_{\mathcal{N}(v)}, z_v).$$

To achieve this, we employ *blocked* backward kernels $B_{K,\nu} \colon \mathbb{X}_K \times \mathrm{Borel}(\mathbb{X}_{\mathcal{N}(K)}) \to [0,1]$, for $x \in \mathbb{X}_K$ given by

$$B_{K,\pi_{t,\mathcal{N}(K)}}(x, \mathrm{d}z) := \frac{p_K(z,x)\pi_{t,\mathcal{N}(K)}(\mathrm{d}z)}{\int_{\mathbb{X}_{\mathcal{N}(K)}} p_K(u,x)\pi_{t,\mathcal{N}(K)}(\mathrm{d}u)}. \tag{11}$$

Again, the filter marginal $\pi_{t,\mathcal{N}(K)}$ is typically intractable and must be replaced by a Monte Carlo approximation $\pi^N_{t,\mathcal{N}(K)}$.

*2) Blocked Forward Smoothing:* In the remainder of this work, for any $K \in \mathcal{K}$, we let $K \subseteq \overline{K} \subseteq \mathbb{V}$ be some enlarged block containing $K$. Note that with this notation, $\pi^N_{t,\mathcal{N}(\overline{K})}$ is an approximation of the marginal filter on the components in the neighbourhood of the enlargement of $K$.

Our proposed blocked FS scheme is outlined in Algorithm 5, where as usual, any action prescribed for *some* $n$ is to be performed conditionally independently for *all* $n \in \{1, \ldots, N\}$.

```
Algorithm 5 (blocked forward smoothing):
1) Perform the following steps (in parallel) for any $K \in \mathcal{K}$.
    i) Set $\alpha^n_{1,\overline{K}} := f_{1,K}(X^n_{1,K})$. For $t > 1$, set
        $$\alpha^n_{t+1,\overline{K}} := \sum_{m=1}^N B_{\overline{K},\pi^N_{t,\mathcal{N}(\overline{K})}}(X^n_{t+1,\overline{K}}, \{X^m_{t,\mathcal{N}(\overline{K})}\}) \times \big[\alpha^m_{t,\overline{K}} + f_{t+1,K}(X^m_{t,\mathcal{N}(K)}, X^n_{t+1,K})\big].$$
2) Approximate $\mathbb{F}_t$ by $\mathbb{F}^N_t := \sum_{K \in \mathcal{K}} \mathbb{F}^N_{t,K}$, where, for any
    $K \in \mathcal{K}$, $\mathbb{F}^N_{t,K} := \sum_{n=1}^N W^n_{t,\overline{K}} \alpha^n_{t,\overline{K}}$.
```

As in the case of standard FS (Algorithm 2), Algorithm 5 can be implemented online. That is, the terms $\alpha^n_{t,\overline{K}}$ can be determined at the $t$th step of the Monte Carlo filter used to generate $(X^n_{t,v}, w^n_{t,v})$. The computational complexity is $\mathcal{O}(N^2 T [\operatorname{card} \mathcal{K}] \max_{K \in \mathcal{K}} \operatorname{card} \mathcal{N}(\overline{K}))$ when running the algorithm up to time $T$. This is slightly higher than that of the (dimensionally unstable) Algorithm 2. However, significant speed-ups should be attainable since the blocked smoothing recursions can be run in parallel on distributed architectures or multiple cores (see [29]).

*3) Blocked Backward Sampling:* As with standard particle smoothing, we may reduce the computational complexity of blocked FS to $\mathcal{O}(NMT[\operatorname{card}\mathcal{K}] \max_{K \in \mathcal{K}} \operatorname{card} \mathcal{N}(\overline{K}))$ via a simple Monte Carlo approximation based on $M < N$ particle paths. Algorithm 6 outlines blocked BS. Here, we use the convention that any action prescribed for *some* $m$ is to be performed conditionally independently for *all* $m \in \{1, \ldots, M\}$.

```
Algorithm 6 (blocked backward sampling):
1) Perform the following steps (in parallel) for any $K \in \mathcal{K}$.
    i) Sample $\widetilde{X}^m_{T,\overline{K}} \sim \pi^N_{T,\overline{K}}$. For $t = T-1, \ldots, 1$, sample
        $$\widetilde{X}^m_{t,\mathcal{N}(\overline{K})} \sim B_{\overline{K},\pi^N_{t,\mathcal{N}(\overline{K})}}(\widetilde{X}^m_{t+1,\overline{K}}, \cdot\,).$$
    ii) Set $\mathbb{F}^N_{T,K} := \frac{1}{M}\sum_{m=1}^M \sum_{t=1}^T f_{t,K}(\widetilde{X}^m_{t-1,\mathcal{N}(K)}, \widetilde{X}^m_{t,K})$.
2) Approximate $\mathbb{F}_T$ by $\mathbb{F}^N_T := \sum_{K \in \mathcal{K}} \mathbb{F}^N_{T,K}$.
```

*4) Bias Reduction:* The blocking strategy introduces a bias but as shown in Proposition 3 below, this bias is *bounded* uniformly in the time horizon and model dimension. In addition, we can *reduce* the bias by defining the enlarged blocks as $\overline{K} := \mathcal{N}_i(K) := \bigcup_{v \in K} \mathcal{N}_i(v)$, for some $i > 0$. That way, blocked particle smoothers do not require evaluating test functions at components near block boundaries (by Proposition 3 below, the bias decays exponentially in the distance to the block boundary).

Though as we will discuss in more detail in Subsection IV-E, any bias reduction attained by choosing larger (enlarged) blocks needs to be carefully balanced against the variance increase this induces. In particular, taking $\overline{K} = \mathbb{V}$ trivially minimises the bias but then the blocked particle smoothers coincide with the (dimensionally unstable) standard particle smoothers.

*5) Marginal Filter Approximations:* As with standard particle smoothers, any stable (Monte Carlo) approximation of filter marginals can be plugged into Algorithms 5 and 6. We consider the following approximations $\pi^N_{t,K'}$ of filter marginals $\pi_{t,K'}$.

*Assumption 5 (marginal filter approximations): For any $t \in \mathbb{N}$ and any block $K' \subseteq \mathbb{V}$, approximate $\pi_{t,K'}$ using a*

*a) standard PF but based on local weights as in* (6),

*b) BPF (via* (8) *as justified in Remark 2).*

*c) suitable marginal of Assumption 3c (IID samples from $\pi_t$),*

*d) suitable marginal of Assumption 3d (IID samples from $\tilde{\pi}_t$).*

### E. Theoretical Analysis

In this subsection, we do not consider enlarged blocks, i.e. the blocks used for smoothing are the same blocks employed by the BPF. Under Assumption 2, for $J \subseteq K = \overline{K} \in \mathcal{K}$, the blocked FS approximation of $\mathbb{Q}_T(F_T)$ then simplifies to

$$\mathbb{Q}^N_{T,\text{FS}(K)}(F_T) = \int f_{r,J}(x_{r,J}) \pi^N_{T,K}(\mathrm{d}x_{T,K}) \times \prod_{t=1}^{T-1} B_{K,\pi^N_{t,\mathcal{N}(K)}}(x_{t+1,K}, \mathrm{d}x_{t,\mathcal{N}(K)}).$$

We now derive uniform (in both the time horizon $T$ and model dimension $V$) bounds on the asymptotic (as $N \to \infty$) bias and variance of the blocked FS estimator $\mathbb{Q}^N_{T,\text{FS}(K)}(F_T)$. Following [13], we also show that this estimator can be made locally consistent by scaling $K$ appropriately with $N$.

*1) Variance:* We now state Proposition 2, proved in Appendix A, which suggests that the variance of blocked FS is bounded in time and in the model dimension but may grow exponentially in the block size for a fixed number of particles (so that controlling the variance requires $N$ to grow exponentially in the block size). To prove this result, we specifically assume that we approximate the filter at any time $t$ using IID samples from $\tilde{\pi}_t$ (Part 1) or from $\pi_t$ (Part 2). In addition, recall that $|\mathcal{K}|_\infty = \max_{K \in \mathcal{K}} \operatorname{card} K$ and let $\mathrm{N}(\mu, \sigma^2)$ be a normal distribution with mean $\mu \in \mathbb{R}$ and variance $\sigma^2 > 0$. The measures $\overline{\mathbb{Q}}_T$ and $\widehat{\mathbb{Q}}_T$ which govern the asymptotic mean in this central limit theorem are specified in Appendix A. We stress that Part 2 of this proposition is included mainly as a direct contrast to the negative result in Proposition 1.

*Proposition 2 (asymptotic variance): Under Assumptions 1 and 2, there exist probability measures $\overline{\mathbb{Q}}_T$ and $\widehat{\mathbb{Q}}_T$ on $\mathbb{X}^T$ such that the following statements hold.*

1) *If filter marginals are approximated per Assumption 5d*

$$\sqrt{N}\big[\mathbb{Q}^N_{T,\mathrm{FS}(K)}(F_T) - \overline{\mathbb{Q}}_T(F_T)\big] \Rightarrow \mathrm{N}(0, \sigma^2_T(F_T)),$$

*as $N \to \infty$, where for $c_1 > 0$ which only depends on $\varepsilon$,*

$$\sigma^2_T(F_T) \le \mathrm{e}^{c_1 \operatorname{card} K} \le \mathrm{e}^{c_1 |\mathcal{K}|_\infty}.$$

2) *Part 1 remains valid (but with asymptotic mean $\widehat{\mathbb{Q}}_T(F_T)$) if filter marginals are approximated per Assumption 5c.*

*2) Bias:* We now state Proposition 3, proved in Appendix A, which shows that the bias of blocked FS is bounded in time and in the model dimension and decays exponentially in the distance to the block boundary. This result assumes that we use either the BPF or IID samples from $\tilde{\pi}_t$ to approximate the filter at each time step. In addition, for any $K, K' \subseteq \mathbb{V}$, we write $d(K, K') := \min_{v \in K} \min_{v' \in K'} d(v, v')$ and $\partial K := \{v \in K \mid \mathcal{N}(v) \nsubseteq K\}$.

