ITERATED FILTERING[1]

By Edward L. Ionides, Anindya Bhadra, Yves Atchadé
and Aaron King

*University of Michigan and Fogarty International Center, National Institutes of
Health, University of Michigan, University of Michigan, and University of
Michigan and Fogarty International Center, National Institutes of Health*

Inference for partially observed Markov process models has been a long-
standing methodological challenge with many scientific and engineering ap-
plications. Iterated filtering algorithms maximize the likelihood function for
partially observed Markov process models by solving a recursive sequence of
filtering problems. We present new theoretical results pertaining to the con-
vergence of iterated filtering algorithms implemented via sequential Monte
Carlo filters. This theory complements the growing body of empirical evi-
dence that iterated filtering algorithms provide an effective inference strategy
for scientific models of nonlinear dynamic systems. The first step in our the-
ory involves studying a new recursive approach for maximizing the likelihood
function of a latent variable model, when this likelihood is evaluated via im-
portance sampling. This leads to the consideration of an iterated importance
sampling algorithm which serves as a simple special case of iterated filtering,
and may have applicability in its own right.

1. Introduction. Partially observed Markov process (POMP) models are
of widespread importance throughout science and engineering. As such, they
have been studied under various names including *state space models* [Durbin
and Koopman (2001)], *dynamic models* [West and Harrison (1997)] and *hidden
Markov models* [Cappé, Moulines and Rydén (2005)]. Applications include ecol-
ogy [Newman et al. (2009)], economics [Fernández-Villaverde and Rubio-Ramírez
(2007)], epidemiology [King et al. (2008)], finance [Johannes, Polson and Stroud
(2009)], meteorology [Anderson and Collins (2007)], neuroscience [Ergün et al.
(2007)] and target tracking [Godsill et al. (2007)].

This article investigates convergence of a Monte Carlo technique for estimating unknown parameters of POMPs, called *iterated filtering*, which was proposed by Ionides, Bretó and King (2006). Iterated filtering algorithms repeatedly carry out a filtering procedure to explore the likelihood surface at increasingly local scales in search of a maximum of the likelihood function. In several case-studies, iterated filtering algorithms have been shown capable of addressing scientific challenges in the study of infectious disease transmission, by making likelihood-based inference computationally feasible in situations where this was previously not the case [King et al. (2008); Bretó et al. (2009); He, Ionides and King (2010); Laneri et al. (2010)]. The partially observed nonlinear stochastic systems arising in the study of disease transmission and related ecological systems are a challenging environment to test statistical methodology [Bjørnstad and Grenfell (2001)], and many statistical methodologies have been tested on these systems in the past fifty years [e.g., Cauchemez and Ferguson (2008); Toni et al. (2008); Keeling and Ross (2008); Ferrari et al. (2008); Morton and Finkenstädt (2005); Grenfell, Bjornstad and Finkenstädt (2002); Kendall et al. (1999); Bartlett (1960); Bailey (1955)]. Since iterated filtering has already demonstrated potential for generating state-of-the-art analyses on a major class of scientific models, we are motivated to study its theoretical justification. The previous theoretical investigation of iterated filtering, presented by Ionides, Bretó and King (2006), did not engage directly in the Monte Carlo issues relating to practical implementation of the methodology. It is relatively easy to check numerically that a maximum has been attained, up to an acceptable level of Monte Carlo uncertainty, and therefore one can view the theory of Ionides, Bretó and King (2006) as motivation for an algorithm whose capabilities were proven by demonstration. However, the complete framework presented in this article gives additional insights into the potential capabilities, limitations and generalizations of iterated filtering.

The foundation of our iterated filtering theory is a Taylor series argument which we present first in the case of a general latent variable model in Section 2. This leads us to propose and analyze a novel *iterated importance sampling* algorithm for maximizing the likelihood function of latent variable models. Our motivation is to demonstrate a relatively simple theoretical result which is then extended to POMP models in Section 3. However, this result also demonstrates the broader possibilities of the underlying methodological approach.

The iterated filtering and iterated importance sampling algorithms that we study have an attractive practical property that the model for the unobserved process enters the algorithm only through the requirement that realizations can be generated at arbitrary parameter values. This property has been called *plug-and-play* [Bretó et al. (2009); He, Ionides and King (2010)] since it permits simulation code to be simply plugged into the inference procedure. A concept closely related to plug-and-play is that of *implicit models* for which the model is specified via an algorithm to generate stochastic realizations [Diggle and Gratton (1984); Bretó et al. (2009)]. In particular, evaluation of the likelihood function for implicit models is

considered unavailable. Implicit models arise when the model is represented by a "black box" computer program. A scientist investigates such a model by inputting parameter values, receiving as output from the "black box" independent draws from a stochastic process, and comparing these draws to the data to make inferences. For an implicit model, only plug-and-play statistical methodology can be employed. Other plug-and-play methods proposed for partially observed Markov models include approximate Bayesian computations implemented via sequential Monte Carlo [Liu and West (2001); Toni et al. (2008)], an asymptotically exact Bayesian technique combining sequential Monte Carlo with Markov chain Monte Carlo [Andrieu, Doucet and Holenstein (2010)], simulation-based forecasting [Kendall et al. (1999)], and simulation-based spectral analysis [Reuman et al. (2006)]. Further discussion of the plug-and-play property is included in the discussion of Section 4.

2. Iterated importance sampling. Let $f_{XY}(x, y; \theta)$ be the joint density of a pair of random variables (X, Y) depending on a parameter $\theta \in \mathbb{R}^p$. We suppose that (X, Y) takes values in some measurable space $\mathfrak{X} \times \mathfrak{Y}$, and $f_{XY}(x, y; \theta)$ is defined with respect to some σ-finite product measure which we denote by $dx\, dy$. We suppose that the observed data consist of a fixed value $y^* \in \mathfrak{Y}$, with X being unobserved. Therefore, $\{f_{XY}(x, y; \theta), \theta \in \mathbb{R}^p\}$ defines a general latent variable statistical model. We write the marginal densities of X and Y as $f_X(x; \theta)$ and $f_Y(y; \theta)$, respectively. The *measurement model* is the conditional density of the observed variable given the latent variable X, written as $f_{Y|X}(y \mid x; \theta)$. The log likelihood function is defined as $\ell(\theta) = \log f_Y(y^*; \theta)$. We consider the problem of calculating the maximum likelihood estimate, defined as $\hat{\theta} = \arg\max_\theta \ell(\theta)$.

