Chin. Ann. Math. Ser. B
40(5), 2019, 811–842
DOI: 10.1007/s11401-019-0161-5

Chinese Annals of Mathematics, Series B

Ergodicity and Accuracy of Optimal Particle Filters for Bayesian Data Assimilation*

David KELLY[1] Andrew M. STUART[2]

Abstract Data assimilation refers to the methodology of combining dynamical models and observed data with the objective of improving state estimation. Most data assimilation algorithms are viewed as approximations of the Bayesian posterior (filtering distribution) on the signal given the observations. Some of these approximations are controlled, such as particle filters which may be refined to produce the true filtering distribution in the large particle number limit, and some are uncontrolled, such as ensemble Kalman filter methods which do not recover the true filtering distribution in the large ensemble limit. Other data assimilation algorithms, such as cycled 3DVAR methods, may be thought of as controlled estimators of the state, in the small observational noise scenario, but are also uncontrolled in general in relation to the true filtering distribution. For particle filters and ensemble Kalman filters it is of practical importance to understand how and why data assimilation methods can be effective when used with a fixed small number of particles, since for many large-scale applications it is not practical to deploy algorithms close to the large particle limit asymptotic. In this paper, the authors address this question for particle filters and, in particular, study their accuracy (in the small noise limit) and ergodicity (for noisy signal and observation) without appealing to the large particle number limit. The authors first overview the accuracy and minorization properties for the true filtering distribution, working in the setting of conditional Gaussianity for the dynamics-observation model. They then show that these properties are inherited by optimal particle filters for any fixed number of particles, and use the minorization to establish ergodicity of the filters. For completeness we also prove large particle number consistency results for the optimal particle filters, by writing the update equations for the underlying distributions as recursions. In addition to looking at the optimal particle filter with standard resampling, they derive all the above results for (what they term) the Gaussianized optimal particle filter and show that the theoretical properties are favorable for this method, when compared to the standard optimal particle filter.
Keywords Particle filters, Data assimilation, Ergodic theory
2000 MR Subject Classification 60G35, 62M20, 37H99, 60F99

1 Introduction

1.1 Background and literature review

Data assimilation describes the blending of dynamical models with data, with the objective of improving state estimation and forecasts. The use of data assimilation originated in the

Manuscript received November 6, 2018.
[1]Courant Institute of Mathematical Sciences, New York University, New York, USA; The Voleon Group, 2150 Dwight Way, Berkeley, CA94704, USA. E-mail: dtbkelly@gmail.com
[2]Department of Computing and Mathematical Sciences, California Institute of Technology, Pasadena, CA91125, USA. E-mail: astuart@caltech.edu
*This work was supported by EPSRC (EQUIPS Program Grant), DARPA (contract W911NF-15-2-0121) and ONR (award N00014-17-1-2079).

geophysical sciences, but is now ubiquitous in engineering and the applied sciences. In numerical weather prediction, large scale ocean-atmosphere models are assimilated with massive data sets, comprising observational data from satellites, ground based weather stations and underwater sensors for example [3]. Data assimilation is prevalent in robotics; the SLAM problem seeks to use sensory data made by robots in an unknown environment to create a map of that environment and locate the robot within it in [46]. It is used in modelling of traffic flow in [50]. And data assimilation is being used in bio-medical applications such as glucose-insulin systems in [41] and the sleep cycle in [42]. These examples serve to illustrate the growth in the use of the methodology, its breadth of applicability and the very different levels of fidelity present in the models and the data in these many applications.

Although typical data assimilation problems can be understood from a Bayesian perspective, for non-linear and potentially high dimensional models it is often infeasible to make useful exact computations with the posterior. To circumvent this problem, practitioners have developed assimilation methods that approximate the true posterior, but for which computations are more feasible. In the engineering communities, particle filters have been developed for this purpose, providing empirical approximations of non-Gaussian posteriors in [12–13]. In the geoscience communities, methods are typically built on Kalman filtering theory, after making suitable Gaussian approximations in [26]; such methods include variational methods like 3DVAR and 4DVAR in [16, 33], the extended Kalman filter (ExKF for short) in [14] and the ensemble Kalman filter (EnKF for short) in [6, 15]. For these methods the underlying Gaussian ansatz render them, in general, invalid as approximations of the true filtering distribution in [29].

Despite their widespread use, many of these algorithms used in geophysical applications remain mysterious from a theoretical perspective. At the heart of the mystery is the fact that data assimilation methods are frequently and successfully implemented in regimes where the approximate filter is not provably valid; it is not known which features of the posterior (the true filtering distribution) are reflected in the approximation and which are not. For example, the ensemble Kalman filter is often implemented with ensemble size several orders of magnitude smaller than needed to reproduce large sample size behaviour, and is applied to problems for which the Gaussian ansatz may not be valid; it nonetheless can still exhibit skillful state estimates, with high correlations between the estimate and true trajectories in [19, 34]. Indeed, the success of the methods in this non-asymptotic regime is the crux of their success; the methods would often be computationally intractable at large ensemble sizes.

This lack of theory has motivated recent efforts to better understand the properties of data assimilation methods in the practical, non-asymptotic regimes. The 3DVAR algorithm has been investigated in the context of toy models for numerical weather prediction, including the Lorenz-63 in [28], Lorenz-96 in [27] and $2d$ Navier-Stokes equations in [4] (see also [36]). These works focus primarily on the question of accuracy – how well does the state estimate track the true underlying signal. Accuracy for the EnKF with fixed ensemble size was first investigated in [23]; the study of accuracy for the EnKF was further developed in [35] using linear models with random coefficients, but much more realistic (practical) assumptions on observations than [23], and moreover focussing on covariance consistency through the Mahalanobis norm. The articles in [48–49] were the first to investigate the stability of EnKF with fixed ensemble size, by formulating the filter as a Markov chain and applying coupling techniques; here by stability

we mean robustness with respect to initialization, and study the issue through the lens of ergodicity. This line of research has been continued in [9] by framing the EnKF as a McKean-Vlasov system. The limitations of the non-practical regimes have also been investigated; in [21] the authors construct simple dissipative dynamical models for which the EnKF is shown to be highly unstable with respect to initial perturbations. This was the first theoretical insight into the frequently observed effect of catastrophic filter divergence [18].

For the nonlinear filtering distribution itself, there has been a great deal of research over the last several decades, particularly on the question of stability. Conditional ergodicity for the filtering distribution for general nonlinear hidden Markov models has been investigated in [25] and later refined in [17]. Ergodicity for nonlinear filters has been discussed in [8, 10, 23] and exponential convergence results were first obtained in [2, 5].

1.2 Our contributions

For particle filters, much of the theoretical literature focuses on the question of consistency in the large ensemble limit, that is, does the empirical approximation converge to the true posterior as the number of particles in the ensemble N approaches infinity. However, in many high dimensional applications such as robotics in [46] and ocean-atmosphere forecasting in [30], particle filters are implemented in the non-asymptotic regime. Indeed in the geosciences, new filtering algorithms have been proposed to beat the curse of dimensionality and are implemented with ensemble sizes many orders of magnitude smaller than the state dimension in [31]. In this article we contribute to the program of analyzing filtering algorithms in practical small ensemble regimes, focusing on the accuracy and stability of particle filters for fixed ensemble sizes. In particular, we address the following question concerning the long-time behaviour of the particle filters: If it is known that the true posterior distribution is accurate and satisfies a minorization condition, can accuracy and conditional ergodicity be proved for the approximate filter?

We focus our attention on the optimal particle filter (OPF for short) in [1, 32, 51]. The OPF is a sequential importance sampling procedure in which particle updates are proposed using a convex combination of the model prediction and the observational data at the next time step. For details on the OPF, including the justification for calling it optimal (see [13]). There are two main reasons that we focus our attention on the OPF. First, the optimal particle filter is known to compare favorably to the standard particle filter, particularly from the perspective of weight degeneracy in high dimensions in [43–44]. Indeed the optimal particle filter can be considered a special case of more complicated filters that have been proposed to beat the curse of dimensionality in [7, 31]. Secondly, under natural assumptions on the dynamics-observation model, the optimal particle filter can be formulated as a random dynamical system which is very similar to the 3DVAR algorithm. This means that techniques for proving accuracy for the 3DVAR filter in earlier literature in [40] can be leveraged for the OPF.

Throughout the article, we make the assumption of conditional Gaussianity for the dynamics-observation model. This framework is frequently employed in practice, particularly in geoscience data assimilation problems. Under this assumption, we show that the true posterior, the filtering distribution, satisfies the properties of long-time accuracy and of minorization. The accuracy result states that, if sufficiently many variables are observed, the posterior will concentrate around the true trajectory in the long time limit. The minorization result shows that

the transition kernel of the nonlinear Markov process generating the filtering distribution is bounded below, uniformly in time, by a time-dependent multiple of a fixed time-independent probability measure. Minorization may be used to prove coupling and ergodicity in linear Markov processes and this fact will be exploited when we study particle filters. Conditional ergodicity results exploiting coupling are obtained under quite general assumptions in [17, 47].

Having introduced concepts in the context of the filtering distribution itself, we go on to show that, under the same model-observation assumptions, the OPF exhibits the long-time properties of conditional ergodicity and accuracy for any fixed ensemble size. For the conditional ergodicity result, we show that two copies of the particle ensembles, initialized differently, but updated with the same observational data, will converge to each other in the long term limit, in a distributional sense. The accuracy result shows that all ensemble members in the particle filter will concentrate near the true signal underlying the data, in the large-time regime. Both the accuracy and ergodicity results use very similar arguments to those employed for the analysis of the filtering distribution itself. In recent work, ergodicity has been used to study the long-run asymptotic behaviour of particle based optimization algorithms in [45], with motivation taken from parameter estimation in partially observed dynamical systems in [38].

In addition, we also establish large ensemble consistency results for the OPF. Here we employ a technique exposed very clearly in [39], which finds a recursion that is approximately satisfied by the bootstrap particle filter, and leverages this fact to obtain an estimate on the distance between the true posterior and the empirical approximation. We show that the same idea can be applied to not only the OPF, but a very large class of sequential importance sampling procedures. We note that large particle consistency results for particle filters should not be considered practical results for high dimensional data assimilation problems, as in practice particle filters are never implemented in this regime. The consistency results are included here as they are practically informative for low dimensional data assimilation problems and moreover as the results are natural consequences of the random dynamical system formulation that has been adopted for accuracy and ergodicity results. For high dimensional data assimilation problems, it may be more practical to look at covariance consistency, as done in [35]. We also note that quite general consistency results are proved for optimal particle filters in [20], using the setting of auxiliary particle filters introduced in [37]; this work was taken further in [11].

As well as obtaining results concerning the stability, accuracy and consistency for the OPF, for which we perform resampling at the end of each assimilation cycle, we also prove the corresponding results for the so called Gaussianized OPF. The terminology Gaussianized OPF was introduced in [22], but the idea was introduced two decades ago in [37] in the context of the auxiliary particle filter (see [32]). The method differs from the OPF only in the implementation of the resampling. Nevertheless, it was shown numerically in [22] that the GOPF compares favorably to the OPF, particularly when applied to high dimensional models. The analysis in this article lends theoretical weight to the advantages of the GOPF over the OPF. In particular we find that the upper bound on the convergence rate for conditional ergodicity for the GOPF has favourable dimension dependence in comparison with the OPF.

1.3 Structure of article and notation

The remainder of the article is structured as follows. At the end of this section we introduce

some notation and terminology that will be useful in the sequel. In Section 2, we formulate the Bayesian problem of data assimilation, introduce the model-observation assumptions under which we work, and prove the accuracy and minorization results for the true posterior. In Section 3, we introduce the bootstrap particle filter, optimal particle filter and Gaussianized optimal particle filter. In Section 4, we prove the conditional ergodicity results for the optimal particle filters. In Section 5, we prove the accuracy results for the optimal particle filters. Finally, in Section 6, we prove the consistency results for the optimal particle filters.

Throughout we let $\mathcal{X}$ denote the finite dimensional Euclidean state space and and we let $\mathcal{Y}$ denote the finite dimensional Euclidean observation space. We write $\mathcal{M}(\mathcal{X})$ for the set of probability measures on $\mathcal{X}$. We denote the Euclidean norm on $\mathcal{X}$ by $|\cdot|$ and for a symmetric positive definite matrix $A \in L(\mathcal{X}, \mathcal{X})$, we define $|\cdot|_A = |A^{-\frac{1}{2}} \cdot |$. The notation $P(a \mid b)$ will denote the density of random variable a, conditioned on known b. Transition kernels $q(a, \cdot)$ will denote transitions from point a, being measures when the second argument is a Borel set in $\mathcal{X}$, and being densities when the second argument is an element of $\mathcal{X}$. The kernel q will also be appended with subscript k when emphasizing that it is inhomogeneous in time. A similar notation will be used for kernels q_k on space $\mathcal{X}^N$. Superscript k will be used in q^k to denote $k-$fold composition of the kernels. Finally we define $S^N : \mathcal{M}(\mathcal{X}) \to \mathcal{M}(\mathcal{X})$ to be the sampling operator $S^N\mu = \frac{1}{N}\sum_{n=1}^{N} \delta_{u^{(n)}}$ where $u^{(n)} \sim \mu$ are i.i.d. random variables.

2 Bayesian Data Assimilation

We describe the set-up which encompasses all the work in this paper, and then study the minorization condition and accuracy for the true filtering distribution.

2.1 Set-up

The state model is taken to be a discrete time Markov chain $\{u_k\}_{k\geq 0}$ taking values in the state space $\mathcal{X}$. We assume that the initial condition u_0 of the chain is distributed according to μ_0, where $\mu_0 \in \mathcal{M}(\mathcal{X})$. The transition kernel for the Markov chain is given by $P(u_{k+1} \mid u_k)$. For each $k \geq 1$, we make an observation of the Markov chain

$$y_{k+1} = h(u_{k+1}) + \eta_{k+1} \ , \tag{2.1}$$

where $h : \mathcal{X} \to \mathcal{Y}$ maps the state space to the observation space, and $\eta_k \sim N(0, \Gamma)$ are centred i.i.d. random variables representing observational noise. We denote by $Y_k = (y_1, \cdots, y_k)$ the accumulated observational data up to time k. We are interested in studying and approximating the filtering distribution $\mu_k(\cdot) = \mathbb{P}(u_k \in \cdot \mid Y_k)$ for all $k \geq 1$. We will denote the density of μ_k by $P(u_k \mid Y_k)$.