*Proposition 3 (asymptotic bias): Under Assumptions 1, 2, and if filter marginals are approximated either per Assumption 5b or per Assumption 5d,*

$$\big|\overline{\mathbb{Q}}_T(F_T) - \mathbb{Q}_T(F_T)\big| \le c_2 \operatorname{card}(J) \mathrm{e}^{-c_3\, d(J, \partial K)},$$

*where $c_2 \in \mathbb{R}$ and $c_3 > 0$ only depend on $\varepsilon$ and $R$.*

*3) Consistency:* Propositions 2 and 3 indicate a bias–variance trade-off: for a fixed number of particles, the bias can be reduced by increasing the size of the blocks but the variance typically grows exponentially in the block size. To minimise the mean-square error (MSE), we must therefore scale the size of the blocks suitably in $N$ to balance the variance and the (squared) bias.

For any $(t, v) \in \{1, \dots, T\} \times \mathbb{V}$, define the function $F_{t,v} \colon \mathbb{X}^T \to \mathbb{R}$ for any $x_{1:T} \in \mathbb{X}^T$ by $F_{t,v}(x_{1:T}) := x_{t,v}$. The *local* MSE for component $v \in K \in \mathcal{K}$ of blocked FS is

$$MSE_{t,v}(N) := \mathbb{E}\big[(\mathbb{Q}^N_{T,\mathrm{FS}(K)}(F_{t,v}) - \mathbb{Q}_T(F_{t,v}))^2\big].$$

The blocked particle smoothing estimate can then be made locally MSE-consistent by growing blocks suitably logarithmically in $N$ because Propositions 2 and 3 imply

$$MSE_{t,v}(N) = \mathcal{O}\big(\mathrm{e}^{-2c_3\, d(v, \partial K)} + N^{-1} \mathrm{e}^{c_1 \operatorname{card} K}\big). \qquad (12)$$

As a simple illustration, assume that the spatial graph $\mathcal{G}$ is a one-dimensional lattice as in Fig. 1 and consider a component $v$ at the centre of $K := \mathcal{N}_i(v)$ so that $d(v, K) = i$ and $\operatorname{card} K = 2i + 1$, for some $i \ge 0$. The local MSE in (12) then vanishes as $N \to \infty$ if the block radius grows as $i = o([\log(N)/c_1 - 1]/2)$. For instance, taking $i = \lfloor \log(N)/(8c_1) \rfloor$ implies that the local MSE is $\mathcal{O}(\exp(-\frac{c_3}{2c_1}\log(N)))$. A straightforward modification of [13, Corollary 2.5] extends such local consistency to components $v$ which are not at the centre of some block or to the case that $\mathcal{G}$ is a $q$-dimensional lattice. Though, in both cases, $i$ must grow even more slowly with $N$.

## V. Simulations

In this section, we compare standard and blocked particle smoothers on a high-dimensional state-space model.

### A. The Model

Blocking strategies have already been successfully applied to perform *filtering* in a functional magnetic resonance imaging (FMRI) application [12] and in military multiple-target tracking scenarios [30]. However, to assess the performance of our *smoothing* algorithms, we consider the more abstract model from [13] which is increasingly popular as a benchmark for SMC algorithms in high dimensions [16], [17], [31]. Purely in order to compare our method against analytical solutions, we let the model be linear and Gaussian.

Let $\mathrm{N}(\,\cdot\,; \mu, \Sigma)$ denote the density of a normal distribution with suitable mean vector $\mu$ and covariance matrix $\Sigma$, let $\mathbf{0}_V \in \mathbb{R}^V$ be a vector of zeros and let $\mathbf{I}_V \in \mathbb{R}^{V \times V}$ be the identity matrix. Then the model is given by $\mathbb{X} = \mathbb{Y} = \mathbb{R}^V$, $m^\theta(x) = \mathrm{N}(x; \mathbf{0}_V, \mathbf{I}_V)$,

$$p^\theta(z, x) = \mathrm{N}(x; Az, \sigma_X^2 \mathbf{I}_V), \quad g^\theta(x, y) = \mathrm{N}(y; x, \sigma_Y^2 \mathbf{I}_V).$$

Here, $\sigma_X, \sigma_Y > 0$ and $A = (a_{i,j})_{(i,j) \in \mathbb{V}^2}$ is a symmetric, banded diagonal (i.e. symmetric Toeplitz) matrix whose diagonal entries are $a_0, a_1, \dots, a_R > 0$ for some $R \in \mathbb{N} \cup \{0\}$:

$$a_{i,j} := \begin{cases} a_r, & \text{if } r \in \{0, 1, \dots, R\} \text{ and } j \in \{i + r, i - r\}, \\ 0, & \text{otherwise.} \end{cases}$$

This induces a local spatial correlation structure because

$$p_v^\theta(z_{\mathcal{N}(v)}, x_v) = \mathrm{N}(x_v; \textstyle\sum_{r=0}^R a_r \sum_{u \in \mathcal{B}_r(v)} z_u, \sigma_X^2),$$

where $\mathcal{B}_r(v) := \{u \in \mathbb{V} \mid d(u, v) = r\}$ denotes the vertices in $\mathbb{V}$ whose distance from vertex $v$ is exactly $r$.

We parametrise the model via

$$\theta := \theta_{0:R+2} := (a_0, a_1, \dots, a_R, \log \sigma_X, \log \sigma_Y).$$

All simulation results use $R = 1$ with true parameter values $a_0 = 0.5$, $a_1 = 0.2$, $\sigma_X = \sigma_Y = 1$.

### B. Smoothing

In this subsection, for fixed $\theta$, we estimate the smoothed sufficient statistic $\mathbb{F}_T$, defined according to (9) with $f_{1,v} \equiv 0$ and, for $t \in \mathbb{N}$, by

$$f_{t+1,v}(x_{t,\mathcal{N}(v)}, x_{t+1,v}) := x_{t+1,v} \textstyle\sum_{u \in \mathcal{B}_r(v)} x_{t,u}.$$

A full list of sufficient statistics for this model is given in Subsection I of the supplementary materials.

We run standard and blocked FS and BS (with $N = 500$ and $M = 100$) for model dimensions up to $V = 500$, for contiguous blocks of size $\operatorname{card} K \in \{1, 2, 20\}$ and for enlarged blocks $\overline{K} := \mathcal{N}_i(K) = \bigcup_{v \in K} \mathcal{N}_i(v)$, for $i \in \{0, 1\}$. The results in Fig. 2 are based on 400 independent repetitions, each using a different observation sequence $y_{1:20}$ sampled from the model.

For each algorithm, we compare the impact of different filter approximations. Standard and blocked PFs used the conditionally, locally optimal proposal kernel $q_{t,v}(x_{\mathcal{N}(v)}, z_v) \propto p_v(x_{\mathcal{N}(v)}, z_v)g(z_v, y_{t,v})$. While this proposal is usually intractable, in the model class considered in this work, it can be 'exactly' approximated [31] (at an additional computational cost which grows in the model dimension). We employed the filter approximations from Assumptions 3a–c for the standard particle smoothers and from Assumptions 5a–c for the blocked particle smoothers. The results suggest the following interpretation.

- Fig. 2a illustrates that the root-mean-square error (RMSE) of estimates of $\mathbb{F}_T/V$ grows in $V$ when using a standard PF irrespective of the type of smoothing algorithm employed. This is consistent with our theory since Propositions 2 and 3 assume dimensionally stable local filter approximations and standard PFs (even with efficient proposals) normally break down in high dimensions. For instance, in a similar model, [17] reported that even the *fully-adapted auxiliary* PF [32], [33] fails to yield useful filter approximations for model dimensions as small as 20.
- Fig. 2b and 2c illustrate that when local filter errors are dimensionally stable, blocked particle smoothers yield estimates of $\mathbb{F}_T/V$ whose error is bounded in $V$. In contrast, standard particle smoothers induce an RMSE that appears to grow with the model dimension.
- The RMSE of both standard and blocked particle smoothers is slightly higher in Fig. 2b than in Fig. 2c. This is due to the additional bias induced by the BPF.
- The first and last column in Fig. 2b and Fig. 2c illustrate the consequence of a suboptimal solution to the bias–variance trade-off discussed at the end of Subsection IV-E. That is, for the given number of particles, the block sizes $\operatorname{card} K = 1$ or $\operatorname{card} K = 20$ induce an error that is larger than for the choice $\operatorname{card} K = 3$ displayed in the second column.

### C. Parameter Estimation

We now combine blocked particle smoothing with stochastic gradient-ascent and stochastic EM algorithms in order to approximate the MLE of the model parameters $\theta$ (see [34] for a comprehensive review).