We consider an iterated importance sampling algorithm which gives a plug-and-play approach to likelihood based inference for implicit latent variable models, based on generating simulations at parameter values in a neighborhood of the current parameter estimate to refine this estimate. This shares broadly similar goals with other Monte Carlo methods proposed for latent variable models [e.g., Johansen, Doucet and Davy (2008); Qian and Shapiro (2006)], and in a more general context has similarities with evolutionary optimization strategies [Beyer (2001)]. We emphasize that the present motivation for proposing and studying iterated importance sampling is to lay the groundwork for the results on iterated filtering in Section 3. However, the successes of iterated filtering methodology on POMP models also raise the possibility that related techniques may be useful in other latent variable situations.

We define the *stochastically perturbed* model to be a triplet of random variables $(\check{X}, \check{Y}, \check{\Theta})$, with perturbation parameter τ and parameter θ, having a joint density on $\mathfrak{X} \times \mathfrak{Y} \times \mathbb{R}^p$ specified as

$$(1) \qquad g_{\check{X}, \check{Y}, \check{\Theta}}(x, y, \check{\vartheta}; \theta, \tau) = f_{XY}(x, y; \check{\vartheta}) \tau^{-p} \kappa_\tau((\check{\vartheta} - \theta)/\tau).$$

Here, $\{\kappa_\tau, \tau > 0\}$ is a collection of mean-zero densities on $\mathbb{R}^p$ (with respect to Lebesgue measure) satisfying condition (A1) below:

(A1) For each $\tau > 0$, κ_τ is supported on a compact set $K_0 \subset \mathbb{R}^p$ independent of τ.

Condition (A1) can be satisfied by the arbitrary selection of κ_τ. At first reading, one can imagine that κ_τ is fixed, independent of τ. However, the additional generality will be required in Section 3.

We start by showing a relationship between conditional moments of $\check{\Theta}$ and the derivative of the log likelihood function, in Theorem 1. We write $\check{\mathbb{E}}_{\theta,\tau}[\cdot]$ to denote expectation with respect to the stochastically perturbed model. We write $u \in \mathbb{R}^p$ to specify a column vector, with u' being the transpose of u. For a function $f = (f_1, \ldots, f_m)' : \mathbb{R}^p \to \mathbb{R}^m$, we write $\int f(u)\, du$ for the vector $(\int f_1(u)\, du, \ldots, \int f_m(u)\, du)' \in \mathbb{R}^m$; For any function $f : \mathbb{R}^p \to \mathbb{R}$, we write $\nabla f(u)$ to denote the column vector gradient of f, with $\nabla^2 f(u)$ being the second derivative matrix. We write $|\cdot|$ for the absolute value of a vector or the largest absolute eigenvalue of a square matrix. We write $B(r) = \{u \in \mathbb{R}^p : |u| \leq r\}$ for the ball of radius r in $\mathbb{R}^p$. We assume the following regularity condition:

(A2) $\ell(\theta)$ is twice differentiable. For any compact set $K_1 \subset \mathbb{R}^p$,

$$\sup_{\theta \in K_1} |\nabla \ell(\theta)| < \infty \quad \text{and} \quad \sup_{\theta \in K_1} |\nabla^2 \ell(\theta)| < \infty.$$

THEOREM 1. *Assume conditions* (A1), (A2). *Let* $h : \mathbb{R}^p \to \mathbb{R}^m$ *be a measurable function possessing constants* $\alpha \geq 0$, $c > 0$ *and* $\varepsilon > 0$ *such that, whenever* $u \in B(\varepsilon)$,

$$|h(u)| \leq c|u|^\alpha. \tag{2}$$

Define $\tau_0 = \sup\{\tau : K_0 \subset B(\varepsilon/\tau)\}$. *For any compact set* $K_2 \subset \mathbb{R}^p$ *there exists* $C_1 < \infty$ *and a positive constant* $\tau_1 \leq \tau_0$ *such that, for all* $0 < \tau \leq \tau_1$,

$$
\begin{aligned}
\sup_{\theta \in K_2} &\left| \check{\mathbb{E}}_{\theta,\tau}[h(\check{\Theta} - \theta) \mid \check{Y} = y^*] - \int h(\tau u)\kappa_\tau(u)\, du \right.\\
&\left. \qquad - \tau \left\{ \int h(\tau u)u' \kappa_\tau(u)\, du \right\} \nabla \ell(\theta) \right| \\
&\leq C_1 \tau^{2+\alpha}.
\end{aligned}
\tag{3}
$$

PROOF. Let $g_{\check{Y}|\check{\Theta}}(y \mid \check{\vartheta}; \theta, \tau)$ denote the conditional density of $\check{Y}$ given $\check{\Theta}$. We note that $g_{\check{Y}|\check{\Theta}}$ does not depend on either τ or θ, and so we omit these dependencies below. Then, $g_{\check{Y}|\check{\Theta}}(y^* \mid \check{\vartheta}) = \int f_{XY}(x, y^*; \check{\vartheta})\, dx = \exp(\ell(\check{\vartheta}))$. Changing variable to $u = (\check{\vartheta} - \theta)/\tau$, we calculate

$$
\begin{aligned}
\check{\mathbb{E}}_{\theta,\tau}[h(\check{\Theta} - \theta) \mid \check{Y} = y^*] &= \frac{\int h(\check{\vartheta} - \theta) g_{\check{Y}|\check{\Theta}}(y^* \mid \check{\vartheta}) \tau^{-p} \kappa_\tau((\check{\vartheta} - \theta)/\tau)\, d\check{\vartheta}}{\int g_{\check{Y}|\check{\Theta}}(y^* \mid \check{\vartheta}) \tau^{-p} \kappa_\tau((\check{\vartheta} - \theta)/\tau)\, d\check{\vartheta}} \\
&= \frac{\int h(\tau u) \exp\{\ell(\theta + \tau u) - \ell(\theta)\} \kappa_\tau(u)\, du}{\int \exp\{\ell(\theta + \tau u) - \ell(\theta)\} \kappa_\tau(u)\, du}.
\end{aligned}
\tag{4}
$$

Applying the Taylor expansion $e^x = 1 + x + (\int_0^1 (1-t)e^{tx}\,dt)x^2$ to the numerator of (4) gives

$$
\int \exp\{\ell(\theta + \tau u) - \ell(\theta)\}h(\tau u)\kappa_\tau(u)\,du
$$

$$
(5) \quad = \int h(\tau u)\kappa_\tau(u)\,du + \int \{\ell(\theta + \tau u) - \ell(\theta)\}h(\tau u)\kappa_\tau(u)\,du
$$

$$
+ \int \left(\int_0^1 (1-t)e^{t(\ell(\theta + \tau u) - \ell(\theta))}\,dt \right)\{\ell(\theta + \tau u) - \ell(\theta)\}^2 h(\tau u)\kappa_\tau(u)\,du.
$$