Although we will not use it for all of the consistency results at the end of the paper, for the ergodicity and accuracy results we will always require the following additional assumptions on the dynamics-observation model.

Assumption 2.1 Let $\psi : \mathcal{X} \mapsto \mathcal{X}$ be bounded. The dynamics-observation model is given

by

$$u_{k+1} = \psi(u_k) + \xi_k, \tag{2.2a}$$

$$y_{k+1} = Hu_{k+1} + \eta_{k+1}, \tag{2.2b}$$

where $u_0 \sim \mu_0, \xi_k \sim N(0,\Sigma)$ i.i.d., $\eta_k \sim N(0,\Gamma)$ i.i.d. and $u_0, \{\xi_k\}$ and $\{\eta_k\}$ are independent. We write $\Sigma = \sigma^2\Sigma_0$ and $\Gamma = \gamma^2\Gamma_0$ and require that Σ_0 and Γ_0 are strictly positive-definite, and that $\sigma, \gamma \in (0,\infty)$ so that $r := \sigma/\gamma \in (0,\infty)$.

For most of the results in this article we will be interested in properties of the conditional distributions $P(u_k \mid Y_k)$, and particle approximations of it, when the observational data Y_k is generated by a fixed realization of the model. For this reason, we introduce the following notation to emphasize that we are considering a fixed realization of the data, generated by a fixed trajectory of the underlying dynamical system.

Assumption 2.2 Fix $u_0^\dagger \in \mathcal{X}$ and positive semi-definite matrices Σ_* and Γ_* on $\mathcal{X}$ and $\mathcal{Y}$ respectively. Let $\{u_k^\dagger\}$ be a realization of the dynamics satisfying

$$u_{k+1}^\dagger = \psi(u_k^\dagger) + r\gamma\xi_k^\dagger,$$

where $u_0^\dagger \in \mathcal{X}$ is fixed and $\xi_k^\dagger \sim N(0,\Sigma_*)$ i.i.d. Similarly define $\{y_k^\dagger\}$ by

$$y_{k+1}^\dagger = Hu_{k+1}^\dagger + \gamma\eta_{k+1}^\dagger, \tag{2.3}$$

where $\eta_{k+1}^\dagger \sim N(0,\Gamma_*)$ i.i.d. and $\{\xi_k\}, \{\eta_k\}$ are independent. We will refer to $\{u_k^\dagger\}_{k\geq 0}$ as the true signal and $\{y_k^\dagger\}_{k\geq 1}$ as the given fixed data. As above, we use the shorthand $Y_k^\dagger = \{y_i^\dagger\}_{i=1}^k$.

Remark 2.1 Note that this data is not necessarily generated from the same statistical model used to define the filtering distribution both since $r^2\gamma^2\Sigma_*$ and $\gamma^2\Gamma_*$ may differ from Σ and Γ, and since the initial condition is fixed. However the covariance structures match if we define $\sigma := r\gamma$, $\Sigma_* = \Sigma_0$ and $\Gamma_* = \Gamma_0$. When studying accuracy, we will consider families of data sets and truths parameterized by $\gamma \to 0$; in this setting it is natural to think of r and the noise sequences $\{\xi_k^\dagger\}_{k\geq 0}$ and $\{\eta_k^\dagger\}_{k\geq 0}$ as fixed, whilst the true signal and fixed data sequences will depend on the value of γ.

The filtering distribution evolves according to the iteration

$$\mu_{k+1} = L_{k+1}P\mu_k, \tag{2.4}$$

where P and L_{k+1} are maps on the space of measures defined as follows. The linear map $P : \mathcal{M}(\mathcal{X}) \to \mathcal{M}(\mathcal{X})$ is the Markov semigroup

$$P\nu(A) = \int_A P(u_{k+1} \mid u_k)\nu(\mathrm{d}u_k).$$

We define the nonlinear likelihood multiplication operator $L_{k+1} : \mathcal{M}(\mathcal{X}) \to \mathcal{M}(\mathcal{X})$ by

$$L_{k+1}\nu(A) = \frac{\int_A P(y_{k+1} \mid u_{k+1})\nu(\mathrm{d}u_{k+1})}{\int_{\mathcal{X}} P(y_{k+1} \mid u_{k+1})\nu(\mathrm{d}u_{k+1})} \tag{2.5}$$

for each $A \subset \mathcal{X}$ measurable. Equation (2.4) simply represents application of Bayes' formula with prior $P\mu_k$ and likelihood $P(y_{k+1} \mid u_{k+1})$.

Expressed in terms of densities (2.4) becomes

$$\begin{aligned} P(u_{k+1} \mid Y_{k+1}) &= P(u_{k+1} \mid Y_k, y_{k+1}) \\ &= \frac{1}{P(y_{k+1} \mid Y_k)} P(y_{k+1} \mid u_{k+1}, Y_k) P(u_{k+1} \mid Y_k) \\ &= \int_{\mathcal{X}} \frac{1}{P(y_{k+1} \mid Y_k)} P(y_{k+1} \mid u_{k+1}, u_k, Y_k) P(u_{k+1} \mid u_k, Y_k) P(u_k \mid Y_k) \mathrm{d}u_k \\ &= \int_{\mathcal{X}} \frac{1}{P(y_{k+1} \mid Y_k)} P(y_{k+1} \mid u_{k+1}) P(u_{k+1} \mid u_k) P(u_k \mid Y_k) \mathrm{d}u_k. \end{aligned} \tag{2.6}$$

Thus we may write

$$P(u_{k+1} \mid Y_{k+1}) = \int_{\mathcal{X}} q_{k+1}(u_k, u_{k+1}) P(u_k \mid Y_k) \mathrm{d}u_k, \tag{2.7}$$

where the transition density q_{k+1} has the form

$$q_{k+1}(u_k, u_{k+1}) = \frac{1}{Z} \exp\Big(-\frac{1}{2} |y_{k+1} - H u_{k+1}|_\Gamma^2 - \frac{1}{2} |u_{k+1} - \psi(u_k)|_\Sigma^2 \Big). \tag{2.8}$$

If we define

$$q_{k+1}^0(u_k, u_{k+1}) = \exp\Big(-\frac{1}{2} |y_{k+1} - H u_{k+1}|_\Gamma^2 - \frac{1}{2} |u_{k+1} - \psi(u_k)|_\Sigma^2 \Big),$$

then we see that

$$Z = \int_{\mathcal{X}} \int_{\mathcal{X}} q_{k+1}^0(u_k, u_{k+1}) P(u_k \mid Y_k) \mathrm{d}u_k \mathrm{d}u_{k+1}. \tag{2.9}$$

Note the inhomogeneous and nonlinear nature of the Markov chain reflected in the fact that Z depends on y_{k+1}, u_k, and on $P(u_k \mid Y_k)$. Despite this dependence, the normalization constant Z can be bounded above independently of k. To see this note that (2.9) gives

$$Z \le \sqrt{(2\pi)^d \det \Sigma} \int_{\mathcal{X}} \int_{\mathcal{X}} \frac{1}{\sqrt{(2\pi)^d \det \Sigma}} \exp\Big(-\frac{1}{2} |u_{k+1} - \psi(u_k)|_\Sigma^2 \Big) P(u_k \mid Y_k) \mathrm{d}u_k \mathrm{d}u_{k+1}.$$

Integrating over u_{k+1} first, and then over u_k, gives

$$Z \le \sqrt{(2\pi)^d \det \Sigma}. \tag{2.10}$$

In the next two subsections we state two theorems concerning the minorization conditions and accuracy of the filtering distribution itself, followed by a subsection which elaborates the connection between the optimal particle filter and the 3DVAR algorithm. The remainder of the paper is devoted to establishing analogous results for various particle filters.

2.2 Minorization condition

The result of this subsection is as following.

Theorem 2.1 *Consider the filtering distributions μ_k under Assumption* 2.1. *Assume moreover that the observational data used to define the filtering distribution is given by $\{y_k^\dagger\}_{k\geq 1}$ from Assumption* 2.2. *Consider the transition kernel $q_{k+1}(u_k,\cdot)$ viewed as a random measure, parameterized by the random observational data. Then there is a probability measure $\mathbb{Q}\in\mathcal{M}(\mathcal{X})$ and a sequence of random constants $\epsilon_k>0$, defined by the observational data, such that, for all Borel sets A in $\mathcal{X}$,*

$$q_{k+1}(u,A)\geq \epsilon_k \mathbb{Q}(A). \tag{2.11}$$

Proof Recall Assumption 2.1 and define $\rho_{k,0}^\dagger = \sigma H\xi_k^\dagger + \gamma\eta_{k+1}^\dagger$ where $\xi_k^\dagger\sim N(0,\Sigma_*)$ i.i.d. , $\eta_k^\dagger\sim N(0,\Gamma_*)$ i.i.d. Recalling (2.8), (2.10) we obtain the lower bound

$$\begin{aligned}
\sqrt{(2\pi)^d\det\Sigma}\; q_{k+1}(u,\mathrm{d}v) &= \frac{\sqrt{(2\pi)^d\det\Sigma}}{Z}\exp\Big(-\frac{1}{2}|y_{k+1}-Hv|_\Gamma^2-\frac{1}{2}|v-\psi(u)|_\Sigma^2\Big)\mathrm{d}v\\
&\geq \exp\Big(-\frac{1}{2}|H\psi(u_k^\dagger)+\rho_{k,0}^\dagger-Hv|_\Gamma^2-\frac{1}{2}|v-\psi(u)|_\Sigma^2\Big)\mathrm{d}v\\
&\geq \exp(-2|H\psi(u_k^\dagger)|_\Gamma^2-|\psi(u)|_\Sigma^2-2|\rho_{k,0}^\dagger|_\Gamma^2-|Hv|_\Gamma^2-|v|_\Sigma^2)\mathrm{d}v\\
&\geq \exp(-\lambda^2-2|\rho_{k,0}^\dagger|_\Gamma^2)\exp\Big(-\frac{1}{2}|v|_D^2\Big)\mathrm{d}v,
\end{aligned}\tag{2.12}$$

where

$$\lambda^2=\sup_{u,v}(2|H\psi(v)|_\Gamma^2+|\psi(u)|_\Sigma^2) \tag{2.13}$$

and

$$\frac{1}{2}D^{-1}=\Sigma^{-1}+H^*\Gamma^{-1}H\ .$$

Thus we have a minorization condition of the form (2.11) where $\mathbb{Q}(\cdot)$ is the Gaussian $N(0,D)$ and the data-dependent random constants ϵ_k are given by

$$\epsilon_k=\frac{\sqrt{\det D}}{\sqrt{\det\Sigma}}\exp(-\lambda^2-2|\rho_{k,0}^\dagger|_\Gamma^2).$$

Remark 2.2 For linear Markov processes the existence of a minorization condition leads in a straightforward way to ergodicity via coupling arguments. In these arguments a new Markov chain, equivalent in law to the original one, is used; in this new Markov chain moves are made according to kernel $\mathbb{Q}$ with probability ϵ_k, being governed by a Bernoulli process. In the linear case it is possible to fix a realization of the Bernoulli process and then average with respect to it and obtain a process equivalent to the original one. This facilitates coupling. For nonlinear Markov processes fixing the Bernoulli process and then averaging does not lead to a process equivalent to the original one and so coupling arguments are more complex (see [10]). The filtering distribution is governed by a nonlinear Markov process and hence we do not prove ergodicity. However for the particle filters studied in later sections we work with linear Markov processes governing the particle ensemble; the proof of minorization is structurally similar to that in the preceding theorem and this is why we include the preceding theorem here.

2.3 Accuracy

We now discuss accuracy of the posterior filtering distribution in the small noise limit $\gamma \ll 1$. The assumptions are somewhat restrictive, but give a flavour of what can be achieved; more careful use of ideas from control theory, such as observability, detectability and stabilizibility may lead to improved results, similar in flavour. The result uses the 3DVAR filter as an upper bound, and in the next subsection we will also show that the 3DVAR filter connects very naturally with the filtering distribution itself, and with the optimal particle filter in the next section of the paper.

Assumption 2.3 Let $r = \frac{\sigma}{\gamma}$ and assume that there is $r_c > 0$ such that, for all $r \in [0, r_c)$, the function $(I - KH)\psi(\cdot)$, with K defined through (2.17) and (2.19), is globally Lipschitz on $\mathcal{X}$ with respect to the norm $\|\cdot\|$ with Lipschitz constant $\alpha = \alpha(r) < 1$.

Theorem 2.2 *Suppose Assumptions* 2.1, 2.3 *hold for some* $r_c > 0$. *Then for all* $r \in [0, r_c)$ *we have*

$$\limsup_{k\to\infty} \mathbb{E}\|u_k - \mathbb{E}^{\mu_k} u_k\|^2 \le c\gamma^2 ,$$

where $\mathbb{E}^{\mu_k}$ *denotes expectation with respect to measure* μ_k *defined through* (2.4) *and* $\mathbb{E}$ *denotes expectation over the dynamical model and the observational data.*

Proof This follows similarly to Corollary 4.3 in [40], using the fact that the mean of the filtering distribution is optimal in the sense that

$$\mathbb{E}\|u_k - \mathbb{E}^{\mu_k} u_k\|^2 \le \mathbb{E}\|u_k - m_k\|^2$$

for any Y_k-measurable sequence $\{m_k\}$. We use for m_k the 3DVAR filter

$$m_{k+1} = (I - KH)\psi(m_k) + Ky_{k+1}.$$

Let $e_k = u_k - m_k$. Following closely Theorem 4.10 of [26] we obtain

$$\mathbb{E}\|e_{k+1}\|_{k+1}^2 \le \alpha^2 \mathbb{E}\|e_k\|^2 + \mathcal{O}(\gamma^2).$$

Application of the Gronwall lemma, plus use of the optimality property, gives the required bound.