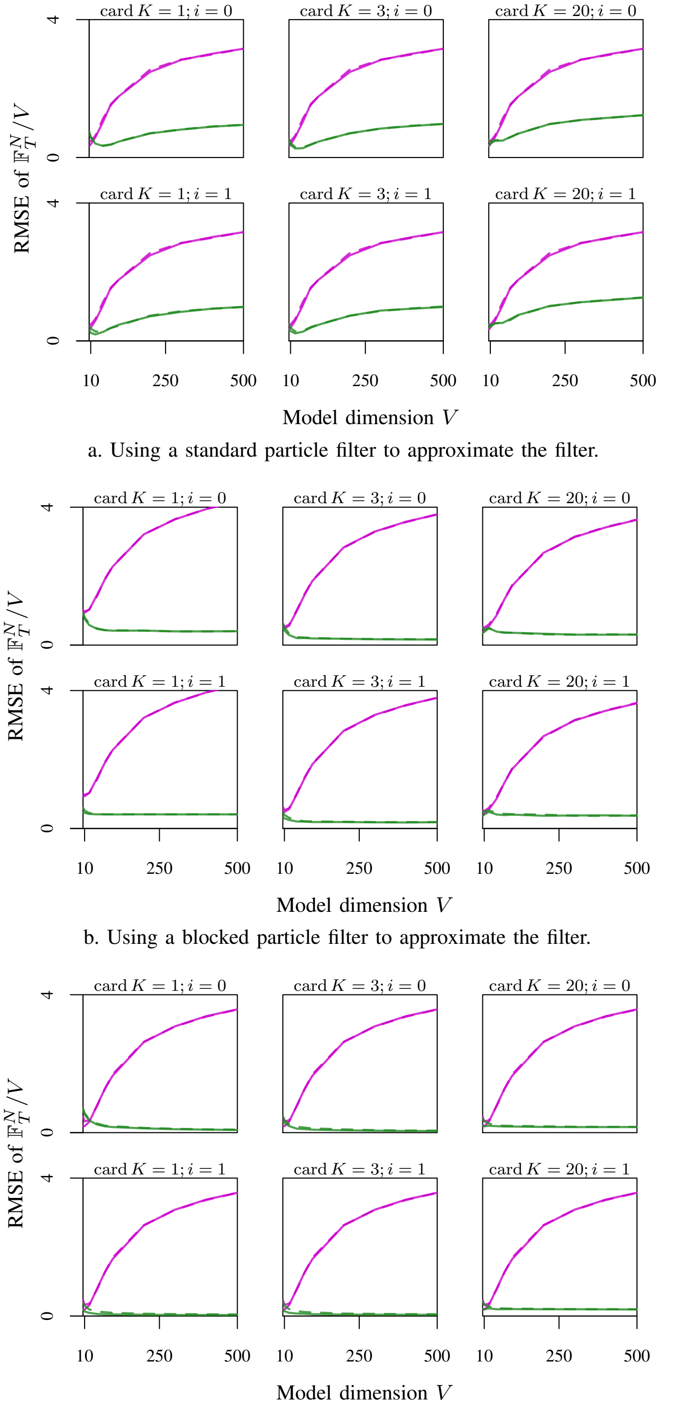
These algorithms generate a sequence of parameter values $(\theta[p])_{p \in \mathbb{N}}$ via some update rule which requires the evaluation of some smoothed sufficient statistics $\mathbb{F}_T^\theta$. As $\mathbb{F}_T^\theta$ is usually intractable, we use (blocked) particle smoothers to estimate it.

*1) Offline Gradient-ascent:* Let $(\gamma[p])_{p \in \mathbb{N}}$ be a step-size sequence which is non-negative, non-increasing, and which satisfies $\sum_{p=1}^\infty \gamma[p] = \infty$ as well as $\sum_{p=1}^\infty \gamma[p]^2 < \infty$. Gradient-ascent algorithms use the update rule

$$\theta[p+1] = \theta[p] + \gamma[p]\mathbb{F}_T^{\theta[p]}, \tag{13}$$

where $\nabla_\vartheta \log \varGamma_T^\vartheta(\mathbf{1})|_{\vartheta=\theta} = \mathbb{F}_T^\theta = \mathbb{Q}_T^\theta(F_T^\theta)$ is the score if the additive functionals $f_{t,v}^\theta$ are as defined in Example 1.

Details on the calculation of the score for the model considered in this section are given in Subsection I of the supplementary materials. The step sizes were $\gamma[p] = p^{-0.8}$. To avoid manual tuning of the step-size sequence, we normalised the gradient approximation by its $L_2$ norm in (13).



a. Using a standard particle filter to approximate the filter.

b. Using a blocked particle filter to approximate the filter.

c. Using IID samples from the filter.

Fig. 2. Root-mean-square error of the estimate of $\mathbb{F}_T/V$. Obtained from 400 simulation runs (each with a different observation sequence) using standard (—) and blocked (—) forward smoothing as well as standard (– –) and blocked (– –) backward sampling. Note that the lines for the standard smoothers are almost indistinguishable from one another and the same is true for the blocked smoothers.

*2) Offline EM:* EM algorithms use the update rule

$$\theta[p+1] := \varLambda(\mathbb{F}_T^{\theta[p]}) := \arg\max_\vartheta \mathbb{E}\Big[\log \frac{\mathrm{d}\varGamma_T^\vartheta}{\mathrm{d}\psi^{\otimes T}}(X_{1:T})\Big], \tag{14}$$

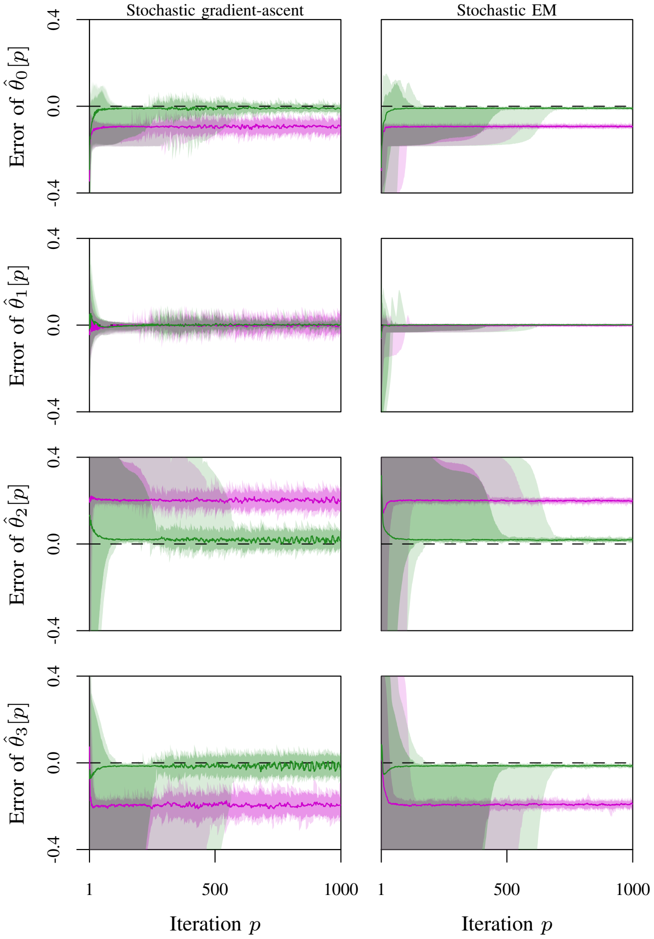


Fig. 3. Average error of the parameter estimates in the 100-dimensional state-space model. Obtained from 45 runs of the stochastic gradient-ascent and stochastic EM algorithms using standard (—) and blocked (—) backward sampling. We used blocks of size card $K = 3$ and enlarged blocks of the form $\mathcal{N}_2(K)$. The shaded areas delimit, respectively, the range all encountered realisations and the $(0.05, 0.95)$-quantiles.

where $X_{1:T} \sim \mathbb{Q}_T^\theta$ and where the expectation on the r.h.s. is usually a function of $\mathbb{F}_T^\theta$.

The exact form of the function $\Lambda$ in (14) for the particular model considered in this section is also given in Subsection I of the supplementary materials.

*3) Results:* We apply both algorithms, based on standard and blocked BS, to approximate the MLE of $\theta$ using $N = 500$, $M = 200$ and $T = 10$. In Fig. 3, we only show results for filter approximations obtained from the BPF (via Assumption 3b for standard BS and via Assumption 5b for blocked BS). As expected, using a standard PF led to large errors in all algorithms.

## VI. Summary

We have presented online and offline smoothing recursions for efficiently estimating smoothed functionals in a class of high-dimensional state-space models (extensions to other models are discussed in Subsection G of the supplementary materials). We combine these backward recursions with existing approximate forward-filtering recursions for high-dimensional models known as blocked particle filters. The resulting algorithms then exploit spatial additivity of the smoothing functionals and spatial ergodicity of the model to produce estimates whose local errors are independent of the model dimension (for a fixed number of particles). Thus, we circumvent the so called curse of dimensionality. This is in contrast to existing methods which require the number of particles to scale exponentially in the model dimension. We have successfully applied our algorithms to perform smoothing and maximum-likelihood estimation.

## Acknowledgements

We wish to thank the four anonymous referees for numerous helpful comments and in particular for suggesting the discussion of ways in which our methodology may be extended to other types of state-space models (Subsection G of the supplementary materials).

## Appendix

## Proof of Proposition 1

This appendix contains the proofs of Propositions 1, 2 and 3. To simplify the notation in all these proofs, we set $g_{t,v}(x_v) := g_v(x_v, y_{t,v})$, $g_{t,K}(x_K) := g_K(x_K, y_{t,K})$ and $g_t(x) := g(x, y_t)$ for any $x \in \mathbb{X}$, any $v \in \mathbb{V}$ and any $K \subseteq \mathbb{V}$.

Our proofs of Propositions 1 and 2 are based on [22, Theorem 3.1] which was formulated for prediction models but whose results can be easily transferred to updated models via [26, Section 2.4.3] as follows. The filter $\pi_t$ of a state-space model defined by the triple $(m, p, g_t)$ can be interpreted as the predictor of a state-space model defined by the triple $(\breve{m}, \breve{p}_t, \breve{g}_t)$, where

- $\breve{m}(x) := m(z)g_1(z)/\int_{\mathbb{X}} m(z)g_1(z)\psi(\mathrm{d}z)$,
- $\breve{p}_t(x, z) := p(x, z)g_t(z)/\int_{\mathbb{X}} p(x, z)g_t(z)\psi(\mathrm{d}z)$,
- $\breve{g}_t(x) := \int_{\mathbb{X}} p(x, z)g_{t+1}(z)\psi(\mathrm{d}z)$.