We now expand the second term on the right-hand side of (5) by making use of the Taylor expansion

$$
(6) \quad \ell(\theta + \tau u) - \ell(\theta) = \tau u' \int_0^1 \nabla\ell(\theta + t\tau u)\,dt
$$

$$
= \tau u' \nabla\ell(\theta) + \tau u' \int_0^1 \{\nabla\ell(\theta + t\tau u) - \nabla\ell(\theta)\}\,dt
$$

and defining $\psi_{\tau,h}(\theta) = \int h(\tau u)\kappa_\tau(u)\,du + \tau\{\int h(\tau u)u'\kappa_\tau(u)\,du\}\nabla\ell(\theta)$. This allows us to rewrite (5) as

$$
(7) \quad \int \exp\{\ell(\theta + \tau u) - \ell(\theta)\}h(\tau u)\kappa_\tau(u)\,du = \psi_{\tau,h}(\theta) + R_1(\theta, \tau),
$$

where

$$
R_1(\theta, \tau)
$$

$$
= \tau \int h(\tau u)u' \left(\int_0^1 \{\nabla\ell(\theta + t\tau u) - \nabla\ell(\theta)\}\,dt \right)\kappa_\tau(u)\,du
$$

$$
+ \int \left(\int_0^1 (1-t)e^{t(\ell(\theta + \tau u) - \ell(\theta))}\,dt \right)\{\ell(\theta + \tau u) - \ell(\theta)\}^2 h(\tau u)\kappa_\tau(u)\,du.
$$

As a consequence of (A2), we have

$$
(8) \quad \sup_{\theta \in K_2} \sup_{u \in K_0} \sup_{t \in [0, \tau_0]} \left(|\nabla\ell(\theta + tu)| + |\nabla^2\ell(\theta + tu)| \right) < \infty.
$$

Combining (8) with the assumption that $K_0 \subset B(\varepsilon/\tau)$, we deduce the existence of a finite constant C_2 such that

$$
(9) \quad \sup_{\theta \in K_2} |R_1(\theta, \tau)| \leq C_2 \tau^{2+\alpha}.
$$

We bound the denominator of (4) by considering the special case of (7) in which h is taken to be the unit function, $h(u) = 1$. Noting that $\int u\kappa_\tau(u)\,du = 0$, we see that $\psi_{\tau,1}(\theta) = 1$ and so (7) yields

$$
\int \exp\{\ell(\theta + \tau u) - \ell(\theta)\}\kappa_\tau(u)\,du = 1 + R_2(\theta, \tau)
$$

with $R_2(\theta, \tau)$ having a bound

$$(10) \qquad \sup_{\theta \in K_2} |R_2(\theta, \tau)| \le C_3 \tau^2$$

for some finite constant C_3. We now note the existence of a finite constant C_4 such that

$$(11) \qquad \sup_{\theta \in K_2} |\psi_{\tau,h}(\theta)| \le C_4 \tau^\alpha$$

implied by (2), (A1) and (A2). Combining (9), (10) and (11) with the identity

$$\breve{\mathbb{E}}_{\theta,\tau}\big(h(\breve{\Theta} - \theta) \mid \breve{Y} = y^*\big) - \psi_{\tau,h}(\theta) = \frac{R_1(\theta, \tau) - R_2(\theta, \tau)\psi_{\tau,h}(\theta)}{1 + R_2(\theta, \tau)},$$

and requiring $\tau_1 < (2C_3)^{-1/2}$, we obtain that

$$(12) \qquad \sup_{\theta \in K_2} \big|\breve{\mathbb{E}}_{\theta,\tau}\big(h(\breve{\Theta} - \theta) \mid \breve{Y} = y^*\big) - \psi_{\tau,h}(\theta)\big| \le C_5 \tau^{2+\alpha}$$

for some finite constant C_5. Substituting the definition of $\psi_{\tau,h}(\theta)$ into (12) gives (3) and hence completes the proof. $\square$

One natural choice is to take $h(u) = u$ in Theorem 1. Supposing that κ_τ has associated positive definite covariance matrix $\breve{\Sigma}$, independent of τ, this leads to an approximation to the derivative of the log likelihood given by

$$(13) \qquad \big|\nabla\ell(\theta) - \{(\tau^2\breve{\Sigma})^{-1}(\breve{\mathbb{E}}_{\theta,\tau}[\breve{\Theta} \mid \breve{Y} = y^*] - \theta)\}\big| \le C_6 \tau$$

for some finite constant C_6, with the bound being uniform over θ in any compact subset of $\mathbb{R}^p$. The quantity $\breve{\mathbb{E}}_{\theta,\tau}[\breve{\Theta} \mid \breve{Y} = y^*]$ does not usually have a closed form, but a plug-and-play Monte Carlo estimate of it is available by importance sampling, supposing that one can draw from $f_X(x; \theta)$ and evaluate $f_{Y|X}(y^* \mid x; \theta)$. Numerical approximation of moments is generally more convenient than approximating derivatives, and this is the reason that the relationship in (13) may be useful in practice. However, one might suspect that there is no "free lunch" and therefore the numerical calculation of the left-hand side of (13) should become fragile as τ becomes small. We will see that this is indeed the case, but that iterated importance sampling methods mitigate the difficulty to some extent by averaging numerical error over subsequent iterations.

A trade-off between bias and variance is to be expected in any Monte Carlo numerical derivative, a classic example being the Kiefer–Wolfowitz algorithm [Kiefer and Wolfowitz (1952); Spall (2003)]. Algorithms which are designed to balance such trade-offs have been extensively studied under the label of *stochastic approximation* [Kushner and Yin (2003); Spall (2003); Andrieu, Moulines and Priouret (2005)]. Algorithm 1 is an example of a basic stochastic approximation algorithm taking advantage of (13). As an alternative, the derivative approximation in (13) could be combined with a stochastic line search algorithm. In order to obtain the plug-and-play property, we consider an algorithm that draws from

---

Algorithm 1 A basic iterated importance sampling procedure. The Monte Carlo random variables required at each iteration are presumed to be drawn independently. Theorem 2 gives sufficient conditions for $\hat{\theta}_M$ to converge to the maximum likelihood estimate as $M \to \infty$.