2.4 Connection with 3DVAR

In the previous subsection we used 3DVAR as a test function to upper bound the error in the true filtering distribution. Here we further develop connections with 3DVAR with an eye on the formulation of the optimal particle filter as a random dynamical system. Consider the general filtering distribution. Application of Bayes' formula in the form

$$q_{k+1}(u_k, \mathrm{d}u_{k+1}) \propto P(y_{k+1}|u_{k+1})P(u_{k+1}|u_k)\mathrm{d}u_{k+1} \tag{2.14}$$

gives

$$q_{k+1}(u_k, \mathrm{d}u_{k+1}) \propto \exp\Big(-\frac{1}{2}|y_{k+1} - h(u_{k+1})|_\Gamma^2 - \frac{1}{2}|u_{k+1} - \psi(u_k)|_\Sigma^2\Big)\mathrm{d}u_{k+1},$$

initialized at the measure μ_0. Assumption 2.1, namely the linearity of the observation operator, introduces a key simplification in this expression: we obtain

$$q_{k+1}(u_k, \mathrm{d}u_{k+1}) \propto \exp\Big(-\frac{1}{2}|y_{k+1} - Hu_{k+1}|^2_\Gamma - \frac{1}{2}|u_{k+1} - \psi(u_k)|^2_\Sigma\Big)\mathrm{d}u_{k+1} \tag{2.15}$$

and a simple completion of the square yields an alternative representation for the transition kernel, namely

$$q_{k+1}(u_k, \mathrm{d}u_{k+1}) \propto \exp\Big(-\frac{1}{2}|y_{k+1} - H\psi(u_k)|^2_S - \frac{1}{2}|u_{k+1} - m_{k+1}|^2_C\Big)\mathrm{d}u_{k+1}\ , \tag{2.16}$$

where

$$\begin{aligned} C^{-1} &= \Sigma^{-1} + H^*\Gamma^{-1}H, \\ S &= H\Sigma H^* + \Gamma, \\ m_{k+1} &= C(\Sigma^{-1}\psi(u_k) + H^*\Gamma^{-1}y_{k+1})\ . \end{aligned} \tag{2.17}$$

The conditional mean m_{k+1} is often given in Kalman filter form

$$m_{k+1} = (I - KH)\psi(u_k) + Ky_{k+1}\ , \tag{2.18}$$

where K is the Kalman gain matrix

$$K = \Sigma H^* S^{-1}. \tag{2.19}$$

The expression (2.15) arises from application of Bayes' formula, derived above in (2.6)–(2.8), in the form (2.14), whilst (2.16) follows from a second application of Bayes' formula to derive the identity

$$P(y_{k+1} \mid u_{k+1})P(u_{k+1} \mid u_k)\mathrm{d}u_{k+1} = P(y_{k+1} \mid u_k)P(u_{k+1} \mid u_k, y_{k+1})\mathrm{d}u_{k+1}.$$

We note that

$$\begin{aligned} P(y_{k+1} \mid u_k) &= Z_S^{-1}\exp\Big(-\frac{1}{2}|y_{k+1} - H\psi(u_k)|^2_S\Big)\ , \\ P(u_{k+1} \mid u_k, y_{k+1}) &= Z_C^{-1}\exp\Big(-\frac{1}{2}|u_{k+1} - m_{k+1}|^2_C\Big)\ . \end{aligned} \tag{2.20}$$

These formulae are prevalent in the data assimilation literature; in particular (2.18) describes the evolution of the mean state estimate in the cycled 3DVAR algorithm, setting $u_{k+1} = m_{k+1}$ in [26]. We will make use of the formulae in Section 3 when describing optimal particle filters as random dynamical systems.

3 Particle Filters with Resampling

In this section we introduce the bootstrap particle filter, and the two optimal particle filters, in all three cases with resampling at every step. Assumption 2.1 ensures that the three particle filters have an elegant interpretation as a random dynamical system (RDS for short) which, in addition, is useful for our analyses. We thus introduce the filters in this way before giving the algorithmic definition which is more commonly found in the literature. The bootstrap particle

filter will not be the focus of subsequent theory, but does serve as an important motivation for the optimal particle filters, and in particular for the consistency results in Section 6.

For each of the three particle filters we will make frequent use of a resampling operator, which draws a sample $u_k^{(n)}$ from $\{\widehat{u}_k^{(m)}\}_{m=1}^N$ with weights $\{w_k^{(m)}\}_{m=1}^N$ which sum to one. To define this operator, we define the intervals $I_k^{(m)} = [\alpha_k^{(m)}, \alpha_k^{(m+1)})$, where

$$\begin{aligned}
&\alpha_k^{(m+1)} = \alpha_k^{(m)} + w_k^{(m+1)}, \quad m = 0, \cdots, N-1, \\
&\alpha_k^{(0)} = 0.
\end{aligned}$$

We then set

$$u_k^{(n)} = \sum_{m=1}^{N} \mathbb{I}_{I_k^{(m)}}(r_k^{(n)}) \widehat{u}_k^{(m)}, \tag{3.1}$$

where $r_k^{(n)} \sim U(0,1)$ i.i.d. Since the weights sum to one, $r_k^{(n)}$ will lie in exactly one of the intervals $I_k^{(i_*)}$ and we will have $u_k^{(n)} = \widehat{u}_k^{(i_*)}$. We also notice that

$$\sum_{m=1}^{N} \frac{1}{N} \delta_{u_k^{(m)}} = S^N \sum_{m=1}^{N} w_k^{(m)} \delta_{\widehat{u}_k^{(m)}},$$

where S^N is the sampling operator defined previously.

3.1 The bootstrap particle filter

The bootstrap particle filter (BPF for short) approximates the filtering distribution μ_k with an empirical measure

$$\rho_k^N = \sum_{n=1}^{N} \frac{1}{N} \delta_{u_k^{(n)}} \,. \tag{3.2}$$

The particle positions $\{u_k^{(n)}\}_{n=1}^N$ are defined as follows:

$$\begin{aligned}
&\widehat{u}_{k+1}^{(n)} = \psi(u_k^{(n)}) + \xi_k^{(n)}, \quad \xi_k^{(n)} \sim N(0, \Sigma) \text{ i.i.d.}\,, \\
&u_{k+1}^{(n)} = \sum_{m=1}^{N} \mathbb{I}_{I_{k+1}^{(m)}}(r_{k+1}^{(n)}) \widehat{u}_{k+1}^{(m)} \,,
\end{aligned} \tag{3.3}$$

where the second equation uses the resampling operator defined in (3.1) with weights computed according to

$$w_{k+1}^{(n),*} = \exp\Big(-\frac{1}{2} |y_{k+1} - H\widehat{u}_{k+1}^{(n)}|_\Gamma^2 \Big), \quad w_{k+1}^{(n)} = \frac{w_{k+1}^{(n),*}}{\sum_{j=1}^{N} w_{k+1}^{(j),*}} \,. \tag{3.4}$$

Thus, for each particle in the RDS, we propagate them forward using the dynamical model and then re-sample from the weighted particles to account for the observation likelihood.

Recall Bayes formula (2.6). The bootstrap particle filter approximates the posterior via a sequential application of importance sampling, using

$$P(u_{k+1} \mid Y_k) = \int P(u_{k+1} \mid u_k) P(u_k \mid Y_k) \mathrm{d}u_k$$

as the proposal and re-weighting according to the likelihood $P(y_{k+1} \mid u_{k+1})$. Thus the method is typically described by the following algorithm for updating the particle positions. The particles are initialized with $u_0^{(n)} \sim \mu_0$ and then updated iteratively as follows:

(1) Draw $\widehat{u}_{k+1}^{(n)} \sim P(u_{k+1} \mid u_k^{(n)})$.

(2) Define the weights $w_{k+1}^{(n)}$ for $n = 1, \cdots, N$ by

$$w_{k+1}^{(n),*} = P(y_{k+1} \mid \widehat{u}_{k+1}^{(n)}), \quad w_{k+1}^{(n)} = \frac{w_{k+1}^{(n),*}}{\sum_{m=1}^{N} w_{k+1}^{(m),*}} . \tag{3.5}$$

(3) Draw $u_{k+1}^{(n)}$ from $\{\widehat{u}_{k+1}^{(n)}\}_{n=1}^N$ with weights $\{w_{k+1}^{(n)}\}_{n=1}^N$.

Under Assumption 2.1, it is clear that the sampling and re-weighting procedures are consistent with (3.3). Note that the normalization factor $P(y_{k+1} \mid Y_k)$ is not required in the algorithm and is instead approximated via the normalization procedure in the second step.

In addition to ρ_k^N it is also useful to define the related measure

$$\widehat{\rho}_k^N = \sum_{n=1}^{N} w_k^{(n)} \delta_{\widehat{u}_k^{(n)}} , \tag{3.6}$$

with $\widehat{\rho}_0^N = \mu_0$, which is related to the bootstrap particle filter by $\rho_k^N = S^N \widehat{\rho}_k^N$. As we shall see in Section 6, the advantage of $\widehat{\rho}_k^N$ is that it has a recursive definition which allows for elegant proofs of consistency results in [39].

3.2 Optimal particle filter

The optimal particle filter with resampling can also be formulated as a RDS. We once again approximate the filtering distribution μ_k with an empirical distribution

$$\mu_k^N = \sum_{n=1}^{N} \frac{1}{N} \delta_{u_k^{(n)}} . \tag{3.7}$$

Under Assumption 2.1 the particles in this approximation are defined as follows. The particle positions are initialized with $u_0^{(n)} \sim \mu_0$. Given a collection of particles $u_k^{(n)}$ the particles are evolved according to the RDS update step

$$\begin{aligned} \widehat{u}_{k+1}^{(n)} &= (I - KH)\psi(u_k^{(n)}) + K y_{k+1} + \zeta_k^{(n)}, \quad \zeta_k^{(n)} \sim N(0, C) \text{ i.i.d.}, \\ u_{k+1}^{(n)} &= \sum_{m=1}^{N} \mathbb{I}_{I_{k+1}^{(m)}}(r_{k+1}^{(n)}) \widehat{u}_{k+1}^{(m)} . \end{aligned} \tag{3.8}$$

Here C, S, K are defined in (2.17), (2.18) and as with the BPF, the second equation uses the resampling operator defined in (3.1) but now using weights computed by

$$w_{k+1}^{(n),*} = \exp\Big(-\frac{1}{2} |y_{k+1} - H\psi(u_k^{(n)})|_S^2 \Big), \quad w_{k+1}^{(n)} = \frac{w_{k+1}^{(n),*}}{\sum_{m=1}^{N} w_{k+1}^{(m),*}} . \tag{3.9}$$

In light of the formulae given in (2.20), which are derived under Assumption 2.1, we see that the optimal particle filter is updating the particle positions by sampling from $P(u_{k+1} \mid u_k^{(n)}, y_{k+1})$ and then re-sampling to account for the likelihood factor $P(y_{k+1} \mid u_k^{(n)})$. In particular, without necessarily making Assumption 2.1, the optimal particle filter is a sequential importance sampling scheme applied to the following decomposition of the filtering distribution

$$\begin{aligned} P(u_{k+1} \mid Y_{k+1}) &= \int_{\mathcal{X}} P(u_{k+1}, u_k \mid Y_{k+1}) \mathrm{d}u_k \\ &= \int_{\mathcal{X}} P(u_{k+1} \mid u_k, y_{k+1}) P(u_k \mid Y_{k+1}) \mathrm{d}u_k \\ &= \int_{\mathcal{X}} \frac{P(y_{k+1} \mid u_k)}{P(y_{k+1} \mid Y_k)} P(u_{k+1} \mid u_k, y_{k+1}) P(u_k \mid Y_k) \mathrm{d}u_k \, . \end{aligned} \tag{3.10}$$

In the algorithmic setting, the filter is initialized with $u_0^{(n)} \sim \mu_0$, then for $k \geq 0$:

(1) Draw $\widehat{u}_{k+1}^{(n)}$ from $P(u_{k+1} \mid u_k^{(n)}, y_{k+1})$;

(2) Define the weights $w_{k+1}^{(n)}$ for $n = 1, \cdots, N$ by

$$w_{k+1}^{(n),*} = P(y_{k+1} \mid u_k^{(n)}), \quad w_{k+1}^{(n)} = \frac{w_{k+1}^{(n),*}}{\sum_{m=1}^{N} w_{k+1}^{(m),*}}; \tag{3.11}$$

(3) Draw $u_{k+1}^{(n)}$ from $\{\widehat{u}_{k+1}^{(m)}\}_{m=1}^N$ with weights $\{w_{k+1}^{(m)}\}_{m=1}^N$.

It is important to note that, although the OPF is well defined in this general setting for any choice of dynamics-obsevation model, it is only implementable under stringent assumptions on the forward and observation model, such as those given in Assumption 2.1; under this assumption the steps (1) and (2) may be implemented using the formulae given in (2.20) and exploited in the derivation of (3.8). We emphasize that models satisfying Assumption 2.1 do arise frequently in practice.

As with the BPF, it is beneficial to consider the related particle filter given by

$$\widehat{\mu}_k^N = \sum_{n=1}^{N} w_k^{(n)} \delta_{\widehat{u}_k^{(n)}} \tag{3.12}$$

for $k \geq 1$ and with $\widehat{\mu}_0^N = \mu_0$. Similarly to the bootstrap particle filter we have that $\mu_k^N = S^N \widehat{\mu}_k^N$.

3.3 Gaussianized optimal particle filter

In [22], an alternative implementation of the OPF is investigated and found to have superior performance on a range of test problems, particularly with respect to the curse of dimensionality. We refer to this filter as the Gaussianized optimal particle filter (GOPF for short), but note that it was first derived in [37]. Once again, we approximate the filtering distribution with an empirical measure

$$\nu_k^N = \sum_{n=1}^{N} \frac{1}{N} \delta_{v_k^{(n)}} \, . \tag{3.13}$$

As in the previous subsection, we first describe the filter under Assumption 2.1. The filter is initialized with $v_0^{(n)} \sim \mu_0$, with subsequent iterates generated by the RDS:

$$\begin{aligned} \tilde{v}_k^{(n)} &= \sum_{m=1}^{N} \mathbb{I}_{I_{k+1}^{(m)}}(r_{k+1}^{(n)}) v_k^{(m)} \ , \\ v_{k+1}^{(n)} &= (I - KH)\psi(\tilde{v}_k^{(n)}) + K y_{k+1} + \zeta_k^{(n)}, \quad \zeta_k^{(n)} \sim N(0, C) \text{ i.i.d.} \end{aligned} \tag{3.14}$$

and the weights appearing in the resampling operator are given by

$$w_{k+1}^{(n),*} = \exp\Big(-\frac{1}{2}|y_{k+1} - H\psi(v_k^{(n)})|_S^2 \Big), \quad w_{k+1}^{(n)} = \frac{w_{k+1}^{(n),*}}{\sum_{m=1}^{N} w_{k+1}^{(m),*}} \ . \tag{3.15}$$

Thus, the update procedure for GOPF is weight-resample-propagate, as opposed to propagate-weight-resample for the OPF. Hence the only difference between the OPF and GOPF is the ordering of the the resampling and propagation steps.