The remainder of this appendix is devoted to the proof of Proposition 1.

*Proof:* Let $\mathring{\pi}_t$ and $\mathring{\mathbb{Q}}_T$ represent, respectively, the marginal of the time-$t$ filter and of the joint smoothing distribution on $\mathring{\mathbb{X}}$, i.e. by Assumption 4, $\pi_t = \bigotimes_{v \in \mathbb{V}} \mathring{\pi}_t$ and $\mathbb{Q}_T = \bigotimes_{v \in \mathbb{V}} \mathring{\mathbb{Q}}_T$. Let $\mathring{B}_{\mathring{\pi}_t}(\mathring{x}, \mathrm{d}\mathring{z}) := \mathring{\pi}_t(\mathrm{d}\mathring{z})\mathring{p}(\mathring{z}, \mathring{x})/\mathring{\pi}_t(\mathring{p}(\,\cdot\,, \mathring{x}))$ be the backward kernel associated with the marginal of the model on one spatial component. Define $\mathring{Q}_s(\mathring{x}, \mathrm{d}\mathring{z}) := \mathring{p}(\mathring{x}, \mathring{z})\mathring{g}_s(\mathring{z})\mathring{\psi}(\mathrm{d}\mathring{z})$, $\mathring{\mathcal{Q}}_{t,T} := \bigotimes_{s=t+1}^T \mathring{Q}_s$ as well as $G_{t,T} := \bigotimes_{v \in \mathbb{V}} \mathring{G}_{t,T}$, with $\mathring{G}_{t,T} := \mathring{\mathcal{Q}}_{t,T}(\mathbf{1})/\mathring{\pi}_t \mathring{\mathcal{Q}}_{t,T}(\mathbf{1})$ and, for $\mathring{z}_t \in \mathring{\mathbb{X}}$,

$$\mathring{D}_{t,T}(\mathring{z}_t, \mathrm{d}\mathring{x}_{1:T}) := \delta_{\mathring{z}_t}(\mathrm{d}\mathring{x}_t)\Bigl[\prod_{s=1}^{t-1} \mathring{B}_{\mathring{\pi}_s}(\mathring{x}_{s+1}, \mathrm{d}\mathring{x}_s)\Bigr] \\ \times \mathring{\mathcal{Q}}_{t,T}(\mathring{x}_t, \mathrm{d}\mathring{x}_{t+1:T}),$$

Finally, $P_{t,T} := \bigotimes_{v \in \mathbb{V}} \mathring{P}_{t,T}$ with $\mathring{P}_{t,T} := \mathring{D}_{t,T}/\mathring{D}_{t,T}(\mathbf{1})$.

We can now prove Part 1. By [22, Theorem 3.1] – transferring the results from prediction to updated models as outlined above – the FS approximation of $\mathbb{Q}_T(F_T)$ is asymptotically normal with asymptotic variance given by

$$\sigma_T^2(F_T) := \sum_{t=1}^T \pi_t\bigl(G_{t,T}^2 P_{t,T}(F_T - \mathbb{Q}_T(F_T))^2\bigr). \quad \text{(A.15)}$$

Without loss of generality, assume that card $J = 1$, i.e. $J = \{u\}$, for some $u \in \mathbb{V}$. Write $F_T(x_{1:T}) \equiv \mathring{F}_T(x_{1:T,u}) \equiv f_{r,u}(x_{r,u})$. The $t$th term on the r.h.s. of (A.15) is then

$$\begin{aligned}\pi_t\bigl(G_{t,T}^2 P_{t,T}(F_T - \mathbb{Q}_T(F_T))^2\bigr) \\ = \mathring{\pi}_t\bigl(\mathring{G}_{t,T}^2 \mathring{P}_{t,T}(\mathring{F}_T - \mathring{\mathbb{Q}}_T(\mathring{F}_T))^2\bigr)\mathring{\pi}_t(\mathring{G}_{t,T}^2)^{V-1}_{.}\end{aligned} \quad \text{(A.16)}$$

Since Jensen's inequality ensures that

$$c_{t,T} := \mathring{\pi}_t(\mathring{G}_{t,T}^2) \geq \mathring{\pi}_t(\mathring{G}_{t,T})^2 = 1, \tag{A.17}$$

and since $f_{r,J}$ is not $\mathring{\psi}$-almost everywhere constant,

$$a_{t,T}(f_{r,J}) := \mathring{\pi}_t\big(\mathring{G}_{t,T}^2 \mathring{P}_{t,T}(\mathring{F}_T - \mathring{\mathbb{Q}}_T(\mathring{F}_T))^2\big)/c_{t,T} > 0.$$

This proves Part 1.

It remains to prove Part 2. If the model transitions are not perfectly mixing, $\mathring{G}_{t,T}$ is not $\mathring{\psi}$-almost everywhere constant. Hence, Jensen's inequality in (A.17) is strict (since $x \mapsto x^2$ is strictly convex). As a result, the r.h.s. in (A.16), and thus $\sigma_T^2(F_T)$, grows exponentially in $V$. □

## Proofs for Subsection IV-E

### A. *Outline*

In this section, we prove Propositions 2 and 3. Central to the proofs are three alternate state-space models which may be viewed as approximations of the original state-space model, i.e. approximations of the model defined by the triple $(m, p, g_t)$.

1) Model $(\widetilde{m}, \tilde{p}_t, \tilde{g}_t)$, defined in Subsection B, represents the asymptotic target distribution of the BPF as stated in [25]. It defines a joint smoothing distribution $\widetilde{\mathbb{Q}}_T$ and filters $\tilde{\pi}_t$.
2) Model $(\widehat{m}, \hat{p}_t, \hat{g}_t)$, defined in Subsection C, defines a joint smoothing distribution $\widehat{\mathbb{Q}}_T$ whose marginal on Block $K \supseteq J$ coincides with the corresponding marginal of $\mathbb{Q}_T$ and its filter coincides with $\pi_t$.
3) Model $(\overline{m}, \bar{p}_t, \bar{g}_t)$, defined in Subsection D, defines a joint smoothing distribution $\overline{\mathbb{Q}}_T$ whose marginal on Block $K \supseteq J$ coincides with the corresponding marginal of $\widetilde{\mathbb{Q}}_T$ and its filter coincides with $\tilde{\pi}_t$.

When approximating the filter using IID samples from $\pi_t$ (respectively $\tilde{\pi}_t$), blocked particle smoothing for $(m, p, g_t)$ coincides with standard particle smoothing for $(\widehat{m}, \hat{p}_t, \hat{g}_t)$ (respectively $(\overline{m}, \bar{p}_t, \bar{g}_t)$). The central limit theorem for standard particle smoothing from [22] then proves Proposition 2.

When approximating the filter using the BPF (or IID samples from $\tilde{\pi}_t$), blocked particle smoothing for $(m, p, g_t)$ coincides with standard particle smoothing for $(\widetilde{m}, \tilde{p}_t, \tilde{g}_t)$. The bound on the bias from [13] then proves Proposition 3.

For probability measures $\mu$ and $\nu$ on $\mathbb{S} = \prod_{i \in \mathbb{I}} \mathbb{S}_i$ and $I \subseteq \mathbb{I} \subseteq \mathbb{N}$, define the *local total variation distance*

$$\|\mu - \nu\|_I := \sup_{f \in \mathcal{S}_I} |\mu(f) - \nu(f)|,$$

where $\mathcal{S}_I$ is the class of measurable functions $f \colon \mathbb{S} \to [-1, 1]$ such that for all $x, z \in \mathbb{S}$, $x_I = z_I$ implies $f(x) = f(z)$.

### B. *Model* $(\widetilde{m}, \tilde{p}_t, \tilde{g}_t)$

The state-space model $(\widetilde{m}, \tilde{p}_t, \tilde{g}_t)$ is defined by $\widetilde{m} := m$, $\tilde{g}_t := g_t$, and $\tilde{p}_t(x, z) := \prod_{K \in \mathcal{K}} \tilde{p}_{t,K}(x_K, z_K)$ with

$$\tilde{p}_{t,K}(x_K, z_K) := \int_{\mathbb{X}} p_K(x_{\mathcal{N}(K)}, z_K)\tilde{\pi}_{t-1,\mathcal{N}(K)\setminus K}(\mathrm{d}x_{\mathcal{N}(K)\setminus K}).$$

Let $\widetilde{\mathbb{Q}}_t$ be the joint smoothing distribution of this model (on $\mathbb{X}^t$) and $\tilde{\pi}_t(A) := \widetilde{\mathbb{Q}}_t(\mathbf{1} \otimes \mathbf{1}_A)$, for $A \subseteq \mathbb{X}$ and $t > 1$, is the blocked filter, with initial condition $\tilde{\pi}_1 := \widetilde{\mathbb{Q}}_1 = \pi_1$.

### C. *Model* $(\widehat{m}, \hat{p}_t, \hat{g}_t)$

The state-space model $(\widehat{m}, \hat{p}_t, \hat{g}_t)$ is defined by

$$\begin{aligned} \widehat{m}(z) &:= m_K(z_K)\varpi_1(z)/\varpi_{1,K}(z_K), \\ \hat{p}_t(x, z) &:= p_K(x_{\mathcal{N}(K)}, z_K)\varpi_t(z)/\varpi_{t,K}(z_K), \quad \text{for } t > 1, \\ \hat{g}_t(x) &:= g_{t,K}(x_K). \end{aligned}$$

Here, writing $K^{\mathrm{c}} := \mathbb{V} \setminus K$ we have defined $\varpi_t(x) := \frac{\mathrm{d}\pi_t}{\mathrm{d}\psi}(x)$ and $\varpi_{t,K}(x_K) := \int_{\mathbb{X}_{K^{\mathrm{c}}}} \varpi_t(x)\psi_{K^{\mathrm{c}}}(\mathrm{d}x_{K^{\mathrm{c}}})$. We let $\widehat{\mathbb{Q}}_t$, $\hat{\pi}_t$ and $\widehat{B}_{t,\hat{\pi}_t}(x, \mathrm{d}z) := \hat{p}_{t+1}(z, x)\hat{\pi}_t(\mathrm{d}z)/\hat{\pi}_t(\hat{p}_{t+1}(\,\cdot\,, x))$ be the joint smoothing distribution, filter and standard backward kernel of $(\widehat{m}, \hat{p}_t, \hat{g}_t)$.