Input:

- Latent variable model described by a latent variable density $f_X(x; \theta)$, measurement model $f_{Y|X}(y \mid x; \theta)$, and data y^*.
- Perturbation density κ having compact support, zero mean and positive definite covariance matrix $\check{\Sigma}$.
- Positive sequences $\{\tau_m\}$ and $\{a_m\}$
- Integer sequence of Monte Carlo sample sizes, $\{J_m\}$
- Initial parameter estimate, $\hat{\theta}_1$
- Number of iterations, M

Procedure:

1 for m in $1 : M$
2   for j in $1 : J_m$
3   draw $Z_{j,m} \sim \kappa(\cdot)$ and set $\check{\Theta}_{j,m} = \hat{\theta}_m + \tau_m Z_{j,m}$
4   draw $\check{X}_{j,m} \sim f_X(\cdot; \check{\Theta}_{j,m})$
5   set $w_j = f_{Y|X}(y^* \mid \check{X}_{j,m}; \check{\Theta}_{j,m})$
6   end for
7   calculate $D_m = \tau_m^{-2} \check{\Sigma}^{-1} \{ (\sum_{j=1}^{J_m} w_j)^{-1} (\sum_{j=1}^{J_m} w_j \check{\Theta}_{j,m}) - \hat{\theta}_m \}$
8   update estimate: $\hat{\theta}_{m+1} = \hat{\theta}_m + a_m D_m$
9 end for

Output:

- parameter estimate $\hat{\theta}_{M+1}$

$f_X(x; \theta)$ for iteratively selected values of θ. This differs from other proposed iterative importance sampling algorithms which aim to construct improved importance sampling distributions [e.g., Celeux, Marin and Robert (2006)]. In principle, a procedure similar to Algorithm 1 could take advantage of alternative choices of importance sampling distribution: the fundamental relationship in Theorem 1 is separate from the numerical issues of computing the required conditional expectation by importance sampling. Theorem 2 gives sufficient conditions for the convergence of Algorithm 1 to the maximum likelihood estimate. To control the variance of the importance sampling weights, we suppose:

(A3) For any compact set $K_3 \subset \mathbb{R}^p$,

$$\sup_{\theta \in K_3, x \in \mathfrak{X}} f_{Y|X}(y^* \mid x; \theta) < \infty.$$

We also adopt standard sufficient conditions for stochastic approximation methods:

(B1) Define $\zeta(t)$ to be a solution to $d\zeta/dt = \nabla \ell(\zeta(t))$. Suppose that $\hat{\theta}$ is an *asymptotically stable equilibrium point*, meaning that (i) for every $\eta > 0$ there exists a $\delta(\eta)$ such that $|\zeta(t) - \hat{\theta}| \leq \eta$ for all $t > 0$ whenever $|\zeta(0) - \hat{\theta}| \leq \delta$, and (ii) there exists a δ_0 such that $\zeta(t) \to \hat{\theta}$ as $t \to \infty$ whenever $|\zeta(0) - \hat{\theta}| \leq \delta_0$.

(B2) With probability one, $\sup_m |\hat{\theta}_m| < \infty$. Further, $\hat{\theta}_m$ falls infinitely often into a compact subset of $\{\zeta(0) : \lim_{t \to \infty} \zeta(t) = \hat{\theta}\}$.

Conditions (B1) and (B2) are the basis of the classic results of Kushner and Clark (1978). Although research into stochastic approximation theory has continued [e.g., Kushner and Yin (2003); Andrieu, Moulines and Priouret (2005); Maryak and Chin (2008)], (B1) and (B2) remain a textbook approach [Spall (2003)]. The relative simplicity and elegance of Kushner and Clark (1978) makes an appropriate foundation for investigating the links between iterated filtering, sequential Monte Carlo and stochastic approximation theory. There is, of course, scope for variations on our results based on the diversity of available stochastic approximation theorems. Although neither (B1) and (B2) nor alternative sufficient conditions are easy to verify, stochastic approximation methods have nevertheless been found effective in many situations. Condition (B2) is most readily satisfied if $\hat{\theta}_m$ is constrained to a neighborhood in which $\hat{\theta}$ is a unique local maximum, which gives a guarantee of local rather than global convergence. Global convergence results have been obtained for related stochastic approximation procedures [Maryak and Chin (2008)] but are beyond the scope of this current paper. The rate assumptions in Theorem 2 are satisfied, for example, by $a_m = m^{-1}$, $\tau_m^2 = m^{-1}$ and $J_m = m^{(\delta+1/2)}$ for $\delta > 0$.

THEOREM 2. *Let $\{a_m\}$, $\{\tau_m\}$ and $\{J_m\}$ be positive sequences with $\tau_m \to 0$, $J_m \tau_m \to \infty$, $a_m \to 0$, $\sum_m a_m = \infty$ and $\sum_m a_m^2 J_m^{-1} \tau_m^{-2} < \infty$. Let $\hat{\theta}_m$ be defined via Algorithm 1. Assuming (A1)–(A3) and (B1) and (B2), $\lim_{m \to \infty} \hat{\theta}_m = \hat{\theta}$ with probability one.*

PROOF. The quantity D_m in line 7 of Algorithm 1 is a self-normalized Monte Carlo importance sampling estimate of

$$(\tau_m^2 \breve{\Sigma})^{-1} \{ \mathbb{E}_{\hat{\theta}_m, \tau_m}[\breve{\Theta} \mid \breve{Y} = y^*] - \hat{\theta}_m \}.$$

We can therefore apply Corollary 8 from Section A.2, writing E_{MC} and Var_{MC} for the Monte Carlo expectation and variance resulting from carrying out Algorithm 1 conditional on the data y^*. This gives

(14)
$$\big| E_{MC}\big[D_m - (\tau_m^2 \breve{\Sigma})^{-1} \{ \mathbb{E}_{\hat{\theta}_m, \tau_m}[\breve{\Theta} \mid \breve{Y} = y^*] - \hat{\theta}_m \}\big] \big|$$
$$\leq \frac{C_7 (\sup_{x \in \mathfrak{X}} f_{Y|X}(y^* \mid x; \hat{\theta}_m))^2}{\tau_m J_m (f_Y(y^*; \hat{\theta}_m))^2},$$

$$(15) \qquad |\operatorname{Var}_{MC}(D_m)| \le \frac{C_8(\sup_{x \in \mathfrak{X}} f_{Y|X}(y^* \mid x; \hat{\theta}_m))^2}{\tau_m^2 J_m(f_Y(y^*; \hat{\theta}_m))^2}$$

for finite constants C_7 and C_8 which do not depend on J, $\hat{\theta}_m$ or τ_m. Having assumed the conditions for Theorem 1, we see from (13) and (14) that D_m provides an asymptotically unbiased Monte Carlo estimate of $\nabla \ell(\hat{\theta}_m)$ in the sense of (B5) of Theorem 6 in Section A.1. In addition, (15) justifies (B4) of Theorem 6. The remaining conditions of Theorem 6 hold by hypothesis. $\square$