In our analysis it is sometimes useful to consider the equivalent RDS:

$$\begin{aligned} \widehat{v}_{k+1}^{(m,n)} &= (I - KH)\psi(v_k^{(m)}) + K y_{k+1} + \zeta_k^{(m,n)}, \quad \zeta_k^{(m,n)} \sim N(0, C) \text{ i.i.d.} \, , \\ v_{k+1}^{(n)} &= \sum_{m=1}^{N} \mathbb{I}_{I_{k+1}^{(m)}}(r_{k+1}^{(n)}) \widehat{v}_{k+1}^{(m,n)} \ . \end{aligned} \tag{3.16}$$

The sequences $v_k^{(n)}$ defined in (3.14) and (3.16) agree because for every n there is exactly one $m = m^*(n)$ such that $\widehat{v}_{k+1}^{(m^*(n),n)}$ survives the resampling step. Writing the algorithm this way allows certain parts of our subsequent analysis to be performed very similarly for both the OPF and GOPF; it is not a formulation to be implemented in practice.

For a general dynamics-observation model, the GOPF is described by the following algorithm:

(1) Define the weights $w_{k+1}^{(n)}$ for $n = 1, \cdots, N$ by

$$w_{k+1}^{(n),*} = P(y_{k+1} \mid v_k^{(n)}), \quad w_{k+1}^{(n)} = \frac{w_{k+1}^{(n),*}}{\sum_{m=1}^{N} w_{k+1}^{(m),*}} \ . \tag{3.17}$$

(2) Draw $\tilde{v}_k^{(n)}$ from $\{v_k^{(m)}\}_{m=1}^N$ with weights $\{w_{k+1}^{(m)}\}_{m=1}^N$.

(3) Draw $v_{k+1}^{(n)}$ from $P(u_{k+1} \mid \tilde{v}_k^{(n)}, y_{k+1})$.

Unlike for the previous filters, there is no need to define an associated 'hatted' measure, as the GOPF can be shown to satisfy a very natural recursion. This will be discussed in Section 6.

4 Ergodicity for Optimal Particle Filters

In this section we study the conditional ergodicity of the two optimal particle filters. The proofs are structurally very similar to one another and so we give details only in one case. The ergodicity results require a metric on probability measures to quantify convergence of differently

initialized posteriors in the long time limit. To this end, we define the total variation metric on $\mathcal{M}(\mathcal{X})$ by

$$d_{TV}(\mu,\nu) = \frac{1}{2}\sup_{|h|\leq 1}|\mu(h)-\nu(h)|, \tag{4.1}$$

where the supremum is taken over all bounded functions $h : \mathcal{X} \to \mathbb{R}$ with $|h| \leq 1$, and where we define $\mu(h) := \int_{\mathcal{X}} h(x)\mu(\mathrm{d}x)$ for any probability measure $\mu \in \mathcal{M}(\mathcal{X})$ and any real-valued test function h bounded by 1 on $\mathcal{X}$. This definition is then readily extended to probability measures on $\mathcal{X}^N$.

4.1 Optimal particle filter

Before stating the conditional ergodicity result, we first need some notation. Define $u_k = (u_k^{(1)}, \cdots, u_k^{(N)})$ to be particle positions defined by the RDS (3.8) with $\mu_0 = \delta_{z_0}$ and similarly $u'_k = (u_k^{(1)'}, \cdots, u_k^{(N)'})$ with $\mu_0 = \delta_{z'_0}$. Then u_k is a Markov chain taking values on $\mathcal{X}^N$, whose inhomogeneous Markov kernel we denote by $q_k(z,\cdot)$. The law of u_k is given by $q^k(z_0,\cdot)$, defined recursively by composing the q_k; and similarly the law of u'_k is given by $q^k(z'_0,\cdot)$. The conditional ergodicity result states that if the two filters u_k, u'_k are driven by the same observational data, then the law of u_k will converge to the law of u'_k exponentially as $k \to \infty$.

Remark 4.1 We abuse notation in this subsection by using $u_k \in \mathcal{X}^N$ to denote the N particles comprising the optimal particle filter; this differs from the notation $u_k \in \mathcal{X}$ used in the remainder of the paper to denote the underlying dynamical model. Similarly q_k is here a one-step linear Markov kernel on $\mathcal{X}^N$ whereas, previously, it denoted a one-step nonlinear Markov kernel on $\mathcal{X}$. We note also the important distinction between q_k and q^k: in the former case q_k is a one-step transition kernel, inhomogeneous and depending on k; in the latter case q^k is kernel found by composing over k steps.

Theorem 4.1 *Suppose that Assumptions* 2.1 *hold. Consider the OPF particles u_k, u'_k defined above. Assume moreover that the observational data used to define each filter is the same, and given by $\{y_k^\dagger\}_{k\geq 1}$ from Assumption* 2.2. *Then there exists $\mathfrak{z}_N \in (0,1)$ such that, almost surely with respect to the randomness generating $\{y_k^\dagger\}_{k\geq 1}$,*

$$\limsup_{k\to\infty}(d_{\mathrm{TV}}(q^k(z_0,\cdot),q^k(z'_0,\cdot)))^{\frac{1}{k}} \leq \mathfrak{z}_N\ . \tag{4.2}$$

Proof Step A Notice that

$$q^{k+1}(z_0,\cdot) = \int_{\mathcal{X}} q_{k+1}(u_k,\cdot)q^k(z_0,\mathrm{d}u_k), \tag{4.3}$$

where the transition kernel $q_{k+1}(u_k,\cdot)$ is here viewed as being a measure.

The heart of the argument is Step B, below, in which we prove a minorization condition for the transition kernel q_{k+1}, as we did for the filtering distribution itself in subsection 2.2. That is, we seek a measure $\mathbb{Q} \in \mathcal{M}(\mathcal{X}^N)$ and a sequence of constants $\epsilon_k > 0$ satisfying

$$q_{k+1}(u,A) \geq \epsilon_k \mathbb{Q}(A) \tag{4.4}$$

for all $u \in \mathcal{X}^N$ and all measurable sets $A \subset \mathcal{X}^N$. Given a minorization condition, we obtain the result via the following standard coupling argument. The minorization condition allows us to define a new Markov kernel

$$\widetilde{q}_{k+1}(x, A) = (1 - \epsilon_k)^{-1}(q_{k+1}(x, A) - \epsilon_k \mathbb{Q}(A)) \, . \tag{4.5}$$

The Markov chain in which transitions occur with probability $1 - \epsilon_k$ according to $\widetilde{q}_{k+1}(x, \cdot)$ and with probability ϵ_k according to $\mathbb{Q}(\cdot)$ is equivalent in law to the Markov chain $\{u_k\}$, and similarly for $\{u'_k\}$. Furthermore we may couple the two Markov chains by using the same random variables to select whether moves are made according to $\widetilde{q}_{k+1}(x, \cdot)$ or $\mathbb{Q}(\cdot)$. We now complete the coupling argument, assuming the minorization condition holds. We compare expectations of the two Markov chains, using the equivalent in law formulation above, and coupled through the random moves according to $\mathbb{Q}(\cdot)$ which occur at each step with probability ε_k. Let A_k be the event that the state independent Markov kernel $\mathbb{Q}(\cdot)$ is not picked at all times $j = 0, \cdots, k-1$. Then we have

$$\begin{aligned} d_{\mathrm{TV}}(q^k(z_0, \cdot), q^k(z'_0, \cdot)) &= \frac{1}{2} \sup_{|f|_\infty \leq 1} |\mathbb{E}(f(u_k) - f(u'_k))| \\ &= \frac{1}{2} \sup_{|f|_\infty \leq 1} |\mathbb{E}((f(u_k) - f(u'_k))\mathbb{I}_{A_k} + (f(u_k) - f(u'_k))\mathbb{I}_{A_k^c})| \, . \end{aligned}$$

Note that for this coupling the second term vanishes, as in the event A_k^c, the two chains Markov kernels $q^k(z_0, \cdot)$ and $q^k(z'_0, \cdot)$ will have become identical to measure $\mathbb{Q}$ at, or before, step k. Once that happens, they remain identical for all future steps. It follows that

$$d_{\mathrm{TV}}(q^k(z_0, \cdot), q^k(z'_0, \cdot)) \leq \mathbb{E}(\mathbb{I}_{A_k}) = \mathbb{P}(A_k) = \prod_{j=1}^{k} (1 - \epsilon_j) \, .$$

To obtain the result (4.2), we need to understand the limiting behaviour of the constants ϵ_j appearing in the minorization condition (4.4). Hence we turn our attention toward obtaining the minorization condition.

Step B Before deriving the minorization, we introduce some preliminaries. Using the fact that

$$y^\dagger_{k+1} = H\psi(u^\dagger_k) + \gamma(rH\xi^\dagger_k + \eta^\dagger_{k+1})$$

and defining

$$\begin{aligned} a_k = ((I - KH)\psi(u_k^{(n)}) + KH\psi(u_k^\dagger))_{n=1}^N, \quad \zeta_k = (\zeta_k^{(n)})_{n=1}^N, \\ \rho^\dagger_{k,0} = \gamma(rH\xi^\dagger_k + \eta^\dagger_{k+1}), \quad \rho^\dagger_k = (\gamma K(rH\xi^\dagger_k + \eta^\dagger_{k+1}))_{n=1}^N, \end{aligned}$$

we see that

$$\widehat{u}_{k+1} = a_k + \rho^\dagger_k + \zeta_k \, .$$

The next element of the sequence, u_{k+1}, is then defined by the second identity in (3.8). We are interested in the conditional ergodicity of $\{u_k\}_{k=1}^\infty$ with the sequence $\{\rho^\dagger_k\}_{k=1}^\infty$ fixed. By Assumption 2.1, a_k is bounded uniformly in k. We define the covariance operator $\mathfrak{C} \in L(\mathcal{X}^N, \mathcal{X}^N)$ to be a block diagonal covariance with each diagonal entry equal to C and then

$$R = \sup_{(u,v)} (|(I - KH)\psi(u) + KH\psi(v)|_C^2) \, ,$$

which is finite by Assumption 2.1.

Now, let E_0 be the event that, upon resampling, every particle survives the resampling. There are $N!$ such permutations. We will do the calculation in the case of a trivial permutation, that is, where each particle is mapped to itself under the resampling. However the bounds which follow work for any permutation because we do not use any information about location of the mean of the particle proposals; we simply use bounds on the drift ψ. If each particle is mapped to itself, then $u_{k+1}^{(n)} = \widehat{u}_{k+1}^{(n)}$ for all $n = 1, \cdots, N$. It follows that

$$\begin{aligned} q_{k+1}(u, A) &= \mathbb{P}(u_{k+1} \in A | u_k = u) \\ &\geq \mathbb{P}(u_{k+1} \in A | u_k = u, E_0)\mathbb{P}(E_0) \\ &= \mathbb{P}(\widehat{u}_{k+1} \in A | u_k = u)\mathbb{P}(E_0) \ . \end{aligned} \tag{4.6}$$

We now note that

$$\begin{aligned} \mathbb{P}(\widehat{u}_{k+1} \in A | u_k = u) &= \frac{1}{\sqrt{(2\pi)^{dN} \det \mathfrak{C}}} \int_A \exp\Big(-\frac{1}{2}|x - a_k - \rho_k^\dagger|_{\mathfrak{C}}^2\Big) \mathrm{d}x \\ &\geq \frac{\exp(-|a_k + \rho_k^\dagger|_{\mathfrak{C}}^2)}{\sqrt{(2\pi)^{dN} \det \mathfrak{C}}} \int_A \exp(-|x|_{\mathfrak{C}}^2) \mathrm{d}x \\ &\geq 2^{-\frac{dN}{2}} \exp(-2|a_k|_{\mathfrak{C}}^2) \exp(-2|\rho_k^\dagger|_{\mathfrak{C}}^2) \mathbb{Q}_{\mathfrak{C}}(A) \\ &\geq 2^{-\frac{dN}{2}} \exp(-2NR^2) \exp(-2|\rho_k^\dagger|_{\mathfrak{C}}^2) \mathbb{Q}_{\mathfrak{C}}(A), \end{aligned}$$

where $\mathbb{Q}_{\mathfrak{C}}(A)$ is the Gaussian measure $N(0, \frac{1}{2}\mathfrak{C})$. Thus we have shown that

$$\mathbb{P}(\widehat{u}_{k+1} \in A | u_k = u) \geq \delta_k \mathbb{Q}_{\mathfrak{C}}(A), \tag{4.7}$$

where

$$\delta_k = 2^{-\frac{dN}{2}} \exp(-2NR^2) \exp(-2|\rho_k^\dagger|_{\mathfrak{C}}^2) \ .$$

Moreover, we have that

$$\mathbb{P}(E_0) = N! \prod_{n=1}^{N} w_{k+1}^{(n)}.$$

Note that we have the bound $w_{k+1}^{(n)} \geq \frac{w_{k+1}^{(n),*}}{N}$ for each $n = 1, \cdots, N$ because each $w_{k+1}^{(m),*}$ is bounded by 1. But we have

$$\begin{aligned} w_{k+1}^{(n),*} &= \exp\Big(-\frac{1}{2}|y_{k+1} - H\psi(u_k^{(n)})|_S^2\Big) \\ &= \exp\Big(-\frac{1}{2}|H\psi(u_k^\dagger) - H\psi(u_k^{(n)}) + \rho_{k,0}^\dagger|_S^2\Big) \\ &\geq \exp(-r^2 - |\rho_{k,0}^\dagger|_S^2), \end{aligned}$$

where

$$r^2 = \sup_{u,v} |H\psi(u) - H\psi(v)|_S^2$$

which is finite by Assumption 2.1. From this we see that

$$\mathbb{P}(E_0) \geq N! \frac{1}{N^N} \exp\big(-Nr^2 - N|\rho_{k,0}^\dagger|_S^2\big).$$

Thus we obtain the minorization condition (4.4) where

$$\epsilon_k = N! \frac{1}{N^N} \exp(-Nr^2 - N|\rho^\dagger_{k,0}|^2_S)\delta_k, \quad \mathbb{Q} = \mathbb{Q}_{\mathfrak{C}} \, .$$

Step C By the argument in Step A we have that

$$d_{\mathrm{TV}}(q^k(z_0,\cdot), q^k(z_0',\cdot))^{\frac{1}{k}} \leq \mathfrak{z}_k, \tag{4.8}$$

where $\mathfrak{z}_k = \big(\prod_{j=1}^{k} (1-\epsilon_j) \big)^{\frac{1}{k}}$. Since the ϵ_k are i.i.d. and integrable, by the law of large numbers, almost surely with respect to the randomness generating the true signal and the data, we have

$$\ln \mathfrak{z}_k = \frac{1}{k} \sum_{j=1}^{k} \ln(1-\epsilon_j) \to \mathbb{E} \ln(1-\epsilon_1) = -\mathbb{E} \sum_{n=1}^{\infty} \frac{1}{n} \epsilon_1^n \, . \tag{4.9}$$

But $\epsilon_1 \leq \exp(-2|\rho^\dagger_{1,0}|^2_\Gamma)$. Since $\rho^\dagger_{1,0}$ is Gaussian it follows that the n^{th} moment of ϵ_1 is bounded above by $\mathcal{O}(n^{-\frac{1}{2}})$ so that the limit of $\ln \mathfrak{z}_k$ is negative and finite; the result follows.