*Lemma 1:*

1) *For any* $t \in \mathbb{N}$, $\|\hat{\pi}_t - \pi_t\|_{\mathbb{V}} = 0$.
2) *For any* $t \in \mathbb{N}$, $\|\widehat{\mathbb{Q}}_t - \mathbb{Q}_t\|_{\{1,\ldots,t\} \times K} = 0$.
3) *Under Assumption 2,*

$$\begin{aligned} \widehat{\mathbb{Q}}_T(F_T) = \int & f_{r,J}(x_{r,J})\pi_{T,K}(\mathrm{d}x_{T,K}) \\ & \times \prod_{t=1}^{T-1} B_{K,\pi_{t,\mathcal{N}(K)}}(x_{t+1,K}, \mathrm{d}x_{t,\mathcal{N}(K)}), \end{aligned}$$

*where* $B_{K,\nu}$ *is the blocked backward kernel from* (11)

The proof of Parts 1 and 2 of Lemma 1 follow by induction. Part 3 follows from Part 1 and the definition of $\widehat{B}_{t,\pi_t}$. For completeness, we detail these proofs in Subsection H of the supplementary materials.

### D. *Model* $(\overline{m}, \bar{p}_t, \bar{g}_t)$

The state-space model $(\overline{m}, \bar{p}_t, \bar{g}_t)$ is defined by

$$\begin{aligned} \overline{m}(z) &:= m_K(z_K)\widetilde{\varpi}_{1,K^{\mathrm{c}}}(z_{K^{\mathrm{c}}}), \\ \bar{p}_t(x, z) &:= p_K(x_{\mathcal{N}(K)}, z_K)\widetilde{\varpi}_{t,K^{\mathrm{c}}}(z_{K^{\mathrm{c}}}), \quad \text{for } t > 1, \\ \bar{g}_t(x) &:= g_{t,K}(x_K). \end{aligned}$$

Here, for any $K \subseteq \mathbb{V}$, we have defined $\widetilde{\varpi}_t(x) := \frac{\mathrm{d}\tilde{\pi}_t}{\mathrm{d}\psi}(x)$ and $\widetilde{\varpi}_{t,K^{\mathrm{c}}}(x_{K^{\mathrm{c}}}) := \int_{\mathbb{X}_K} \widetilde{\varpi}_t(x)\psi_K(\mathrm{d}x_K)$. We let $\overline{\mathbb{Q}}_t$, $\bar{\pi}_t$ and $\overline{B}_{t,\bar{\pi}_t}(x, \mathrm{d}z) := \bar{p}_{t+1}(z, x)\bar{\pi}_t(\mathrm{d}z)/\bar{\pi}_t(\bar{p}_{t+1}(\,\cdot\,, x))$ be the joint smoothing distribution, filter and standard backward kernel of $(\overline{m}, \bar{p}_t, \bar{g}_t)$.

*Lemma 2:*

1) *For any* $t \in \mathbb{N}$, $\|\bar{\pi}_t - \tilde{\pi}_t\|_{\mathbb{V}} = 0$.
2) *For any* $t \in \mathbb{N}$, $\|\overline{\mathbb{Q}}_t - \widetilde{\mathbb{Q}}_t\|_{\{1,\ldots,t\} \times K} = 0$.
3) *Under Assumption 2,*

$$\begin{aligned} \overline{\mathbb{Q}}_T(F_T) = \int & f_{r,J}(x_{r,J})\tilde{\pi}_{T,K}(\mathrm{d}x_{T,K}) \\ & \times \prod_{t=1}^{T-1} B_{K,\tilde{\pi}_{t,\mathcal{N}(K)}}(x_{t+1,K}, \mathrm{d}x_{t,\mathcal{N}(K)}), \end{aligned}$$

*where* $B_{K,\nu}$ *is the blocked backward kernel from* (11).

The proof of Lemma 2 is similar to the proof of Lemma 1. For completeness, we detail the proof in Subsection H of the supplementary materials.

### E. Central Limit Theorem

*Proof (of Proposition 2):* We first prove Part 2. By Lemma 1, we are performing standard particle smoothing for $(\widehat{m}, \hat{p}_t, \hat{g}_t)$. For any $t \le q \le T$ and any $x_t, z_t \in \mathbb{X}$, define

$$\mathcal{Q}_{t,q}(x_t, \mathrm{d}x_{t+1:T}) := \prod_{s=t+1}^{q} \hat{p}_s(x_{s-1}, x_s)\hat{g}_s(x_s)\psi(\mathrm{d}x_s),$$

$$D_{t,T}(z_t, \mathrm{d}x_{1:T}) := \delta_{z_t}(\mathrm{d}x_t)\Bigg[\prod_{s=1}^{t-1} \widehat{B}_{s,\hat{\pi}_s}(x_{s+1}, \mathrm{d}x_s)\Bigg] \times \mathcal{Q}_{t,T}(x_t, \mathrm{d}x_{t+1:T}).$$

Furthermore, we write $Q_{t,q}(f_q)(x_t) := \mathcal{Q}_{t,q}(\mathbf{1} \otimes f_q)(x_t)$, and also define $G_{t,T} := Q_{t,T}(\mathbf{1})/\hat{\pi}_t Q_{t,T}(\mathbf{1})$ as well as $P_{t,T}(F_T) := D_{t,T}(F_T)/D_{t,T}(\mathbf{1})$.

By [22, Theorem 3.1] (again transferred to updated models as described above) the blocked FS approximation of $\widehat{\mathbb{Q}}_T(F_T)$ is asymptotically normal with asymptotic variance

$$\sigma_T^2(F_T) \le \varepsilon^{4 \operatorname{card} K} \sum_{t=1}^{T} \hat{\pi}_t\big(P_{t,T}(F_T - \widehat{\mathbb{Q}}_T(F_T))^2\big),$$

since it is easy to check that $G_{t,T} \le \varepsilon^{2 \operatorname{card} K}$. Define the Markov kernels $R_{t,r} \colon \mathbb{X} \times \operatorname{Borel}(\mathbb{X}) \to [0, 1]$ by

$$R_{t,r}(f_r)(x_t) := \begin{cases} \delta_{x_t}(f_r) = f_r(x_t), & \text{if } t = r, \\ Q_{t,r}(f_r)(x_t)/Q_{t,r}(\mathbf{1})(x_t), & \text{if } t < r, \\ \int_{\mathbb{X}^{t-r}} f_r(x_r) \prod_{s=t-1}^{r} \widehat{B}_{s,\hat{\pi}_s}(x_{s+1}, \mathrm{d}x_s), & \text{if } t > r. \end{cases}$$

Let $\operatorname{osc}(f) := \sup_{(x,y)\in\mathbb{X}^2} |f(x) - f(y)|$ denote the oscillations of some function $f \colon \mathbb{X} \to \mathbb{R}$ and let $\operatorname{Osc}_1(\mathbb{X})$ denote the set of all real-valued functions with domain $\mathbb{X}$ whose oscillations are less than 1. Furthermore, let $\beta(M) := \sup_{f\in\operatorname{Osc}_1(\mathbb{X})} \operatorname{osc}(M(f)) \in [0, 1]$ denote the Dobrushin coefficient of a Markov kernel $M \colon \mathbb{X} \times \operatorname{Borel}(\mathbb{X}) \to [0, 1]$. Then

$$\hat{\pi}_t\big(P_{t,T}(F_T - \widehat{\mathbb{Q}}_T(F_T))^2\big) \le 4\beta(R_{t,r}), \tag{A.18}$$

where we have used that since $\|f_{r,J}\| \le 1$,

$$\operatorname{osc}\big((F_T - \widehat{\mathbb{Q}}_T(F_T))/2\big) = \operatorname{osc}\big((f_{r,J} - \widehat{\mathbb{Q}}_T(F_T))/2\big) \le 1.$$

For any $t \in \mathbb{N}$, we bound $\beta(R_{t,r})$ in (A.18) as follows:

$$\beta(R_{t,r}) \le (1 - \varepsilon^{-4 \operatorname{card} K})^{|t-r|}. \tag{A.19}$$

If $t = r$, (A.19) holds because $\beta(R_{t,r}) \le 1$. If $t < r$, (A.19) is implied by [22, Proposition 4.3.3]. Finally, it can be easily checked that $\widehat{B}_{t,\hat{\pi}_t}(x, \cdot\,) \ge \varepsilon^{-4 \operatorname{card} K} \widehat{B}_{t,\hat{\pi}_t}(z, \cdot\,)$, for any $(x, z) \in \mathbb{X}^2$, so that $\beta(\widehat{B}_{t,\hat{\pi}_t}) \le 1 - \varepsilon^{-4 \operatorname{card} K}$. Hence, if $t > r$, $\beta(R_{t,r}) \le \prod_{s=t-1}^{r} \beta(\widehat{B}_{s,\hat{\pi}_s}) \le (1 - \varepsilon^{-4 \operatorname{card} K})^{|t-r|}$. As a result,

$$\sum_{t=1}^{T} \beta(R_{t,r}) \le 2 \sum_{t=0}^{\infty} (1 - \varepsilon^{-4 \operatorname{card} K})^t = 2\varepsilon^{4 \operatorname{card} K},$$

so that $\sigma_T^2(F_T) \le 8\varepsilon^{8 \operatorname{card} K} = \mathrm{e}^{c_1 \operatorname{card} K} \le \mathrm{e}^{c_1 |\mathcal{K}|_\infty}$, with $c_1 := \log(8\varepsilon^8)$. This completes the proof of Part 2.