3. Iterated filtering for POMP models. Let $\{X(t), t \in T\}$ be a Markov process [Rogers and Williams (1994)] with $X(t)$ taking values in a measurable space $\mathfrak{X}$. The time index set $T \subset \mathbb{R}$ may be an interval or a discrete set, but we are primarily concerned with a finite subset of times $t_1 < t_2 < \cdots < t_N$ at which $X(t)$ is observed, together with some initial time $t_0 < t_1$. We write $X_{0:N} = (X_0, \ldots, X_N) = (X(t_0), \ldots, X(t_N))$. We write $Y_{1:N} = (Y_1, \ldots, Y_N)$ for a sequence of random variables taking values in a measurable space $\mathfrak{Y}^N$. We assume that $X_{0:N}$ and $Y_{1:N}$ have a joint density $f_{X_{0:N}, Y_{1:N}}(x_{0:n}, y_{1:n}; \theta)$ on $\mathfrak{X}^{N+1} \times \mathfrak{Y}^N$, with θ being an unknown parameter in $\mathbb{R}^p$. A POMP model may then be specified by an initial density $f_{X_0}(x_0; \theta)$, conditional transition densities $f_{X_n|X_{n-1}}(x_n \mid x_{n-1}; \theta)$ for $1 \le n \le N$, and the conditional densities of the observation process which are assumed to have the form $f_{Y_n|Y_{1:n-1}, X_{0:n}}(y_n \mid y_{1:n-1}, x_{0:n}; \theta) = f_{Y_n|X_n}(y_n \mid x_n; \theta)$. We use subscripts of f to denote the required joint and conditional densities. We write f without subscripts to denote the full collection of densities and conditional densities, and we call such an f the *generic density* of a POMP model. The data are a sequence of observations by $y_{1:N}^* = (y_1^*, \ldots, y_N^*) \in \mathfrak{Y}^N$, considered as fixed. We write the log likelihood function of the data for the POMP model as $\ell_N(\theta)$ where

$$\ell_n(\theta) = \log \int f_{X_0}(x_0; \theta) \prod_{k=1}^n f_{X_k, Y_k|X_{k-1}}(x_k, y_k^* \mid x_{k-1}; \theta) \, dx_{0:n}$$

for $1 \le n \le N$. Our goal is to find the maximum likelihood estimate, $\hat{\theta} = \arg \max_\vartheta \ell_N(\theta)$.

It will be helpful to construct a POMP model $\bar{f}$ which expands the model f by allowing the parameter values to change deterministically at each time point. Specifically, we define a sequence $\vartheta_{0:N} = (\vartheta_0, \ldots, \vartheta_N) \in \{\mathbb{R}^p\}^{N+1}$. We then write $(\overline{X}_{0:N}, \overline{Y}_{1:N}) \in \mathfrak{X}^{N+1} \times \mathfrak{Y}^N$ for a POMP with generic density $\bar{f}$ specified by the joint density

$$(16) \qquad \begin{aligned} &\bar{f}_{\overline{X}_{0:N}, \overline{Y}_{1:N}}(x_{0:N}, y_{1:N}; \vartheta_{0:n}) \\ &= f_{X_0}(x_0; \vartheta_0) \prod_{k=1}^N f_{Y_k, X_k|X_{k-1}}(y_k, x_k \mid x_{k-1}; \vartheta_k). \end{aligned}$$

We write the log likelihood of $y_{1:n}^*$ for the model $\overline{f}$ as $\overline{\ell}_n(\vartheta_{0:n}) = \log \bar{f}_{\overline{Y}_{1:n}}(y_{1:n}^*; \vartheta_{0:n})$. We write $\theta^{[k]}$ to denote k copies of $\theta \in \mathbb{R}^p$, concatenated in a column vector, so that $\ell_n(\theta) = \overline{\ell}_n(\theta^{[n+1]})$. We write $\nabla_i \overline{\ell}_n(\vartheta_{0:n})$ for the partial derivative of $\overline{\ell}_n$ with respect to ϑ_i, for $i = 0, \ldots, n$. An application of the chain rule gives the identity

$$(17) \qquad \nabla \ell_n(\theta) = \sum_{i=0}^{n} \nabla_i \overline{\ell}_n(\theta^{[n+1]}).$$

The regularity condition employed for Theorem 3 below is written in terms of this deterministically perturbed model:

(A4) For each $1 \le n \le N$, $\overline{\ell}_n(\theta_{0:n})$ is twice differentiable. For any compact subset K of $\{\mathbb{R}^p\}^{n+1}$ and each $0 \le i \le n$,

$$(18) \qquad \sup_{\theta_{0:n} \in K} |\nabla_i \overline{\ell}_n(\theta_{0:n})| < \infty \quad \text{and} \quad \sup_{\theta_{0:n} \in K} |\nabla_i^2 \overline{\ell}_n(\theta_{0:n})| < \infty.$$

Condition (A4) is a nonrestrictive smoothness assumption. However, the relationship between smoothness of the likelihood function, the transition density $f_{X_k|X_{k-1}}(x_k \mid x_{k-1}; \theta)$, and the observation density $f_{Y_k|X_k}(y_k \mid x_k; \theta)$ is simple to establish only under the restrictive condition that $\mathfrak{X}$ is a compact set. Therefore, we note an alternative to (A4) which is more restrictive but more readily checkable:

(A4′) $\mathfrak{X}$ is compact. Both $f_{X_k|X_{k-1}}(x_k \mid x_{k-1}; \theta)$ and $f_{Y_k|X_k}(y_k \mid x_k; \theta)$ are twice differentiable with respect to θ. These derivatives are continuous with respect to x_{k-1} and x_k.