4.2 Gaussianized optimal filter

As in the last section, we define $v_k = (v_k^{(1)}, \cdots, v_k^{(N)})$ and similarly for v_k' using the RDS but now for the GOPF (3.14) (or alternatively (3.16)) with distinct initializations $\mu_0 = \delta_{z_0}$ and $\mu_0 = \delta_{z_0'}$. Similarly to Theorem 4.1, we let $q^k(z_0,\cdot)$ denote the law of v_k.

Theorem 4.2 *Suppose that Assumptions* 2.1 *hold. Consider the GOPF particles v_k, v_k' defined above. Assume moreover that the observational data used to define each filter is the same, and given by $\{y_k^\dagger\}_{k\geq 1}$ from Assumption* 2.2. *Then there exists $\mathfrak{z}_N \in (0,1)$ such that, almost surely with respect to the randomness generating $\{y_k^\dagger\}_{k\geq 1}$,*

$$\limsup_{k\to\infty} (d_{\mathrm{TV}}(q^k(z_0,\cdot), q^k(z_0',\cdot)))^{\frac{1}{k}} \leq \mathfrak{z}_N \, . \tag{4.10}$$

Proof The proof follows similarly to that of Theorem 4.1, in particular it suffices to obtain a minorization condition for $q_{k+1}(v,\cdot)$. We will use the RDS representation (3.16), which we now recall

$$\begin{aligned} \widehat{v}_{k+1}^{(m,n)} &= (I-KH)\psi(v_k^{(m)}) + Ky_{k+1} + \zeta_k^{(m,n)}, \quad \zeta_k^{(m,n)} \sim N(0,C) \text{ i.i.d.}\,, \\ v_{k+1}^{(n)} &= \sum_{m=1}^{N} \mathbb{I}_{I_{k+1}^{(m)}}(r_{k+1}^{(n)}) \widehat{v}_{k+1}^{(m,n)} \, . \end{aligned} \tag{4.11}$$

In this formulation, note that for each n there is one and only one $m = m^*(n)$ such that $\mathbb{I}_{I_k^{(m)}}(r_k^{(n)}) = 1$. We see that

$$v_k := (v_k^{(n)})_{n=1}^N = (\widehat{v}_k^{(m^*(n),n)})_{n=1}^N \, .$$

Using the fact that

$$y_{k+1}^\dagger = H\psi(u_k^\dagger) + \gamma(rH\xi_k^\dagger + \eta_{k+1}^\dagger)$$

and defining

$$a_k = ((I - KH)\psi(v_k^{(m^*(n))}) + KH\psi(u_k^\dagger))_{n=1}^N, \quad \zeta_k = (\zeta_k^{(m^*(n),n)})_{n=1}^N, \tag{4.12}$$

$$\rho_{k,0}^\dagger = \gamma(rH\xi_k^\dagger + \eta_{k+1}^\dagger), \quad \rho_k^\dagger = (\gamma K(rH\xi_k^\dagger + \eta_{k+1}^\dagger))_{n=1}^N, \tag{4.13}$$

we see that

$$v_{k+1} = a_k + \rho_k^\dagger + \zeta_k \ .$$

Now notice that

$$\begin{aligned} q_{k+1}(v, A) = \mathbb{P}(v_{k+1} \in A | v_k = v) &= \mathbb{P}((\widehat{v}_{k+1}^{(m^*(n),n)})_{n=1}^N \in A | v_k = v) \\ &= \frac{1}{\sqrt{(2\pi)^{dN} \det \mathfrak{C}}} \int_A \exp\Big(-\frac{1}{2}|x - a_k - \rho_k^\dagger|_{\mathfrak{C}}^2 \Big) \mathrm{d}x \\ &\geq \frac{\exp(-|a_k + \rho_k^\dagger|_{\mathfrak{C}}^2)}{\sqrt{(2\pi)^d \det \mathfrak{C}}} \int_A \exp(-|x|_{\mathfrak{C}}^2) \mathrm{d}x \\ &\geq 2^{-\frac{dN}{2}} \exp(-2|a_k|_{\mathfrak{C}}^2) \exp(-2|\rho_k^\dagger|_{\mathfrak{C}}^2) \mathbb{Q}_{\mathfrak{C}}(A) \\ &\geq 2^{-\frac{dN}{2}} \exp(-2NR^2) \exp(-2|\rho_k^\dagger|_{\mathfrak{C}}^2) \mathbb{Q}_{\mathfrak{C}}(A), \end{aligned} \tag{4.14}$$

where $\mathbb{Q}_{\mathfrak{C}}$ is the Gaussian measure $N(0, \frac{1}{2}\mathfrak{C})$. Thus we have shown that

$$q_{k+1}(v, A) \geq \delta_k \mathbb{Q}_{\mathfrak{C}}(A), \tag{4.15}$$

where

$$\delta_k = 2^{-\frac{dN}{2}} \exp(-2NR^2) \exp(-2|\rho_k^\dagger|_{\mathfrak{C}}^2) \ .$$

The remainder of the proof (step C) follows identically to Theorem 4.1.

Remark 4.2 We can compare our (upper bounds on the) rates of convergence for the two optimal filters, using the minorization constants. For the OPF we have

$$\epsilon_1 = N! \frac{1}{N^N} \exp\big(-Nr^2 - N|\rho_{1,0}^\dagger|_S^2\big)\delta_1 \ ,$$

where

$$\delta_1 = 2^{-\frac{dN}{2}} \exp(-2NR^2) \exp(-2|\rho_1^\dagger|_{\mathfrak{C}}^2) \ ,$$

for the GOPF we simply have $\epsilon_1 = \delta_1$. The extra N dependence in the OPF clearly leads to a slower (upper bound on the) rate of convergence for the OPF. Thus, by this simple argument, we obtain a better convergence rate for the GOPF than for the OPF. This suggests that the GOPF may have a better rate of convergence for fixed ensemble sizes; further analysis or experimental study of this point would be of interest.

5 Accuracy for Optimal Particle Filters

In this section we study the accuracy of the optimal particle filters, in the small noise limit $\gamma \to 0$. The expectation appearing in the theorem statements is with respect to the noise generating the data, and with respect to the randomness within the particle filter itself. Note that this situation differs from that in the accuracy result for the filter itself which uses data generated by the statistical model. Assumption 2.2 relaxes this assumption.

5.1 Optimal particle filter

Theorem 5.1 *Suppose that Assumptions* 2.1, 2.3 *hold and consider the OPF with particles* $\{u_k^{(n)}\}_{n=1}^N$ *defined by* (3.8) *with data* $\{y_k^\dagger\}$ *given by Assumption* 2.2. *It follows that there is constant* c *such that*

$$\limsup_{k\to\infty} \mathbb{E}\big(\max_n \|u_k^{(n)} - u_k^\dagger\|^2\big) \le c\gamma^2.$$

Proof First recall the notation $\Sigma = \sigma\Sigma_0$, $\Gamma = \gamma\Gamma_0$ and $r = \frac{\sigma}{\gamma}$. Now define

$$S_0 = r^2 H\Sigma_0 H^* + \Gamma_0,$$
$$C_0 = r^2(I - KH)\Sigma_0$$

and note that

$$S = \gamma^2 S_0, \quad C = \gamma^2 C_0, \quad K = r^2\Sigma_0 H^* S_0^{-1}\,.$$

We will use the RDS representation

$$\begin{aligned} \widehat{u}_{k+1}^{(n)} &= (I - KH)\psi(u_k^{(n)}) + Ky_{k+1}^\dagger + \gamma\zeta_{0,k}^{(n)}, \quad \zeta_k^{(n)} \sim N(0, C) \text{ i.i.d.}\,, \\ u_{k+1}^{(n)} &= \sum_{m=1}^N \mathbb{I}_{I_{k+1}^{(m)}}(r_{k+1}^{(n)})\widehat{u}_{k+1}^{(n)}\,, \end{aligned} \tag{5.1}$$

where $\zeta_{0,k}^{(n)} \sim N(0, C_0)$ i.i.d. Hence we have

$$u_{k+1}^\dagger = (I - KH)\psi(u_k^\dagger) + KH\psi(u_k^\dagger) + r\gamma\xi_k^\dagger, \tag{5.2}$$
$$\widehat{u}_{k+1}^{(n)} = (I - KH)\psi(u_k^{(n)}) + K(H\psi(u_k^\dagger) + \gamma\eta_{k+1}^\dagger) + \gamma\zeta_{0,k}^{(n)}, \tag{5.3}$$

where $\zeta_{0,k}^{(n)} \sim N(0, C_0)$ i.i.d. Subtracting, we obtain

$$\widehat{u}_{k+1}^{(n)} - u_{k+1}^\dagger = (I - KH)\big(\psi(u_k^{(n)}) - \psi(u_k^\dagger)\big) + \gamma\iota_k^{(n)}, \tag{5.4}$$

where $\iota_k^{(n)} := (K\eta_{k+1}^\dagger + \zeta_{0,k}^{(n)} - r\xi_k^\dagger)$. Moreover we have the identity

$$u_{k+1}^\dagger = \sum_{m=1}^N \mathbb{I}_{I_{k+1}^{(m)}}(r_{k+1}^{(n)})u_{k+1}^\dagger\,. \tag{5.5}$$

Thus, defining

$$e_k^{(n)} = u_k^{(n)} - u_k^\dagger, \quad \widehat{e}_k^{(n)} = \widehat{u}_k^{(n)} - u_k^\dagger,$$

we have from (5.1) and (5.5)

$$e_{k+1}^{(n)} = \sum_{m=1}^N \mathbb{I}_{I_{k+1}^{(m)}}(r_{k+1}^{(n)})\widehat{e}_{k+1}^{(m)}\,.$$

Thus

$$\max_n \|e_{k+1}^{(n)}\|^2 \le \max_m \|\widehat{e}_{k+1}^{(m)}\|^2\,,$$

where the norm is the one in which we have a contraction. Using (5.4), the Lipschitz property of $(I - KH)\psi(\cdot)$, taking expectations and using independence, yields

$$\mathbb{E}\big(\max_n \|u_{k+1}^{(n)} - u_{k+1}^\dagger\|^2\big) \le \alpha^2\mathbb{E}\big(\max_n \|u_k^{(n)} - u_k^\dagger\|^2\big) + \mathcal{O}(\gamma^2)$$

and the result follows by Gronwall.

5.2 Gaussianized optimal filter

Theorem 5.2 *Let Assumptions* 2.1, 2.3 *hold and consider the GOPF with particles* $\{v_k^{(n)}\}_{n=1}^N$ *defined by* (3.14) (*or* (3.16)) *with data* $\{y_k^\dagger\}$ *given by Assumption* 2.2. *It follows that there is constant* c *such that*

$$\limsup_{k\to\infty} \mathbb{E}\big(\max_n \|v_k^{(n)} - u_k^\dagger\|^2\big) \le c\gamma^2.$$

Proof Recall the notation defined at the beginning of the proof of Theorem 5.1. Recall also the RDS representation of the GOPF (3.16)

$$\begin{aligned} \widehat{v}_{k+1}^{(m,n)} &= (I-KH)\psi(v_k^{(m)}) + Ky_{k+1} + \gamma\zeta_{0,k}^{(m,n)}, \\ v_{k+1}^{(n)} &= \sum_{m=1}^N \mathbb{I}_{I_{k+1}^{(m)}}(r_{k+1}^{(n)})\widehat{v}_{k+1}^{(m,n)} \ , \end{aligned} \tag{5.6}$$

where we now have $\zeta_{0,k}^{(m,n)} \sim N(0, C_0)$ i.i.d. , recalling that $C = \gamma^2 C_0$. We also have the identity

$$u_{k+1}^\dagger = (I-KH)\psi(u_k^\dagger) + KH\psi(u_k^\dagger) + r\gamma\xi_k^\dagger \ , \tag{5.7}$$

where $\zeta_{0,k}^{(n)} \sim N(0, C_0)$ i.i.d. Subtracting, we obtain

$$\widehat{v}_{k+1}^{(m,n)} - u_{k+1}^\dagger = (I-KH)\big(\psi(v_k^{(m)}) - \psi(u_k^\dagger)\big) + \gamma\iota_k^{(n)}, \tag{5.8}$$

where $\iota_k^{(n)} := (K\eta_{k+1}^\dagger + \zeta_{0,k}^{(n)} - r\xi_k^\dagger)$. Note that

$$u_{k+1}^\dagger = \sum_{m=1}^N \mathbb{I}_{I_{k+1}^{(m)}}(r_{k+1}^{(n)})u_{k+1}^\dagger, \tag{5.9}$$

so that, defining

$$e_k^{(n)} = v_k^{(n)} - u_k^\dagger, \quad \widehat{e}_k^{(m,n)} = \widehat{v}_k^{(m,n)} - u_k^\dagger,$$

we have from (5.6), (5.8) and (5.9)

$$e_{k+1}^{(n)} = \sum_{m=1}^N \mathbb{I}_{I_{k+1}^{(m)}}(r_{k+1}^{(n)})\widehat{e}_{k+1}^{(m,n)}.$$

Thus

$$\max_n \|e_{k+1}^{(n)}\|^2 \le \max_m \|\widehat{e}_{k+1}^{(m,n)}\|^2 \ ,$$

where the norm is the one in which we have a contraction. Using (5.8), using the Lipschitz property of $(I-KH)\psi(\cdot)$, taking expectations and using independence, gives

$$\mathbb{E}\big(\max_n \|v_{k+1}^{(n)} - u_{k+1}^\dagger\|^2\big) \le \alpha^2\mathbb{E}\big(\max_n \|v_k^{(n)} - u_k^\dagger\|^2\big) + \mathcal{O}(\gamma^2) \ .$$

The result follows by Gronwall.