To prove Part 1, note that by Lemma 2, we are performing standard particle smoothing for $(\overline{m}, \bar{p}_t, \bar{g}_t)$. Part 2 is then proved exactly as Part 1 but with $(\widehat{m}, \hat{p}_t, \hat{g}_t, \widehat{\mathbb{Q}}_T, \hat{\pi}_t, \widehat{B}_{t,\hat{\pi}_t})$ replaced by $(\overline{m}, \bar{p}_t, \bar{g}_t, \overline{\mathbb{Q}}_T, \bar{\pi}_t, \overline{B}_{t,\bar{\pi}_t})$. □

### F. Asymptotic Bias

*Proof (of Proposition 3):* By Lemma 2 we are performing standard particle smoothing for $(\widetilde{m}, \tilde{p}_t, \tilde{g}_t)$. In particular,

$$\|\overline{\mathbb{Q}}_T - \mathbb{Q}_T\|_{\{t\}\times J} = \|\widetilde{\mathbb{Q}}_T - \mathbb{Q}_T\|_{\{t\}\times J},$$

for any $t \in \{1, \ldots, T\}$ and any $J \subseteq K$. Then by [25, Theorem 4.3] (which is stated only for the bias of the filter, i.e. for $t = T$, but whose proof is established for $t < T$),

$$\|\widetilde{\mathbb{Q}}_T - \mathbb{Q}_T\|_{\{t\}\times J} \le c_2 \operatorname{card}(J)\mathrm{e}^{-c_3\, d(J,\partial K)}.$$

This completes the proof. □

## References

[1] O. Cappé, E. Moulines, and T. Rydén, *Inference in Hidden Markov Models*, ser. Springer Series in Statistics. Springer, 2005.

[2] A. Doucet and A. M. Johansen, "A tutorial on particle filtering and smoothing: Fifteen years later," in *The Oxford Handbook of Nonlinear Filtering*, ser. Oxford Handbooks, D. Crisan and B. Rozovskii, Eds. Oxford University Press, 2011, ch. 24, pp. 656–704.

[3] P. Del Moral, A. Doucet, and A. Jasra, "Sequential Monte Carlo samplers," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 68, no. 3, pp. 411–436, 2006.

[4] C. Andrieu, A. Doucet, and R. Holenstein, "Particle Markov chain Monte Carlo methods," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 72, no. 3, pp. 269–342, 2010, with discussion.

[5] N. Chopin, P. E. Jacob, and O. Papaspiliopoulos, "SMC$^2$: an efficient algorithm for sequential analysis of state space models," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 75, no. 3, pp. 397–426, 2013.

[6] G. Poyiadjis, A. Doucet, and S. S. Singh, "Particle approximations of the score and observed information matrix in state space models with application to parameter estimation," *Biometrika*, vol. 98, no. 1, pp. 65–80, 2011.

[7] T. Bengtsson, P. Bickel, and B. Li, "Curse-of-dimensionality revisited: collapse of the particle filter in very large scale systems," in *Probability and statistics: Essays in honor of David A. Freedman*, D. Nolan and T. Speed, Eds. Institute of Mathematical Statistics, 2008, vol. 2, pp. 316–334.

[8] X. Boyen and D. Koller, "Tractable inference for complex stochastic processes," in *Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence*. Morgan Kaufmann Publishers, 1998, pp. 33–42.

[9] ——, "Exploiting the architecture of dynamic systems," in *Proceedings of the 16th National Conference on Artificial Intelligence*, 1999, pp. 313–320.

[10] B. Ng, L. Peshkin, and A. Pfeffer, "Factored particles for scalable monitoring," in *Proceedings of the 18th Conference on Uncertainty in Artificial Intelligence*. Morgan Kaufmann Publishers, 2002, pp. 370–377.

[11] B. C. Brandao, J. Wainer, and S. K. Goldenstein, "Subspace hierarchical particle filter," in *19th Brazilian Symposium on Computer Graphics and Image Processing*. IEEE, 2006, pp. 194–204.

[12] E. Besada-Portas, S. M. Plis, M. Jesus, and T. Lane, "Parallel subspace sampling for particle filtering in dynamic Bayesian networks," in *Machine Learning and Knowledge Discovery in Databases*. Springer, 2009, pp. 131–146.

[13] P. Rebeschini and R. van Handel, "Can local particle filters beat the curse of dimensionality?" *The Annals of Applied Probability*, vol. 25, no. 5, pp. 2809–2866, 2015.

[14] J. MacCormick and A. Blake, "A probabilistic exclusion principle for tracking multiple objects," *International Journal of Computer Vision*, vol. 39, no. 1, pp. 57–71, 2000.

[15] M. Wüthrich, J. Bohg, D. Kappler, C. Pfreundt, and S. Schaal, "The coordinate particle filter – a novel particle filter for high dimensional systems," in *2015 IEEE International Conference on Robotics and Automation (ICRA)*, May 2015, pp. 2454–2461.

[16] A. Beskos, D. Crisan, A. Jasra, K. Kamatani, and Y. Zhou, "A stable particle filter for a class of high-dimensional state-space models," *Advances in Applied Probability*, vol. 49, no. 1, pp. 24–48, 2017.

[17] P. Guarniero, A. M. Johansen, and A. Lee, "The iterated auxiliary particle filter," *Journal of the American Statistical Association*, 2016, to be published.

[18] X. Boyen and D. Koller, "Approximate learning of dynamic models," in *Proceedings of the Eleventh Conference on Advances in Neural Information Processing Systems*. MIT Press, 1999, p. 396.

[19] J. Ala-Luhtala, N. Whiteley, R. Piché, and S. Ali-Löytty, "Factorized approximations in high-dimensional hidden Markov models," n.d., in preparation.

[20] S. J. Godsill, A. Doucet, and M. West, "Monte Carlo smoothing for nonlinear time series," *Journal of the American Statistical Association*, vol. 99, no. 465, 2004.

[21] P. Del Moral, A. Doucet, and S. Singh, "Forward smoothing using sequential Monte Carlo," *ArXiv e-prints*, Dec. 2010.

[22] P. Del Moral, A. Doucet, and S. S. Singh, "A backward particle interpretation of Feynman–Kac formulae," *ESAIM: Mathematical Modelling and Numerical Analysis*, vol. 44, no. 5, pp. 947–975, 2010.

[23] R. Douc, A. Garivier, E. Moulines, and J. Olsson, "Sequential Monte Carlo smoothing for general state space hidden Markov models," *The Annals of Applied Probability*, vol. 21, no. 6, pp. 2109–2145, 2011.

[24] J. Olsson and J. Westerborn, "Efficient particle-based online smoothing in general hidden Markov models: the PaRIS algorithm," *Bernoulli*, vol. 23, no. 3, pp. 1951–1996, 2017.

[25] P. Rebeschini and R. van Handel, "Comparison theorems for Gibbs measures," *Journal of Statistical Physics*, vol. 157, no. 2, pp. 234–281, 2014.

[26] P. Del Moral, *Feynman-Kac Formulae: Genealogical and Interacting Particle Systems with Applications*. Springer, 2004.

[27] R. Douc, E. Moulines, and J. Olsson, "Long-term stability of sequential Monte Carlo methods under verifiable conditions," *The Annals of Applied Probability*, vol. 24, no. 5, pp. 1767–1802, 2014.

[28] N. Whiteley, "Stability properties of some particle filters," *The Annals of Applied Probability*, vol. 23, no. 6, pp. 2500–2537, 2013.

[29] A. Lee, C. Yau, M. B. Giles, A. Doucet, and C. C. Holmes, "On the utility of graphics cards to perform massively parallel simulation of advanced Monte Carlo methods," *Journal of Computational and Graphical Statistics*, vol. 19, no. 4, pp. 769–789, 2010.

[30] S. Das, D. Lawless, B. Ng, and A. Pfeffer, "Factored particle filtering for data fusion and situation assessment in urban environments," in *Seventh International Conference on Information Fusion*, vol. 2. IEEE, 2005, pp. 955–962.

[31] C. A. Naesseth, F. Lindsten, and T. B. Schön, "Nested sequential Monte Carlo methods," in *Proceedings of the 32nd International Conference on Machine Learning*, 2015.

[32] M. K. Pitt and N. Shephard, "Filtering via simulation: Auxiliary particle filters," *Journal of the American Statistical Association*, vol. 94, no. 446, pp. 590–599, 1999.

[33] A. M. Johansen and A. Doucet, "A note on auxiliary particle filters," *Statistics & Probability Letters*, vol. 78, no. 12, pp. 1498–1504, 2008.

[34] N. Kantas, A. Doucet, S. S. Singh, J. Maciejowski, and N. Chopin, "On particle methods for parameter estimation in state-space models," *Statistical Science*, vol. 30, no. 3, pp. 328–351, 2015.

**Axel Finke** received a B.S. (2010) in economics from Münster Univ., Germany and an M.S. (with distinction) and Ph.D. in statistics from Warwick Univ., UK in 2011 and 2015, respectively. Subsequently, he joined the Signal Processing and Communications Laboratory, Dept. of Engineering, Univ. of Cambridge, UK, as research associate. He is currently working as a research associate at the Dept. of Statistical Science, University College London, UK His research interests include computational statistics, in particular sequential Monte Carlo and Markov chain Monte Carlo methods for both Bayesian and Frequentist inference.