Iterated filtering involves introducing an auxiliary POMP model in which a time-varying parameter process $\{\breve{\Theta}_n, 0 \le n \le N\}$ is introduced. Let κ be a probability density function on $\mathbb{R}^p$ having compact support, zero mean and covariance matrix Σ. Let $Z_0, \ldots, Z_N$ be N independent draws from κ. We introduce two perturbation parameters, σ and τ, and construct a process $\breve{\Theta}_{0:N}$ by setting $\breve{\Theta}_0 = \vartheta_0 + \tau Z_0$ and $\breve{\Theta}_k = \vartheta_k + \tau Z_0 + \sigma \sum_{j=1}^{k} Z_j$ for $1 \le k \le N$. The joint density of $\breve{\Theta}_{0:N}$ is written as $g_{\breve{\Theta}_{0:N}}(\breve{\vartheta}_{0:N}; \vartheta_{0:N}, \sigma, \tau)$. We define the stochastically perturbed POMP model g with a Markov process $\{(\breve{X}_n, \breve{\Theta}_n), 0 \le n \le N\}$, observation process $\breve{Y}_{1:N}$ and parameter $(\vartheta_{0:N}, \sigma, \tau)$ by the joint density

$$g_{\breve{X}_{0:N}, \breve{\Theta}_{0:N}, \breve{Y}_{1:N}}(x_{0:N}, \breve{\vartheta}_{0:N}, y_{1:N}; \vartheta_{0:N}, \sigma, \tau)$$
$$= g_{\breve{\Theta}_{0:N}}(\breve{\vartheta}_{0:N}; \vartheta_{0:N}, \sigma, \tau) \bar{f}_{\overline{X}_{0:N}, \overline{Y}_{1:N}}(x_{0:N}, y_{1:N}; \breve{\vartheta}_{0:N}).$$

We seek a result analogous to Theorem 1 which takes into account the specific structure of a POMP. Theorem 3 below gives a way to approximate $\nabla \ell_N(\theta)$ in terms of moments of the filtering distributions for g. We write $\breve{\mathbb{E}}_{\vartheta_{0:n}, \sigma, \tau}$ and

$\check{\mathrm{Var}}_{\vartheta_{0:n},\sigma,\tau}$ for the expectation and variance, respectively, for the model g. We will be especially interested in the situation where $\vartheta_{0:n} = \theta^{[n+1]}$, which leads us to define the following filtering means and prediction variances:

$$\check{\theta}_n^F = \check{\theta}_n^F(\theta, \sigma, \tau) = \check{\mathbb{E}}_{\theta^{[n+1]},\sigma,\tau}[\check{\Theta}_n \mid \check{Y}_{1:n} = y_{1:n}^*]$$

(19)
$$= \int \check{\vartheta}_n g_{\check{\Theta}_n|\check{Y}_{1:n}}(\check{\vartheta}_n \mid y_{1:n}^*; \theta^{[n+1]}, \sigma, \tau) \, d\check{\vartheta}_n,$$

$$\check{V}_n^P = \check{V}_n^P(\theta, \sigma, \tau) = \check{\mathrm{Var}}_{\theta^{[n+1]},\sigma,\tau}(\check{\Theta}_n \mid \check{Y}_{1:n-1} = y_{1:n-1}^*)$$

for $n = 1, \ldots, N$, with $\check{\theta}_0^F = \theta$.

THEOREM 3. *Suppose condition* (A4). *Let σ be a function of τ with* $\lim_{\tau \to 0} \sigma(\tau)/\tau = 0$. *For any compact set $K_4 \subset \mathbb{R}^p$, there exists a finite constant C_9 such that for all τ small enough,*

(20)
$$\sup_{\theta \in K_4} |\tau^{-2}\Sigma^{-1}(\check{\theta}_N^F - \theta) - \nabla \ell_N(\theta)| \le C_9(\tau + \sigma^2/\tau^2)$$

and

(21)
$$\sup_{\theta \in K_4} \left| \sum_{n=1}^N (\check{V}_n^P)^{-1}(\check{\theta}_n^F - \check{\theta}_{n-1}^F) - \nabla \ell_N(\theta) \right| \le C_9(\tau + \sigma^2/\tau^2).$$

PROOF. For each $n \in \{1, \ldots, N\}$, we map onto the notation of Section 2 by setting $X = \overline{X}_{0:n}$, $Y = \overline{Y}_{1:n}$, $\theta = \vartheta_{0:n}$, $\check{\Theta} = \check{\Theta}_{0:n}$, $y^* = y_{1:n}^*$ and $h(\vartheta_{0:n}) = \vartheta_n$. We note that, by construction, this implies $\check{X} = \check{X}_{0:n}$, $\check{Y} = \check{Y}_{1:n}$, and $\kappa_\tau(\vartheta_{0:n}) = \kappa(\vartheta_0) \prod_{i=1}^n (\sigma/\tau)^{-p} \kappa((\vartheta_i - \vartheta_{i-1})/(\sigma/\tau))$. For this choice of h, the integral $\tau \int h(\tau u) u' \kappa_\tau(u) \, du$ is a $p \times p(n+1)$ matrix for which the ith $p \times p$ sub-matrix is $\check{\mathbb{E}}_{\vartheta_{0:n},\sigma,\tau}[(\check{\Theta}_n - \vartheta_n)(\check{\Theta}_{i-1} - \vartheta_{i-1})'] = \{\tau^2 + (i-1)\sigma^2\}\Sigma$. Thus,

$$\left(\tau \int h(\tau u) u' \kappa_\tau(u) \, du \right) \nabla \bar{\ell}_n(\vartheta_{0:n})$$

(22)
$$= \sum_{i=0}^n \check{\mathbb{E}}_{\vartheta_{0:n},\sigma,\tau}[(\check{\Theta}_n - \vartheta_n)(\check{\Theta}_i - \vartheta_i)'] \nabla_i \bar{\ell}_n(\vartheta_{0:n})$$

$$= \sum_{i=0}^n (\tau^2 \Sigma + i\sigma^2 \Sigma) \nabla_i \bar{\ell}_n(\vartheta_{0:n}).$$

Applying Theorem 1 in this context, the second term in (3) is zero and the third term is given by (22). We obtain that for any compact $K_5 \subset \mathbb{R}^{p(n+1)}$ there is a

$C_{10} < \infty$ such that

$$\sup_{\vartheta_{0:n} \in K_5} \left| \tau^{-2} \Sigma^{-1} (\breve{\mathbb{E}}_{\vartheta_{0:n},\sigma,\tau}[\breve{\Theta}_n \mid Y_{1:n} = y_{1:n}^*] - \vartheta_n) \right.$$