6 Consistency in the Large Particle Limit

In this section we state and prove consistency results for the BPF, OPF and GOPF introduced in Section 3, in a simple unified framework. For the BPF the result is well known but we reproduce it here as the prove serves as an ideological template for the more complicated proofs to follow; furthermore we present the clean proof given in [39] (see also [26, Chapter 4]) as this particular approach to the result generalizes naturally to the OPF and GOPF. We also note that the more general analysis of the consistency of the auxiliary particle filter in [20] implies the result that we prove here about the GOPF.

6.1 Bootstrap particle filter

In the following, we let $f_{k+1} : \mathcal{X} \to \mathbb{R}$ be any function with $f_{k+1}(u_{k+1}) \propto P(y_{k+1} \mid u_{k+1})$; any proportionality constant will suffice, but the normalization constant is of course natural. As in previous sections, we let μ_k denote the filtering distribution. The following theorem is stated and then proved through a sequence of lemmas in the remainder of the subsection.

Theorem 6.1 *Let $\widehat{\rho}_k^N, \rho_k^N$ be the BPFs defined by* (3.6), (3.2) *respectively, and suppose that there exists a constant $\kappa \in (0, 1]$ such that*

$$\kappa \leq f_{k+1}(u_{k+1}) \leq \kappa^{-1} \tag{6.1}$$

for all $u_{k+1} \in \mathcal{X}$, $y_{k+1} \in \mathcal{Y}$ and $k \in \{0, \cdots, K-1\}$. Then we have

$$d(\widehat{\rho}_K^N, \mu_K) \leq \sum_{k=1}^{K} (2\kappa^{-2})^k N^{-\frac{1}{2}} \tag{6.2}$$

and

$$d(\rho_K^N, \mu_K) \leq \sum_{k=0}^{K} (2\kappa^{-2})^k N^{-\frac{1}{2}} \tag{6.3}$$

for all $K, N \geq 1$.

Remark 6.1 Note that the constant κ^{-2} appearing in the estimates above arises as the ratio of the upper and lower bounds in (6.1). In particular, we cannot optimize κ by choosing a different proportionality constant for f_{k+1}.

Recall formulation (2.4) of the iteration for the filtering distribution. In terms of understanding the approximation properties of the BPF, the key observation is that the measures $\{\widehat{\rho}_k^N\}_{k\geq 0}$ satisfy the recursion

$$\widehat{\rho}_{k+1}^N = L_{k+1} S^N P \widehat{\rho}_k^N, \quad \widehat{\rho}_0^N = \mu_0, \tag{6.4}$$

where $P : \mathcal{M}(\mathcal{X}) \to \mathcal{M}(\mathcal{X})$ is the Markov semigroup and, as defined in subsection 1.3, $S^N : \mathcal{M}(\mathcal{X}) \to \mathcal{M}(\mathcal{X})$ is the sampling operator. The convergence of the measures is quantified by the metric on random elements of $\mathcal{M}(\mathcal{X})$ defined by

$$d(\mu, \nu) = \sup_{|f|_\infty \leq 1} \sqrt{\mathbb{E}^\omega |\mu(f) - \nu(f)|^2}\,,$$

where, in our setting, $\mathbb{E}^\omega$ will always denote expectation with respect to the randomness in the sampling operator S^N. This metric reduces to twice the total variation metric, used in studying ergodicity, when the measures are not random. The main ingredients for the proof are the following three estimates for the operators P, S^N and L_{k+1} with respect to the metric d.

Lemma 6.1 *We have the following*:
(1) $\sup_{\nu\in\mathcal{M}(\mathcal{X})} d(S^N\nu,\nu) \le N^{-\frac{1}{2}}$.
(2) $d(P\mu,P\nu) \le d(\mu,\nu)$ *for all* $\mu,\nu\in\mathcal{M}(\mathcal{X})$.

Proof See [26, Lemma 4.7, Lemma 4.8].

We state the following lemma in a slightly more general form than necessary for the BPF, as it will be applied in different contexts for the optimal particle filters.

Lemma 6.2 *Let $\mathcal{Z}$ be a finite dimensional Euclidean space. Suppose that $g_{k+1}:\mathcal{Z}\to[0,\infty)$ is bounded and that there exists $\kappa\in(0,1]$ such that*

$$\kappa \le g_{k+1}(u) \le \kappa^{-1} \tag{6.5}$$

for all $u\in\mathcal{Z}$ and define $G_{k+1}:\mathcal{M}(\mathcal{Z})\to\mathcal{M}(\mathcal{Z})$ by $G_{k+1}(\nu)(\varphi)=\nu(g_{k+1}\varphi)/\nu(g_{k+1})$. Then

$$d(G_{k+1}\mu, G_{k+1}\nu) \le (2\kappa^{-2})d(\mu,\nu)$$

for all $\mu,\nu\in\mathcal{M}(\mathcal{Z})$.

Proof See [26, Lemma 4.9].

We can now prove the consistency result.

Proof of Theorem 6.1 First note that, taking $\mathcal{Z}=\mathcal{X}$ and $g_{k+1}=f_{k+1}$ in Lemma 6.2, we obtain $G_{k+1}\nu = L_{k+1}\nu$. Thus, by (6.1), it follows that $d(L_{k+1}\mu, L_{k+1}\nu)\le(2\kappa^{-2})d(\mu,\nu)$ for all $\mu,\nu\in\mathcal{M}(\mathcal{X})$. Combining this fact with the recursions given in (6.4), (2.4) and the estimates given in Lemmas 6.1 we have

$$\begin{aligned} d(\widehat{\rho}^N_{k+1},\mu_{k+1}) &= d(L_{k+1}S^N P\widehat{\rho}^N_k, L_{k+1}P\mu_k) \\ &\le 2\kappa^{-2} d(S^N P\widehat{\rho}^N_k, P\mu_k) \\ &\le 2\kappa^{-2}(d(S^N P\widehat{\rho}^N_k, P\widehat{\rho}^N_k) + d(P\widehat{\rho}^N_k, P\mu_k)) \\ &\le 2\kappa^{-2}N^{-\frac{1}{2}} + 2\kappa^{-2} d(\widehat{\rho}^N_k,\mu_k)\,. \end{aligned} \tag{6.6}$$

And since $\widehat{\rho}^N_0=\mu_0$, we obtain (6.2) by induction. Moreover, since $\rho_k = S^N\widehat{\rho}^N_k$

$$d(\rho_k,\mu_k) = d(S^N\widehat{\rho}^N_k,\mu_k) \le d(S^N\widehat{\rho}^N_k,\widehat{\rho}^N_k) + d(\widehat{\rho}^N_k,\mu_k) \tag{6.7}$$

and (6.3) follows.

6.2 Sequential importance resampler

In this section we will apply the above strategy to prove the corresponding consistency result for the OPF. Instead of restricting to the OPF, we will obtain results for the sequential

importance resampler (SIR for short), for which the OPF is a special case. See [13, sections II, III] for background in sequential importance sampling, and on the use of resampling. As with the OPF, the SIR is an empirical measure

$$\mu_k^N = \sum_{n=1}^{N} \frac{1}{N} \delta_{u_k^{(n)}} \,. \tag{6.8}$$

We will abuse notation slightly by keeping the same notation for the OPF and the SIR. The particle positions are drawn from a proposal distribution $\pi(u_{k+1}|u_k, y_{k+1})$ and re-weighted accordingly. As usual, the positions are initialized with $u_0^{(n)} \sim \mu_0$ and updated by

(1) Draw $\widehat{u}_{k+1}^{(n)}$ from $\pi(u_{k+1} \mid u_k^{(n)}, y_{k+1})$.

(2) Define the weights $w_{k+1}^{(n)}$ for $n = 1, \cdots, N$ by

$$w_{k+1}^{(n),*} = \frac{P(y_{k+1}|\widehat{u}_{k+1}^{(n)}) P(\widehat{u}_{k+1}^{(n)}|u_k^{(n)})}{\pi(\widehat{u}_{k+1}^{(n)}|u_k^{(n)}, y_{k+1})}, \quad w_{k+1}^{(n)} = \frac{w_{k+1}^{(n),*}}{\sum_{m=1}^{N} w_{k+1}^{(m),*}} \,. \tag{6.9}$$

(3) Draw $u_{k+1}^{(n)}$ from $\{\widehat{u}_{k+1}^{(m)}\}_{m=1}^N$ with weights $\{w_{k+1}^{(m)}\}_{m=1}^N$.

Thus, if we take the proposal to be $\pi(u_{k+1} \mid u_k, y_{k+1}) = P(u_{k+1} \mid u_k, y_{k+1})$ then we obtain the OPF (3.7). Without being more specific about the proposal π, it is not possible to represent the SIR as a random dynamical system in general.

Precisely as with the OPF, for the SIR we define the related filter

$$\widehat{\mu}_k^N = \sum_{n=1}^{N} w_k^{(n)} \delta_{\widehat{u}_k^{(n)}} \tag{6.10}$$

with $\widehat{\mu}_0^N = \mu_0$ and note the important identity $\mu_k^N = S^N \widehat{\mu}_k^N$. The following theorem, and corollary, are proved in the remainder of the subsection, through a sequence of lemmas.

Theorem 6.2 *Let $\widehat{\mu}^N, \mu^N$ be the SIR filters defined by* (6.10), (6.8) *respectively, with proposal distribution π. Suppose that there exists $f_{k+1} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ with*

$$f_{k+1}(u_{k+1}, u_k) \propto \frac{P(y_{k+1} \mid u_{k+1}) P(u_{k+1} \mid u_k)}{\pi(u_{k+1} \mid u_k, y_{k+1})} \tag{6.11}$$

and satisfying

$$\kappa \le f_{k+1}(u_{k+1}, u_k) \le \kappa^{-1} \tag{6.12}$$

for all $u_{k+1}, u_k \in \mathcal{X}$, $k \in \{0, \cdots, K-1\}$ and some $\kappa \in (0, 1]$. Then we have

$$d(\widehat{\mu}_K^N, \mu_K) \le \sum_{k=1}^{K} (2\kappa^{-2})^k N^{-\frac{1}{2}} \tag{6.13}$$

and

$$d(\mu_K^N, \mu_K) \le \sum_{k=0}^{K} (2\kappa^{-2})^k N^{-\frac{1}{2}} \tag{6.14}$$

for all $K, N \ge 1$.

Remark 6.2 As for the boostrap particle filter, the appearance of κ^{-2} reflects the ratio of the upper and lower bounds in (6.11); hence there is nothing to be gained from optimizing over the constant of proportionality. If we let

$$f_{k+1}(u_{k+1}, u_k) = \frac{P(y_{k+1} \mid u_{k+1})P(u_{k+1} \mid u_k)}{\pi(u_{k+1} \mid u_k, y_{k+1})},$$

then the estimate (6.12) is equivalent to

$$\kappa\pi(u_{k+1} \mid u_k, y_{k+1}) \leq P(y_{k+1} \mid u_{k+1})P(u_{k+1} \mid u_k) \leq \kappa^{-1}\pi(u_{k+1} \mid u_k, y_{k+1}) \ .$$

This can thus be interpreted as a quantification of equivalence between measures π and the optimal proposal $P(y_{k+1} \mid u_{k+1})P(u_{k+1} \mid u_k)$.

Remark 6.3 It is important to note that Assumption 2.1 on the dynamics-observation model is not required by Theorem 6.2. However Assumption 2.1 can be used to ensure that (6.12) holds. This observation leads to the following corollary.

Corollary 6.1 *Let $\widehat{\mu}^N, \mu^N$ be the OPFs defined in* (3.12), (3.7) *respectively and satisfying Assumption* 2.1. *Then there is $\kappa = \kappa(Y_K)$ such that we have*

$$d(\widehat{\mu}_K^N, \mu_K) \leq \sum_{k=1}^{K} (2\kappa^{-2})^k N^{-\frac{1}{2}} \tag{6.15}$$

and

$$d(\mu_K^N, \mu_K) \leq \sum_{k=0}^{K} (2\kappa^{-2})^k N^{-\frac{1}{2}} \tag{6.16}$$

for all $K, N \geq 1$ and where $\kappa^{-1} = \exp\big(\max_{0\leq j\leq K-1} |y_{j+1}|^2 + \sup_v |H\psi(v)|_S^2\big)$.

Although similar to the argument for the BPF, the recursion argument for the SIR is necessarily more complicated than that for the BPF, as the weights $w_{k+1}^{(n)}$ can potentially depend on both $u_{k+1}^{(n)}$ and $u_k^{(n)}$. This suggests that we must build a recursion which updates measures on a joint space $(u_{k+1}, u_k) \in \mathcal{X} \times \mathcal{X}$. This would also be necessary if we restricted our attention to the OPF, as the weights are defined using $u_k^{(n)}$ and not the particle positions $\widehat{u}_{k+1}^{(n)}$ after the proposal.

The recursion is defined using the following three operators.

(1) First P_{k+1}^{π} maps probability measures on $\mathcal{X}$ to probability measures on $\mathcal{X} \times \mathcal{X}$ by

$$P_{k+1}^{\pi}\mu(A) = \int\!\!\int_A \pi(u_{k+1}|u_k, y_{k+1})\mu(\mathrm{d}u_k)\mathrm{d}u_{k+1}, \tag{6.17}$$

where A is a measurable subset of $\mathcal{X} \times \mathcal{X}$.