**Sumeetpal S. Singh** received the B.E. (with first-class honours) and Ph.D. degrees from the Dept. of Electrical Engineering, University of Melbourne, Australia, in 1997 and 2002, respectively. After having worked in Industry for a number of years, he joined the Cambridge University Engineering Department in 2004 as a Research Associate and is currently a Reader in Engineering Statistics. He is also a Fellow and Director of Studies at Churchill College, a Fellow of the Alan Turing Institute, an Affiliated Lecturer of the Statistics Laboratory and an Associate Editor of Statistics and Computing. His research focus is on statistical signal processing, in particular, using Monte Carlo methods, covering algorithmic development for applications and theoretical analysis. He has been recognised for his work on Multi-target Tracking (awarded the IEEE M. Barry Carlton Award in 2013.)

# Supplementary Materials: *Approximate Smoothing and Parameter Estimation in High-Dimensional State-Space Models*

Axel Finke, Sumeetpal S. Singh

In these supplementary materials, we first discuss ways in which the methods proposed in the main document may be extended to other models than those described in Subsection III-A in the main document. We also provide the proofs of Lemmata 1 and 2 as well as additional details about the simulation study conducted in Section V of the main document.

## *G. Extension to other models*

To develop dimensionally stable particle smoothers, it is necessary to exploit some specific model structure. Indeed, as in the case of particle filters, it seems unlikely that there exists a way of circumventing the curse of dimensionality which is directly applicable to *arbitrary* state-space models.

In this work, we exploit such structure by considering the canonical class of high-dimensional state-space models from [1] (reviewed in Section III-A of the main document). In this class of models, we can exploit the spatial-decorrelation to circumvent the curse of dimensionality suffered by particle *smoothers* via suitable blocking approximations (i.e. using the idea through which [1] devised dimensionally stable particle *filters*).

Of course, many existing high-dimensional state-space models do not exactly fit into the class of models from Subsection III-A in the main document (hereafter, we will simply refer to the latter as 'canonical models') but will nonetheless exhibit (or can be modified to exhibit) some spatial decorrelation. To perform smoothing for such models, we propose two options.

1) In some cases, it may be possible to devise blocked particle methods such that they still exploit spatial decorrelation in the spirit of [1] but can be implemented for the original model. This approach is discussed in Example 2 below, where we describe a way of modifying our blocked particle smoothers to accommodate a different class of model.
2) Alternatively, we advocate taking the bias–variance trade-off which is already at the heart of blocking approximations (see [1] and Subsection IV-E of the main document) one step further. That is, we advocate accepting slightly more model misspecification error (i.e. bias) in exchange for a practical (i.e. dimensionally stable) way of estimating the model. Example 3 below describes how learning two more parameters (one of which is a suitable neighbourhood size $R$) for the canonical model allows us to relax the assumption that the state transition of the fitted model fully factorise as specified in Subsection III-A of the main document. Our proposed methodology and analysis applies without modification in this case.

*Example 2 (modified algorithm): As before, let $X_t$ and $Y_t$ be $V$-dimensional vectors. Consider the multivariate stochastic volatility model defined through*

$$\begin{aligned} Y_t &= V_t \eta_t, \\ X_t &= A X_{t-1} + \epsilon_t, \end{aligned}$$

*where $V_t$ is a $V \times V$ diagonal matrix whose entries on the diagonal are given by the volatilities $\exp(X_{t,1}), \dots, \exp(X_{t,V})$. The joint error term $(\epsilon_t, \eta_t)$ is* multivariate Gaussian with covariance matrix

$$\begin{bmatrix} \Sigma_\epsilon & \Sigma_{\epsilon,\eta} \\ \Sigma_{\epsilon,\eta} & \Sigma_\eta \end{bmatrix}.$$

We assume that order of the assets, i.e. the order of the components of $X_t$ and $Y_t$, has been chosen in such a way that highly correlated assets have been grouped together (e.g. by applying some clustering technique to the data). This ensures spatial decorrelation by implying that $A$, $\Sigma_\epsilon$ and $\Sigma_\eta$ have little mass far away from the main diagonal. Note, however, that we do not assume that the entries off the main diagonal (or off the first few diagonals in the case of $A$) are zero so that the transition densities $p(x,z)$ and $g(z,y_t)$ (with respect to the usual dominating measure $\psi = \bigotimes_{v \in \mathbb{V}} \psi_v$ and $\varphi = \bigotimes_{v \in \mathbb{V}} \varphi_v$ for $\mathbb{V} := \{1, \dots, V\}$) do not factorise. As a result, the stochastic volatility model cannot be viewed as a special case of the canonical model.

The model does, however, satisfy the following weakened assumption.

Assumption 6: For any time $t$ and any block $K \subseteq \mathbb{V}$, it is possible to evaluate the marginal densities

a) $g_K(z, y_{t,K}) := \int g(z, y_t) \prod_{v \notin K} \varphi_v(\mathrm{d}y_{t,v})$,
b) $p_K(x, z_K) := \int p(x,z) \prod_{v \notin K} \psi_v(\mathrm{d}z_v)$.

Under Assumption 6a, we can approximate the filter via a modified version of the BPF (for simplicity, we only consider a bootstrap version, i.e. $q_t(x,z) = p(x,z)$). The modified BPF remains exactly as in Algorithm 4 of the main document except that the local weights associated with block $K$ are now calculated as

$$w^n_{t,K} := g_K(x^n_t, y_{t,K}). \tag{A.20}$$

Likewise, under Assumption 6b, we can perform smoothing via modified versions of blocked FS and blocked BS. The modified blocked particle smoothers remain exactly as Algorithms 5 and 6 of the main document, except that we replace the blocked backward kernels $B_{K,\pi^N_{t,\mathcal{N}(K)}}(x_K, \mathrm{d}z_{\mathcal{N}(K)})$ by

$$\sum_{n=1}^{N} \frac{W^n_{t,K} p(X^n_t, x_K)}{\sum_{k=1}^{N} W^k_{t,K} p(X^k_t, x_K)} \delta_{X^n_{t,K}}(\mathrm{d}z_K). \tag{A.21}$$

Note that these kernels now map from $K$ to $K$ (in a suitable sense) whereas Algorithms 5 and 6 (main document) used kernels that map from $K$ to $\mathcal{N}(K)$. Though if required by the test functions which we wish to integrate, we could easily make (A.21) map from $K$ to some $i$-neighbourhood $\mathcal{N}_i(K)$ by replacing $X^n_{t,K}$ and $W^n_{t,K}$ by $X^n_{t,\mathcal{N}_i(K)}$ and $W^n_{t,\mathcal{N}_i(K)}$ (computed via (A.20)) in (A.21). As usual, we may also replace $K \in \mathcal{K}$ by some enlarged block $\overline{K} \supseteq K$ to reduce bias.

Example 3 (modified model): Note that even though the canonical model assumes that spatial components interact only locally (in some $R$-neighbourhood) over a single time step, all spatial components interact with one another after sufficiently many time steps $L$, where $L$ is a function of the neighbourhood size $R$.

For instance, in the one-dimensional lattice example shown in Fig. 1 – where $R = 1$, i.e. $\mathcal{N}(v) = \{v-1, v, v+1\}$ – all spatial components interact over $L = V - 1$ steps.

*This motivates approximating a model with non-local spatial interactions (and observations $y_{1:T'}$ over $T'$ time steps) by a canonical model with only local interactions and with $T := LT'$ and where the latter takes*

$$g(x_t, y_t) := \begin{cases} g(x_t, y_t'), & \textit{for } t \in \{L, 2L, \dots, LT'\}, \\ 1, & \textit{otherwise.} \end{cases}$$

*Given the choice of approximating canonical model, our theoretical results immediately guarantee dimensional stability of the blocked particle smoothers as the latter can be applied without any change. In particular, the parameter-estimation algorithms from Subsection V-C of the main document can then be used to guide the selection of $R$ and $L$.*

### H. Additional proofs

In this subsection, we present the proofs of Lemma 1 and Lemma 2 which were omitted from the main document.

*Proof (of Lemma 1):* We prove Part 1 by induction. Clearly,

$$\begin{aligned} \hat{\pi}_1(\mathrm{d}z) &\propto \psi(\mathrm{d}z)\varpi_1(z) m_K(z_K) g_{1,K}(z_K)/\varpi_{1,K}(z_K) \\ &\propto \psi(\mathrm{d}z)\varpi_1(z) \propto \pi_1(\mathrm{d}z), \end{aligned}$$

since $\varpi_{1,K}(z_K) \propto m_K(z_K) g_{1,K}(z_K)$. Assume now that the statement holds at some time $t-1$. Then

$$\begin{aligned} \hat{\pi}_t(\mathrm{d}z) &\propto \int_{\mathbb{X}} \hat{p}_t(x, z)\hat{g}(z, y_t)\psi(\mathrm{d}z)\pi_{t-1}(\mathrm{d}x) \\ &= \frac{\int_{\mathbb{X}} p_K(x_{\mathcal{N}(K)}, z_K) g_{t,K}(z_K)\pi_{t-1}(\mathrm{d}x)}{\varpi_{t,K}(z_K)} \pi_t(\mathrm{d}z) \\ &\propto \pi_t(\mathrm{d}z), \end{aligned}$$

since $\varpi_{t,K}(z_K) \propto \int_{\mathbb{X}} p_K(x_{\mathcal{N}(K)}, z_K) g_{t,K}(z_K)\pi_{t-1}(\mathrm{d}x)$.