(23)
$$\left. - \sum_{i=0}^{n} (1 + \sigma^2 \tau^{-2} i) \nabla_i \bar{\ell}_n(\vartheta_{0:n}) \right|$$

$$\leq C_{10} \tau.$$

Applying (23) to the special case of $\vartheta_{0:n} = \theta^{[n+1]}$, making use of (17) and (19), we infer the existence of finite constants C_{11} and C_{12} such that

$$\sup_{\theta \in K_4} |\tau^{-2} \Sigma^{-1} (\breve{\theta}_n^F - \theta) - \nabla \ell_n(\theta)| \leq C_{11} \tau + \sup_{\theta \in K_4} \sigma^2 \tau^{-2} \sum_{i=1}^{n} i |\nabla_i \ell_n(\theta^{[n+1]})|$$

(24)
$$\leq C_{12}(\tau + \sigma^2 \tau^{-2}),$$

which establishes (20). To show (21), we write

$$\sum_{k=1}^{n} \{\breve{V}_k^P\}^{-1} (\breve{\theta}_k^F - \breve{\theta}_{k-1}^F)$$

(25)
$$= \sum_{k=1}^{n} \tau^{-2} \Sigma^{-1} (\breve{\theta}_k^F - \breve{\theta}_{k-1}^F) + \tau^{-2} \sum_{k=1}^{n} (\{\tau^{-2} \breve{V}_k^P\}^{-1} - \Sigma^{-1})(\breve{\theta}_k^F - \breve{\theta}_{k-1}^F)$$

$$= \tau^{-2} \Sigma^{-1} (\breve{\theta}_n^F - \theta) + \tau^{-2} \sum_{k=1}^{n} (\{\tau^{-2} \breve{V}_k^P\}^{-1} - \Sigma^{-1})(\breve{\theta}_k^F - \breve{\theta}_{k-1}^F).$$

We note that (24) implies the existence of a bound

(26)
$$|\breve{\theta}_n^F - \theta| \leq C_{13} \tau^2.$$

Combining (25), (24) and (26), we deduce

$$\sup_{\theta \in K_4} \left| \sum_{k=1}^{n} \{\breve{V}_k^P\}^{-1} (\breve{\theta}_k^F - \breve{\theta}_{k-1}^F) - \nabla \ell_n(\theta) \right|$$

$$\leq C_{14}(\tau + \sigma^2 \tau^{-2}) + C_{15} \sum_{k=1}^{n} \sup_{\theta \in K_4} |\{\tau^{-2} \breve{V}_k^P\}^{-1} - \Sigma^{-1}|$$

for finite constants C_{14} and C_{15}. For invertible matrices A and B, we have the bound

(27)
$$|A^{-1} - B^{-1}| \leq |B^{-1}|^2 (1 - |(B - A)B^{-1}|)^{-1} |B - A|$$

provided that $|(B - A)B^{-1}| < 1$. Applying (27) with $A = \{\tau^{-2} \breve{V}_k^P\}$ and $B = \Sigma$, we see that the theorem will be proved once it is shown that

$$\sup_{\theta \in K_4} |\tau^{-2} \breve{V}_k^P - \Sigma| \leq C_0(\tau + \sigma^2 \tau^{-2}).$$

Now, it is easy to check that

$$\tau^{-2}\breve{V}_n^P - \Sigma = \tau^{-2}\breve{\mathbb{E}}_{\theta^{[n+1]},\sigma,\tau}[(\breve{\Theta}_{n-1} - \theta)(\breve{\Theta}_{n-1} - \theta)' \mid \breve{Y}_{1:n-1} = y_{1:n-1}^*] - \Sigma$$
$$- \tau^{-2}(\breve{\theta}_{n-1}^F - \theta)(\breve{\theta}_{n-1}^F - \theta)' + \sigma^2\tau^{-2}\Sigma.$$

Applying Theorem 1 again with $h(\vartheta_{0:n}) = (\vartheta_n - \theta)(\vartheta_n - \theta)'$, and making use of (26), we obtain

$$\sup_{\theta \in K_4} \left| \tau^{-2}\breve{\mathbb{E}}_{\theta^{[n+1]},\sigma,\tau}[(\breve{\Theta}_n - \theta)(\breve{\Theta}_n - \theta)' \mid \breve{Y}_{1:n} = y_{1:n}^*] - \Sigma \right|$$

(28)
$$\leq C_{16}(\tau + \sigma^2\tau^{-2}),$$

which completes the proof. □

The two approximations to the derivative of the log likelihood in (20) and (21) are asymptotically equivalent in the theoretical framework of this paper. However, numerical considerations may explain why (21) has been preferred in practical applications. To be concrete, we suppose henceforth that numerical filtering will be carried out using the basic sequential Monte Carlo method presented as Algorithm 2. Sequential Monte Carlo provides a flexible and widely used class of filtering algorithms, with many variants designed to improve numerical efficiency [Cappé, Godsill and Moulines (2007)]. The relatively simple sequential Monte Carlo method in Algorithm 2 has, however, been found adequate for previous data analyses using iterated filtering [Ionides, Bretó and King (2006); King et al. (2008); Bretó et al. (2009); He, Ionides and King (2010); Laneri et al. (2010)].

When carrying out filtering via sequential Monte Carlo, the resampling involved has a consequence that all surviving particles can descend from only few recent ancestors. This phenomenon, together with the resulting shortage of diversity in the Monte Carlo sample, is called *particle depletion* and can be a major obstacle for the implementation of sequential Monte Carlo techniques [Arulampalam et al. (2002); Andrieu, Doucet and Holenstein (2010)]. The role of the added variation on the scale of σ_m in the iterated filtering algorithm is to rediversify the particles and hence to combat particle depletion. Mixing considerations suggest that the new information about θ in the nth observation may depend only weakly on $y_{1:n-k}^*$ for sufficiently large k [Jensen and Petersen (1999)]. The actual Monte Carlo particle diversity of the filtering distribution, based on (21), may therefore be the best guide when sequentially estimating the derivative of the log likelihood. Future theoretical work on iterated filtering algorithms should study formally the role of mixing, to investigate this heuristic argument. However, the theory presented in Theorems 3, 4 and 5 does formally support using a limit random walk perturbations without any mixing conditions. Two influential previous proposals to use stochastic perturbations to reduce numerical instabilities arising in plug-and-play inference for POMPS via sequential Monte Carlo [Kitagawa (1998); Liu and West (2001)] lack even this level of theoretical support.