(2) The reweighting operator L_{k+1}^{π} maps probability measures on $\mathcal{X} \times \mathcal{X}$ to probability measures on $\mathcal{X} \times \mathcal{X}$ and is defined by

$$L_{k+1}^{\pi}Q(A) = Z^{-1} \int\!\!\int_A w_{k+1}(u_{k+1}, u_k)Q(\mathrm{d}u_{k+1}, \mathrm{d}u_k), \tag{6.18}$$

where Z is the normalization constant of the resulting measure. The weight function is given by

$$w_{k+1}(u_{k+1}, u_k) = \frac{P(y_{k+1}|u_{k+1})P(u_{k+1}|u_k)}{\pi(u_{k+1}|u_k, y_{k+1})} \,. \tag{6.19}$$

(3) Finally, M maps probability measures on $\mathcal{X} \times \mathcal{X}$ into probability measures on $\mathcal{X}$ via marginalization onto the first component:

$$MQ(B) = \int \int_{B \times \mathcal{X}} Q(\mathrm{d}u_{k+1}, \mathrm{d}u_k).$$

It is easy to see that the posterior μ_k satisfies a natural recursion in terms of these operators.

Lemma 6.3 $\mu_{k+1} = ML^{\pi}_{k+1} P^{\pi}_{k+1} \mu_k$.

Proof Let $P(u_k \mid Y_k)$ denote the density of μ_k, then $P^{\pi}_{k+1}\mu_k$ is a measure on $\mathcal{X} \times \mathcal{X}$ with density

$$\pi(u_{k+1} \mid u_k, y_{k+1}) P(u_k \mid Y_k) \,. \tag{6.20}$$

And $L^{\pi}_{k+1} P^{\pi}_{k+1} \mu_k$ is a measure on $\mathcal{X} \times \mathcal{X}$ with density

$$\begin{aligned} &Z^{-1} w_{k+1}(u_{k+1}, u_k) \pi(u_{k+1} \mid u_k, y_{k+1}) P(u_k \mid Y_k) \\ &= Z^{-1} P(y_{k+1} \mid u_{k+1}) P(u_{k+1} \mid u_k) P(u_k \mid Y_k) \,. \end{aligned} \tag{6.21}$$

Finally, $ML^{\pi}_{k+1} P^{\pi}_{k+1} \mu_k$ is a measure on $\mathcal{X}$ with density

$$\int_{\mathcal{X}} Z^{-1} P(y_{k+1} \mid u_{k+1}) P(u_{k+1} \mid u_k) P(u_k \mid Y_k) \mathrm{d}u_k = Z^{-1} P(y_{k+1} \mid u_{k+1}) P(u_{k+1} \mid Y_k) \,. \tag{6.22}$$

Similarly, for the normalization factor, we have

$$\begin{aligned} Z &= \int \int_{\mathcal{X} \times \mathcal{X}} P(y_{k+1} \mid u_{k+1}) P(u_{k+1} \mid u_k) P(u_k \mid Y_k) \mathrm{d}u_k \mathrm{d}u_{k+1} \\ &= \int_{\mathcal{X}} P(y_{k+1} \mid u_{k+1}) P(u_{k+1} \mid Y_k) \mathrm{d}u_k \end{aligned} \tag{6.23}$$

and thus by Bayes' formula (6.22) is equal to $P(u_{k+1} \mid Y_{k+1})$ as required.

We now show that an associated recursion is satisfied by the SIR filter $\widehat{\mu}^N$.

Lemma 6.4 *Let $\widehat{\mu}^N$ be the SIR filter given by* (6.10)*, then*

$$\widehat{\mu}^N_{k+1} = ML^{\pi}_{k+1} S^N P^{\pi}_{k+1} \widehat{\mu}^N_k \tag{6.24}$$

for all $k \geq 0$ and $N \geq 1$, where S^N denotes the sampling operator acting on $\mathcal{M}(\mathcal{X} \times \mathcal{X})$.

Proof By definition, $\widehat{\mu}^N_k = \sum_{n=1}^{N} w^{(n)}_k \delta_{\widehat{u}^{(n)}_k}$ so that $P^{\pi}_{k+1} \widehat{\mu}^N_k \in \mathcal{M}(\mathcal{X} \times \mathcal{X})$ with density

$$\sum_{n=1}^{N} w^{(n)}_k \pi(u_{k+1} \mid \widehat{u}^{(n)}_k, y_{k+1}) \delta(u_k - \widehat{u}^{(n)}_k) \,. \tag{6.25}$$

Note that a sample $U \sim P^{\pi}_{k+1}\widehat{\mu}^N_k$ is a pair $(\widehat{u}^{(n)}_{k+1}, u^{(n)}_k)$ obtained as follows: first draw a sample $u^{(n)}_k$ from $\{\widehat{u}^{(n)}_k\}^N_{n=1}$ with weights $\{w^{(n)}_k\}^N_{n=1}$ and then draw sample $\widehat{u}^{(n)}_{k+1}$ from $\pi(u_{k+1} \mid u^{(n)}_k, y_{k+1})$. Thus, by definition of the $\widehat{u}^{(n)}_{k+1}$ sequence we see that $S^N P^{\pi}_{k+1}\widehat{\mu}^N_k$ has density

$$\frac{1}{N}\sum_{n=1}^{N} \delta(u_{k+1} - \widehat{u}^{(n)}_{k+1})\delta(u_k - u^{(n)}_k)\,. \tag{6.26}$$

It follows that $L^{\pi}_{k+1}S^N P^{\pi}_{k+1}\mu^N_k$ has density

$$\sum_{n=1}^{N} Z^{-1} w_{k+1}(\widehat{u}^{(n)}_{k+1}, u^{(n)}_k)\delta(u_{k+1} - \widehat{u}^{(n)}_{k+1})\delta(u_k - u^{(n)}_k) \tag{6.27}$$

and $ML^{\pi}_{k+1}S^N P^{\pi}_{k+1}\mu^N_k$ has density

$$\sum_{n=1}^{N} Z^{-1} w_{k+1}(\widehat{u}^{(n)}_{k+1}, u^{(n)}_k)\delta(u_{k+1} - \widehat{u}^{(n)}_{k+1})\,. \tag{6.28}$$

Lastly, the normalization factor is given by

$$\begin{aligned} Z &= \int\int_{\mathcal{X}\times\mathcal{X}} \sum_{n=1}^{N} w_{k+1}(\widehat{u}^{(n)}_{k+1}, u^{(n)}_k)\delta(u_{k+1} - \widehat{u}^{(n)}_{k+1})\delta(u_k - u^{(n)}_k)\mathrm{d}u_k\mathrm{d}u_{k+1} \\ &= \sum_{n=1}^{N} w_{k+1}(\widehat{u}^{(n)}_{k+1}, u^{(n)}_k)\,, \end{aligned} \tag{6.29}$$

so that $Z^{-1}w_{k+1}(\widehat{u}^{(n)}_{k+1}, u^{(n)}_k) = w^{(n)}_{k+1}$ and we obtain the result.

In the final step before proving Theorem 6.2, we state some simple properties for the operators appearing in the recursions. Note that these are similar but not (all) immediately implied by the corresponding results for the BPF, Lemma 6.1.

Lemma 6.5 *We have the following simple estimates*:
(1) $d(M\nu, M\mu) \le d(\nu, \mu)$,
(2) $d(P^{\pi}_{k+1}\nu, P^{\pi}_{k+1}\mu) \le d(\nu, \mu)$,
(3) $\sup_{\nu\in\mathcal{M}(\mathcal{X}\times\mathcal{X})} d(S^N\nu, \nu) \le N^{-\frac{1}{2}}$.

Proof Let $\tilde{f}(x, y) = f(x)$ and let $g(x, y)$ denote an arbitrary function. Then

$$M\nu(f) - M\mu(f) = \nu(\tilde{f}) - \mu(\tilde{f}). \tag{6.30}$$

The first inequality follows immediately from taking supremum over all $|f| \le 1$, which is necessarily smaller than the supremum of $\nu(g) - \mu(g)$ over all $|g| \le 1$.

We also have

$$P^{\pi}_{k+1}\nu(g) - P^{\pi}_{k+1}\mu(g) = \nu(g^{\pi}) - \mu(g^{\pi}), \tag{6.31}$$

where $g^{\pi}(u_k) = \int g(u_{k+1}, u_k)\pi(u_{k+1} \mid u_k, Y_{k+1})\mathrm{d}u_{k+1}$. And since $|g^{\pi}|_{\infty} \le 1$, the second inequality follows. The third inequality is proven in [26, Lemma 4.7], simply replacing $\mathcal{X}$ with $\mathcal{X} \times \mathcal{X}$.

We can now proceed with the main result.

Proof of Theorem 6.2 In the context of Lemma 6.2, take $\mathcal{Z} = \mathcal{X} \times \mathcal{X}$ and $g_{k+1} = f_{k+1}$, it follows that $G_{k+1}\nu = L^{\pi}_{k+1}\nu$. Indeed, for any $\varphi : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and with $g_{k+1} = Z^{-1}w_{k+1}$ we have

$$G_{k+1}\nu(\varphi) = \frac{\nu(g_{k+1}\varphi)}{\nu(g_{k+1})} = \frac{\nu(w_{k+1}\varphi)}{\nu(w_{k+1})} = L^{\pi}_{k+1}\nu(\varphi) \, .$$

We therefore obtain from Lemma 6.2 that

$$d(L^{\pi}_{k+1}\mu, L^{\pi}_{k+1}\nu) \leq (2\kappa^{-2})d(\mu, \nu)$$

for all $\mu, \nu \in \mathcal{M}(\mathcal{X} \times \mathcal{X})$.

Thus, using the recursions given in Lemmas 6.3, 6.4 and the estimates given in Lemma 6.5, we obtain

$$\begin{aligned} d(\widehat{\mu}^N_{k+1}, \mu_{k+1}) &= d(ML^{\pi}_{k+1}S^N P^{\pi}_{k+1}\widehat{\mu}^N_k, ML^{\pi}_{k+1}P^{\pi}_{k+1}\mu_k) \\ &\leq d(L^{\pi}_{k+1}S^N P^{\pi}_{k+1}\widehat{\mu}^N_k, L^{\pi}_{k+1}P^{\pi}_{k+1}\mu_k) \\ &\leq 2\kappa^{-2}d(S^N P^{\pi}_{k+1}\widehat{\mu}^N_k, P^{\pi}_{k+1}\mu_k) \\ &\leq 2\kappa^{-2}\left(d(S^N P^{\pi}_{k+1}\widehat{\mu}^N_k, P^{\pi}_{k+1}\widehat{\mu}^N_k) + d(P^{\pi}_{k+1}\widehat{\mu}^N_k, P^{\pi}_{k+1}\mu_k)\right) \\ &\leq 2\kappa^{-2}N^{-\frac{1}{2}} + 2\kappa^{-2}d(\widehat{\mu}^N_k, \mu_k) \, , \end{aligned} \tag{6.32}$$

and since $\widehat{\mu}^N_0 = \mu_0$, we obtain (6.15) by induction. Moreover, since $\mu^N_k = S^N\widehat{\mu}^N_k$,

$$d(\mu^N_k, \mu_k) = d(S^N\widehat{\mu}^N_k, \mu_k) \leq d(S^N\widehat{\mu}^N_k, \widehat{\mu}^N_k) + d(\widehat{\mu}^N_k, \mu_k) \tag{6.33}$$

and (6.16) follows.

The corollary follows immediately.

Proof of Corollary 6.1 For the OPF we have

$$\pi(u_{k+1} \mid u_k, y_{k+1}) = P(u_{k+1} \mid u_k, y_{k+1}) = \frac{P(y_{k+1} \mid u_{k+1})P(u_{k+1} \mid u_k)}{P(y_{k+1} \mid u_k)}, \tag{6.34}$$

where we have applied Bayes formula in the final equality. But under Assumption 2.1 we have that

$$P(y_{k+1} \mid u_k) = Z_S^{-1} \exp\Big(-\frac{1}{2}|y_{k+1} - H\psi(u_k)|^2_S \Big) \, . \tag{6.35}$$

Thus we define f_{k+1} by

$$f_{k+1}(u_{k+1}, u_k) = Z_S \frac{P(y_{k+1} \mid u_{k+1})P(u_{k+1} \mid u_k)}{\pi(u_{k+1} \mid u_k, y_{k+1})} = \exp\Big(-\frac{1}{2}|y_{k+1} - H\psi(u_k)|^2_S \Big) \tag{6.36}$$

and hence (6.12) holds with $\kappa^{-1} = \exp\Big(\max_{0 \leq j \leq K-1} |y_{j+1}|^2 + \sup_v |H\psi(v)|^2_S \Big)$, which, for each Y_k, is finite by Assumption 2.1. The result follows from Theorem 6.2.

6.3 Gaussianized optimal particle filter

In this section we derive the consistency result for the GOPF.

Theorem 6.3 *Let ν^N be the GOPF defined by* (3.13) *and let Assumption* 2.1 *hold. Then there is $\kappa = \kappa(Y_K)$ such that*

$$d(\nu_K^N, \mu_K) \leq \sum_{k=0}^{K} (2\kappa^{-2})^k N^{-\frac{1}{2}} \tag{6.37}$$

for all $K, N \geq 1$, where $\kappa^{-1} = \exp\Big(\max_{0 \leq j \leq K-1} |y_{j+1}|^2 + \sup_v |H\psi(v)|_S^2 \Big)$.

For the GOPF, the consistency proof uses the same strategy, but turns out to be much more straightforward. First note that the decomposition of the filtering distribution given in (3.10) gives the recursion formula

$$\mu_{k+1} = Q_{k+1} K_{k+1} \mu_k, \tag{6.38}$$

where $K_{k+1} : \mathcal{M}(\mathcal{X}) \to \mathcal{M}(\mathcal{X})$ is defined by

$$K_{k+1}\mu(A) = Z^{-1} \int_A P(y_{k+1} \mid u_k) \mu(\mathrm{d}u_k) \tag{6.39}$$

for all measurable $A \subset \mathcal{X}$ where Z is the normalization constant, and $Q_{k+1} : \mathcal{M}(\mathcal{X}) \to \mathcal{M}(\mathcal{X})$ is the Markov semigroup with kernel $P(u_{k+1} \mid u_k, y_{k+1})$.