Part 2 is also proved by induction. By Part 1, since $\pi_1 = \mathbb{Q}_1$ and $\hat{\pi}_1 = \widehat{\mathbb{Q}}_1$, the statement holds at time $t = 1$. Assume now that the statement holds at some time $t-1$. Then for any $A \subseteq \mathbb{X}_K^t$, by Part 1,

$$\begin{aligned} &\int_{\mathbb{X}^t} \mathbf{1}_A(x_{1:t,K})\widehat{\mathbb{Q}}_t(\mathrm{d}x_{1:t}) \\ &\qquad \propto \int_{\mathbb{X}^t} \mathbf{1}_A(x_{1:t,K}) \frac{\varpi_t(x_t)}{\varpi_{t,K}(x_{t,K})} p_K(x_{t-1,\mathcal{N}(K)}, x_{t,K}) \\ &\qquad\quad \times g_{t,K}(x_{t,K})\mathbb{Q}_{t-1}(\mathrm{d}x_{1:t-1})\psi(\mathrm{d}x_t) \\ &\qquad \propto \int_{\mathbb{X}^t} \mathbf{1}_A(x_{1:t,K})\mathbb{Q}_t(\mathrm{d}x_{1:t}). \end{aligned}$$

To prove Part 3, note that for any $A \subseteq \mathbb{X}_{\mathcal{N}(K)}$, by Part 1,

$$\begin{aligned} &\int_{\mathbb{X}} \mathbf{1}_A(z_{\mathcal{N}(K)})\widehat{B}_{t,\hat{\pi}_t}(x, \mathrm{d}z) \\ &\qquad = \frac{\int_A p_K(z_{\mathcal{N}(K)}, x_K)\pi_{t,\mathcal{N}(K)}(\mathrm{d}z_{\mathcal{N}(K)})}{\int_{\mathbb{X}_{\mathcal{N}(K)}} p_K(u_{\mathcal{N}(K)}, x_K)\pi_{t,\mathcal{N}(K)}(\mathrm{d}u_{\mathcal{N}(K)})} \\ &\qquad = \int_A B_{K,\pi_{t,\mathcal{N}(K)}}(x_K, \mathrm{d}z_{\mathcal{N}(K)}). \end{aligned}$$

Noting that the above expression is constant in $x_{K^{\mathrm{c}}}$ then completes the proof of Part 3. □

*Proof (of Lemma 2):* The proof is immediate if we replace the quantities $(\widehat{m}, \hat{p}_t, \hat{g}_t, \widehat{\mathbb{Q}}_t, \hat{\pi}_t, \widehat{B}_{t,\hat{\pi}_t}, \pi_t, \mathbb{Q}_t)$ in the proof of Lemma 1 by the quantities $(\overline{m}, \bar{p}_t, \bar{g}_t, \overline{\mathbb{Q}}_t, \bar{\pi}_t, \overline{B}_{t,\bar{\pi}_t}, \tilde{\pi}_t, \widetilde{\mathbb{Q}}_t)$. □

### *I. Sufficient Statistics*

In this subsection, we derive the sufficient statistics needed for performing parameter estimation in the high-dimensional linear-Gaussian state-space model from Section V of the main document. For $q, r \in \{0, 1, \dots, R\}$, define

$$f_{1,v}^{\theta,(1,r,q)} = f_{1,v}^{\theta,(2,r)} \equiv 0,$$

and, for any $t \in \mathbb{N}$,

$$\begin{aligned} f_{t+1,v}^{\theta,(1,r,q)}(x_{t,\mathcal{N}(v)}, x_{t+1,v}) &:= \textstyle\sum_{u \in \mathcal{B}_q(v)} x_{t,u} \sum_{w \in \mathcal{B}_r(v)} x_{t,w}, \\ f_{t+1,v}^{\theta,(2,r)}(x_{t,\mathcal{N}(v)}, x_{t+1,v}) &:= x_{t+1,v} \textstyle\sum_{u \in \mathcal{B}_r(v)} x_{t,u}, \\ f_{t,v}^{\theta,(3)}(x_{t-1,\mathcal{N}(v)}, x_{t,v}) &:= x_{t,v}^2, \\ f_{t,v}^{\theta,(4)}(x_{t-1,\mathcal{N}(v)}, x_{t,v}) &:= x_{t,v} y_{t,v}. \end{aligned}$$

For any superscript $(\star)$ in the previous equation, we may then define the following sums of smoothed sufficient statistics

$$\mathbb{F}_T^{\theta,(\star)} := \sum_{t=1}^{T} \sum_{v \in \mathbb{V}} \mathbb{E}\big[f_{t,v}^{\theta,(\star)}(X_{t-1,\mathcal{N}(v)}, X_{t,v})\big], \quad X_{1:T} \sim \mathbb{Q}_T^{\theta}.$$

To simplify the notation, we define the following matrix and column vector

$$\begin{aligned} \mathbb{F}_T^{\theta,(1)} &:= (\mathbb{F}_T^{\theta,(1,r,q)})_{(r,q) \in \{0,\dots,R\}^2} \in \mathbb{R}^{(R+1)\times(R+1)}, \\ \mathbb{F}_T^{\theta,(2)} &:= (\mathbb{F}_T^{\theta,(2,l)})_{l \in \{0,\dots,R\}} \in \mathbb{R}^{R+1}, \end{aligned}$$

and finally, we collect all of these smoothed sufficient statistics in the ordered set

$$\mathbb{F}_T^{\theta} := (\mathbb{F}_T^{\theta,(1)}, \dots, \mathbb{F}_T^{\theta,(4)}). \tag{A.22}$$

### *J. Parameter Estimation*

In this subsection, we state the ways in which the update rules of (stochastic) gradient-ascent and EM algorithms depend on the smoothed sufficient statistics from (A.22). To simplify the notation, we write $y^2 := \sum_{t=1}^{T} \sum_{v \in \mathbb{V}} y_{t,v}^2$.

*1) Gradient-ascent Algorithm:* While the score may be directly written as an additive function as shown in Example 1 of the main document, we can alternatively write it as $\nabla_{\vartheta} \log \varGamma_T^{\vartheta}(\mathbf{1})|_{\vartheta=\theta} =: \varPsi^{\theta}(\mathbb{F}_T^{\theta})$. Here, for $T > 1$, writing

$$\varPsi^{\theta}(\mathbb{F}_T^{\theta}) = (\varPsi_r^{\theta}(\mathbb{F}_T^{\theta}))_{r \in \{0,\dots,R+2\}},$$

we have

$$\begin{aligned} \varPsi_{0:R}^{\theta}(\mathbb{F}_T^{\theta}) &= \mathrm{e}^{-2\theta_{R+1}} \big(\mathbb{F}_T^{\theta,(2)} - \mathbb{F}_T^{\theta,(1)} \theta_{0:R}\big), \\ \varPsi_{R+1}^{\theta}(\mathbb{F}_T^{\theta}) &= \mathrm{e}^{-2\theta_{R+1}} \big(\mathbb{F}_T^{\theta,(3)} - \mathbb{F}_1^{\theta,(3)} - 2\theta_{0:R}^{\intercal} \mathbb{F}_T^{\theta,(2)} \\ &\qquad + \theta_{0:R}^{\intercal} \mathbb{F}_T^{\theta,(1)} \theta_{0:R}\big) - V(T-1), \\ \varPsi_{R+2}^{\theta}(\mathbb{F}_T^{\theta}) &= \mathrm{e}^{-2\theta_{R+2}} \big(\mathbb{F}_T^{\theta,(3)} - 2\mathbb{F}_T^{\theta,(4)} + y^2\big) - VT. \end{aligned}$$

*2) EM Algorithm:* For (stochastic) EM algorithms, the vector (of length $R+3$) needed for the update rule in (14) in the main document,

$$\varLambda(\mathbb{F}_T^{\theta}) := (\varLambda_r(\mathbb{F}_T^{\theta}))_{r \in \{0,\dots,R+2\}},$$

is given by

$$\begin{aligned} \varLambda_{0:R}(\mathbb{F}_T^{\theta}) &= (\mathbb{F}_T^{\theta,(1)})^{-1} \mathbb{F}_T^{\theta,(2)}, \\ \varLambda_{R+1}(\mathbb{F}_T^{\theta}) &= \tfrac{1}{2} \log\big(\tfrac{1}{V(T-1)} \big[\mathbb{F}_T^{\theta,(3)} - \mathbb{F}_1^{\theta,(3)} \\ &\qquad - (\mathbb{F}_T^{\theta,(2)})^{\intercal} (\mathbb{F}_T^{\theta,(1)})^{-1} \mathbb{F}_T^{\theta,(2)}\big]\big), \\ \varLambda_{R+2}(\mathbb{F}_T^{\theta}) &= \tfrac{1}{2} \log\big(\tfrac{1}{VT} \big[\mathbb{F}_T^{\theta,(3)} - 2\mathbb{F}_T^{\theta,(4)} + y^2\big). \end{aligned}$$

Finally, we note that since this model is in the exponential family, the maximisation problem admits a closed-form solution.

### *K. Implementation Details*

All the computations in this paper were implemented in the programming language C++ using the Armadillo linear algebra library [2]. All the algorithms were called from the R programming language [3] using various Rcpp [4], [5] libraries.

## References

[1] P. Rebeschini, Patrick and R. van Handel, “Can local particle filters beat the curse of dimensionality?”, The Annals of Applied Probability, vol. 25, no. 5, pp. 2809–2866.

[2] C. Sanderson and R. Curtin, “Armadillo: a template-based C++ library for linear algebra,” Journal of Open Source Software, vol. 1, p. 26, 2016.

[3] R Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, 2015. [Online]. Available: http://www.R-project.org/

[4] D. Eddelbuettel and R. Franois, “Rcpp: Seamless R and C++ integration,” Journal of Statistical Software, vol. 40, no. 8, pp. 1–18, 2011. [Online]. Available: http://www.jstatsoft.org/v40/i08/

[5] D. Eddelbuettel, Seamless R and C++ Integration with Rcpp. Springer, 2013.