Moreover, we have the following recursion for the GOPF. Let $\nu_k^N = \frac{1}{N} \sum_{n=1}^{N} \delta_{v_k^{(n)}}$.

Lemma 6.6 *The GOPF ν_k^N satisfies the recursion*

$$\nu_{k+1}^N = S^N Q_{k+1} K_{k+1} \nu_k^N \tag{6.40}$$

with $\nu_0^N = S^N \mu_0$.

Proof Note that $K_{k+1} \in \mathcal{M}(\mathcal{X})$ with density

$$\sum_{n=1}^{N} Z^{-1} P(y_{k+1} \mid v_k^{(n)}) \delta(v_k - v_k^{(n)}) \, . \tag{6.41}$$

The normalization constant is given by

$$Z = \int_{\mathcal{X}} \sum_{n=1}^{N} P(y_{k+1} \mid v_k^{(n)}) \delta(v_k - v_k^{(n)}) \mathrm{d}v_k = \sum_{n=1}^{N} P(y_{k+1} \mid v_k^{(n)}) \tag{6.42}$$

and thus $Z^{-1} P(y_{k+1} \mid v_k^{(n)}) = w_{k+1}^{(n)}$. We then have $Q_{k+1} K_{k+1} \nu_k^N \in \mathcal{M}(\mathcal{X})$ with density

$$\sum_{n=1}^{N} w_{k+1}^{(n)} P(v_{k+1} \mid v_k^{(n)}, y_{k+1}) \, . \tag{6.43}$$

To draw a sample $v_{k+1}^{(n)}$ from this mixture model, we draw $\tilde{v}_k^{(n)}$ from $\{v_k^{(m)}\}_{m=1}^N$ with weights $\{w_{k+1}^{(m)}\}_{m=1}^N$ and then draw $v_{k+1}^{(n)}$ from $P(v_{k+1} \mid \tilde{v}_k^{(n)}, y_{k+1})$. It follows that $S^N Q_{k+1} K_{k+1} \nu_k^N = \nu_{k+1}^N$.

Proof of Theorem 6.3 If we let

$$g_{k+1}(v_k) := Z_S P(y_{k+1}|v_k) = \exp\Big(-\frac{1}{2}|y_{k+1} - H\psi(v_k)|_S^2\Big)\,,$$

then g_{k+1} satisfies the assumptions of Lemma 6.2 with

$$\kappa^{-1} = \exp\Big(\max_{0\le j\le K-1}|y_{j+1}|^2 + \sup_v |H\psi(v)|_S^2\Big).$$

In particular, since $G_{k+1}\nu = K_{k+1}\nu$, it follows from Lemma 6.2 that

$$d(K_{k+1}\mu, K_{k+1}\nu) \le (2\kappa^{-2})d(\mu,\nu)$$

for all $\mu,\nu \in \mathcal{M}(\mathcal{X})$.

Using the recursions (6.38), (6.40) and the estimates from Lemma 6.1, we see that

$$\begin{aligned} d(\nu_{k+1}^N, \mu_{k+1}) &= d(S^N Q_{k+1}K_{k+1}\nu_k^N, Q_{k+1}K_{k+1}\mu_k) \\ &\le d(S^N Q_{k+1}K_{k+1}\nu_k^N, Q_{k+1}K_{k+1}\nu_k^N) + d(Q_{k+1}K_{k+1}\nu_k^N, Q_{k+1}K_{k+1}\mu_k) \\ &\le N^{-\frac{1}{2}} + d(K_{k+1}\nu_k^N, K_{k+1}\mu_k) \\ &\le N^{-\frac{1}{2}} + 2\kappa^{-2}d(\nu_k^N, \mu_k^N) \end{aligned} \tag{6.44}$$

by induction, we obtain

$$d(\nu_{k+1}^N, \mu_{k+1}^N) \le \sum_{j=0}^{k}(2\kappa^{-2})^j N^{-\frac{1}{2}} + (2\kappa^{-2})^{k+1}d(\nu_0^N, \mu_0)\,. \tag{6.45}$$

And the result follows from the fact $d(\nu_0^N, \mu_0) = d(S^N\mu_0, \mu_0) \le N^{-\frac{1}{2}}$.

Acknowledgements DK was supported as an NYU-Courant instructor. The authors are grateful to an anonymous referee for comments on an earlier version of this paper which have led to a correction of the presentation in Subsection 2.2, and pointers to relevant literature that were previously omitted; they are also grateful to Jonathan Mattingly for discussions relating to ergodicity of nonlinear Markov processes, and to Arnaud Doucet and Adam Johansen for helpful pointers to the literature. The authors also highlight, on the occasion of his 70th birthday, the inspirational work of Andy Majda, in applied mathematics in general, and in the area of data assimilation and filtering (the subject of this paper) in particular.

References

[1] Akashi, H. and Kumamoto, H., Random sampling approach to state estimation in switching environments, *Automatica*, **13**(4), 1977, 429–434.

[2] Atar, R. and Zeitouni, O., Exponential stability for nonlinear filtering, *Annales de l'IHP Probabilitiés et statistiques*, **33**, 1997, 697–725,.

[3] Bauer, P., Thorpe, A. and Brunet, G., The quiet revolution of numerical weather prediction, *Nature*, **525**(7567), 2015, 47–55.

[4] Blömker, D., Law, K., Stuart, A. M. and Zygalakis, K. C., Accuracy and stability of the continuous-time 3DVAR filter for the Navier-Stokes equation, *Nonlinearity*, **26**(8), 2013, 2193.

[5] Budhiraja, A. and Ocone, D., Exponential stability of discrete-time filters for bounded observation noise, *Systems & Control Letters*, **30**(4), 1997, 185–193.

[6] Burgers, G., Leeuwen, P. J. V. and Evensen, G.. Analysis scheme in the ensemble Kalman filter, *Monthly Weather Review*, **126**(6), 1998, 1719–1724.

[7] Chorin, A. J. and Tu, X. M., Implicit sampling for particle filters, *Proceedings of the National Academy of Sciences*, **106**(41), 2009, 17249–17254.

[8] Crisan, D. and Heine, K., Stability of the discrete time filter in terms of the tails of noise distributions, *Journal of the London Mathematical Society*, **78**(2), 2008, 441–458.

[9] Moral, P. D., Kurtzmann, A. and Tugaut, J., On the stability and the uniform propagation of chaos of extended ensemble Kalman-Bucy filters, 2016, arXiv:1606.08256.

[10] Douc, R., Fort, G., Moulines, E. and Priouret, P., Forgetting the initial distribution for hidden Markov models, *Stochastic Processes and Their Applications*, **119**(4), 2009, 1235–1256.

[11] Douc, R., Moulines, E. and Olsson, J., Optimality of the auxiliary particle filter, *Probability and Mathematical Statistics*, **29**(1), 2009, 1–28.

[12] Doucet, A., Freitas, N. D. and Gordon, N., An introduction to sequential Monte Carlo methods, Sequential Monte Carlo Methods in Practice, Springer-Verlag, New York, 2010.

[13] Doucet, A., Godsill, S. and Andrieu, C., On sequential Monte Carlo sampling methods for Bayesian filtering, *Statistics and Computing*, **10**(3), 2000, 197–208.

[14] Evensen, G., Using the extended Kalman filter with a multilayer quasi-geostrophic ocean model, *Journal of Geophysical Research: Oceans*, **97**(C11), 1992, 17905–17924.

[15] Evensen, G., The ensemble Kalman filter: Theoretical formulation and practical implementation, *Ocean Dynamics*, **53**(4), 2003, 343–367.

[16] Ghil, M. and Malanotte-Rizzoli, P., Data assimilation in meteorology and oceanography, *Advances in Geophysics*, **33**, 1991, 141–266.

[17] Handel, R. V., The stability of conditional Markov processes and Markov chains in random environments, *The Annals of Probability*, **37**(5), 2009, 1876–1925.

[18] Harlim, J. and Majda, A. J., Catastrophic filter divergence in filtering nonlinear dissipative systems, *Communications in Mathematical Sciences*, **8**(1), 2010, 27–43.

[19] Houtekamer, P. L. and Mitchell, H. L., Ensemble Kalman filtering, *Quarterly Journal of the Royal Meteorological Society*, **131**(613), 2005, 3269–3289.

[20] Johansen, A. M. and Doucet, A., A note on auxiliary particle filters, *Statistics & Probability Letters*, **78**(12), 2008, 1498–1504.

[21] Kelly, D., Majda, A. J. and Tong, X. T., Concrete ensemble Kalman filters with rigorous catas-trophic filter divergence, *Proceedings of the National Academy of Sciences*, **112**(34), 2015, 10589–10594.

[22] Kelly, D., Vanden-Eijnden, E. and Weare, J., Personal Communications, 2016.

[23] Kelly, D. T. B., Law, K. J. H. and Stuart, A. M., Well-posedness and accuracy of the ensemble Kalman filter in discrete and continuous time, *Nonlinearity*, **27**(10), 2014, 2579.

[24] Kleptsyna, M. L. and Veretennikov, A. Y., On discrete time ergodic filters with wrong initial data, *Probability Theory and Related Fields*, **141**(3–4), 2008, 411–444.

[25] Kunita, H., Asymptotic behavior of the nonlinear filtering errors of Markov processes, *Journal of Multivariate Analysis*, **1**(4), 1971, 365–393.

[26] Law, K., Stuart, A. and Zygalakis, K., Data Assimilation, Springer-Verlag, New York, 2015.

[27] Law, K. J. H., Sanz-Alonso, D., Shukla, A. and Stuart, A. M., Filter accuracy for the Lorenz 96 model: fixed versus adaptive observation operators, *Physica D: Nonlinear Phenomena*, **325**, 2016, 1–13.

[28] Law, K. J. H., Shukla, A. and Stuart, A. M., Analysis of the 3DVAR filter for the partially observed Lorenz '63 model, *Discrete and Continuous Dynamical Systems A*, **34**, 2014, 1061–1078.

[29] Law, K. J. H., Tembine, H. and Tempone, R., Deterministic mean-field ensemble Kalman filtering, 2014, arXiv:1409.0628.

[30] Leeuwen, P. J. V., Particle filtering in geophysical systems, *Monthly Weather Review*, **137**(12), 2009, 4089–4114.

[31] Leeuwen, P. J. V., Nonlinear data assimilation in geosciences: An extremely efficient particle filter, *Quarterly Journal of the Royal Meteorological Society*, **136**(653), 2010, 1991–1999.

[32] Liu, J. S. and Chen, R., Blind deconvolution via sequential imputations, *Journal of the American Statistical Association*, **90**(430), 1995, 567–576.

[33] Lorenc, A. C., Analysis methods for numerical weather prediction, *Quarterly Journal of the Royal Meteorological Society*, **112**(474), 1986, 1177–1194.

[34] Majda, A. J. and Harlim, J., Filtering Complex Turbulent Systems, Cambridge University Press, Cambridge, 2012.

[35] Majda, A. J. and Tong, X. T., Robustness and accuracy of finite ensemble Kalman filters in large dimensions, 2016, arXiv:1606.09321.

[36] Moodey, A. J. F., Lawless, A. S., Potthast, R. W. E. and Leeuwen, P. J. V., Nonlinear error dynamics for cycled data assimilation methods, *Inverse Problems*, **29**(2), 2013, 025002.

[37] Pitt, M. K. and Shephard, N., Filtering via simulation: Auxiliary particle filters, *Journal of the American Statistical Association*, **94**(446), 1999, 590–599.

[38] Poyiadjis, G., Doucet, A. and Singh, S. S., Particle approximations of the score and observed information matrix in state space models with application to parameter estimation, *Biometrika*, **98**(1), 2011, 65–80.

[39] Rebeschini, P. and Handel, R. V., Can local particle filters beat the curse of dimensionality? *The Annals of Applied Probability*, **25**(5), 2015, 2809–2866.

[40] Sanz-Alonso, D. and Stuart, A. M., Long-time asymptotics of the filtering distribution for partially observed chaotic dynamical systems, *SIAM/ASA Journal on Uncertainty Quantification*, **3**(1), 2015, 1200–1220.

[41] Sedigh-Sarvestani, M., Albers, D. J., and Gluckman, B. J., Data assimilation of glucose dynamics for use in the intensive care unit, 2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society, IEEE, 2012, 5437–5440.

[42] Sedigh-Sarvestani, M., Schiff, S. J. and Gluckman, B. J., Reconstructing mammalian sleep dynamics with data assimilation, *PLoS Comput Biol*, **8**(11), 2012, e1002788.

[43] Snyder, C., Particle filters, the "optimal" proposal and high-dimensional systems, Proceedings of the ECMWF Seminar on Data Assimilation for Atmosphere and Ocean, 2011.

[44] Snyder, C., Bengtsson, T. and Morzfeld, M., Performance bounds for particle filters using the optimal proposal, *Monthly Weather Review*, **143**(11), 2015, 4750–4761.

[45] Tadic, V. Z. B. and Doucet, A., Asymptotic properties of recursive maximum likelihood estimation in non-linear state-space models, 2018, arXiv:1806.09571.

[46] Thrun, S., Particle filters in robotics, Proceedings of the Eighteenth Conference on Uncertainty in artificial Intelligence, Morgan Kaufmann Publishers Inc., San Francisco, CA, 2002, 511–518.

[47] Tong, X. T. and Handel, R. V., Ergodicity and stability of the conditional distributions of nondegenerate Markov chains, *Annals of Applied Probability*, **22**(4), 2012, 1495–1540.

[48] Tong, X. T., Majda, A. J. and Kelly, D., Nonlinear stability of the ensemble Kalman filter with adaptive covariance inflation, *Communications in Mathematical Sciences*, **14**(5), 2016, 1283–1313.

[49] Tong, X. T., Majda, A. J. and Kelly, D., Nonlinear stability and ergodicity of ensemble based Kalman filters, *Nonlinearity*, **29**(2), 2016, 657.

[50] Work, D. B., Blandin, S., Tossavainen, O. P. and et al., A traffic model for velocity data assimilation, *Applied Mathematics Research Express*, **2010**(1), 2010, 1–35.

[51] Zaritskii, V. S., Svetnik, V. B. and Šimelevič, L. I., Monte-Carlo technique in problems of optimal information processing, *Avtomatika i Telemekhanika*, **36**(12), 1975, 95–103.