Algorithm 2 A basic sequential Monte Carlo procedure for a discrete-time Markov process. For the unperturbed model, set $\tilde{X}_n = X_n$, $\tilde{Y}_n = Y_n$, $\tilde{f} = f$ and $\tilde{\theta} = \theta$. For the stochastically perturbed model, set $\tilde{X}_n = (\check{X}_n, \check{\Theta}_n)$, $\tilde{Y}_n = Y_n$, $\tilde{f} = g$ and $\tilde{\theta} = (\theta^{[N+1]}, \sigma, \tau)$. It is neither necessary nor computationally optimal to draw from the density $\tilde{f}_{\tilde{X}_n | \tilde{X}_{n-1}}$ in step 3 [e.g., Arulampalam et al. (2002)] however only this choice leads to the plug-and-play property. The resampling in step 5 is taken to follow a multinomial distribution to build on previous theoretical results making this assumption [Del Moral and Jacod (2001); Crisan and Doucet (2002)]. An alternative is the systematic procedure in Arulampalam et al. [(2002), Algorithm 2] which has less Monte Carlo variability. We support the use of systematic sampling in practice, and we suppose that all our results would continue to hold in such situations.

Input:

- POMP model described by a generic density $\tilde{f}$ having parameter vector $\tilde{\theta}$ and corresponding to a Markov process $\tilde{X}_{0:N}$, observation process $\tilde{Y}_{1:N}$, and data $y_{1:N}^*$
- Number of particles, J

Procedure:

1 initialize filter particles $\tilde{X}_{0,j}^F \sim \tilde{f}_{\tilde{X}_0}(\tilde{x}_0; \tilde{\theta})$ for j in $1:J$
2 for n in $1:N$
3   for j in $1:J$ draw prediction particles $\tilde{X}_{n,j}^P \sim \tilde{f}_{\tilde{X}_n | \tilde{X}_{n-1}}(\tilde{x}_n \mid \tilde{X}_{n-1,j}^F; \tilde{\theta})$
4   set $w(n,j) = \tilde{f}_{\tilde{Y}_n | \tilde{X}_n}(y_n^* \mid \tilde{X}_{n,j}^P; \tilde{\theta})$
5   draw $k_1, \ldots, k_J$ such that $\mathbb{P}\{k_j = i\} = w(n,i) / \sum_\ell w(n,\ell)$
6   set $\tilde{X}_{n,j}^F = \tilde{X}_{n,k_j}^P$
7 end for

To calculate Monte Carlo estimates of the quantities in (19), we apply Algorithm 2 with $\tilde{f} = g$, $\tilde{X}_n = (\check{X}_n, \check{\Theta}_n)$, $\tilde{\theta} = (\theta^{[N+1]}, \sigma, \tau)$ and J particles. We write $\tilde{X}_{n,j}^P = (X_{n,j}^P, \Theta_{n,j}^P)$ and $\tilde{X}_{n,j}^F = (X_{n,j}^F, \Theta_{n,j}^F)$ for the Monte Carlo samples from the prediction and filtering and calculations in steps 3 and 6 of Algorithm 2. Then we define

$$(29) \quad \begin{aligned} \tilde{\theta}_n^F &= \tilde{\theta}_n^F(\theta, \sigma, \tau, J) = \frac{1}{J} \sum_{j=1}^{J} \Theta_{n,j}^F, \\ \tilde{V}_n^P &= \tilde{V}_n^P(\theta, \sigma, \tau, J) = \frac{1}{J-1} \sum_{j=1}^{J} (\Theta_{n,j}^P - \tilde{\theta}_{n-1}^F)(\Theta_{n,j}^P - \tilde{\theta}_{n-1}^F)'. \end{aligned}$$

In practice, a reduction in Monte Carlo variability is possible by modifying (29) to estimate θ_n^F and V_n^P from weighted particles prior to resampling [Chopin (2004)].

We now present, as Theorem 4, an analogue to Theorem 3 in which the filtering means and prediction variances are replaced by their Monte Carlo counterparts. The stochasticity in Theorem 4 is due to Monte Carlo variability, conditional on the data $y_{1:N}^*$, and we write E_{MC} and Var_{MC} to denote Monte Carlo means and variances. The Monte Carlo random variables required to implement Algorithm 2 are presumed to be drawn independently each time the algorithm is evaluated. To control the Monte Carlo bias and variance, we assume:

(A5) For each n and any compact set $K_6 \subset \mathbb{R}^p$,

$$\sup_{\theta \in K_6, x \in \mathfrak{X}} f_{Y_n | X_n}(y_n^* \mid x_n; \theta) < \infty.$$

THEOREM 4. *Let* $\{\sigma_m\}$, $\{\tau_m\}$ *and* $\{J_m\}$ *be positive sequences with* $\tau_m \to 0$, $\sigma_m \tau_m^{-1} \to 0$ *and* $\tau_m J_m \to \infty$. *Define* $\widetilde{\theta}_{n,m}^F = \widetilde{\theta}_n^F(\theta, \sigma_m, \tau_m, J_m)$ *and* $\widetilde{V}_{n,m}^P = \widetilde{V}_n^P(\theta, \sigma_m, \tau_m, J_m)$ *via* (29). *Suppose conditions* (A4) *and* (A5) *and let* K_7 *be an arbitrary compact subset of* $\mathbb{R}^p$. *Then,*

$$(30) \qquad \lim_{m \to \infty} \sup_{\theta \in K_7} \left| E_{MC}\left[\sum_{n=1}^N (\widetilde{V}_{n,m}^P)^{-1}(\widetilde{\theta}_{n,m}^F - \widetilde{\theta}_{n-1,m}^F) \right] - \nabla \ell_N(\theta) \right| = 0,$$

$$(31) \qquad \lim_{m \to \infty} \sup_{\theta \in K_7} \left| \tau_m^2 J_m \, \mathrm{Var}_{MC}\left(\sum_{n=1}^N (\widetilde{V}_{n,m}^P)^{-1}(\widetilde{\theta}_{n,m}^F - \widetilde{\theta}_{n-1,m}^F) \right) \right| < \infty.$$

PROOF. Let K_8 be a compact subset of $\mathbb{R}^2$ containing $\{(\sigma_m, \tau_m), m = 1, 2, \ldots\}$. Set $\theta \in K_7$ and $(\sigma, \tau) \in K_8$. Making use of the definitions in (19) and (29), we construct $u_n = (\check{\theta}_n^F - \check{\theta}_{n-1}^F)/\tau$ and $v_n = \check{V}_n^P/\tau^2$, with corresponding Monte Carlo estimates $\widetilde{u}_n = (\widetilde{\theta}_n^F - \widetilde{\theta}_{n-1}^F)/\tau$ and $\widetilde{v}_n = \widetilde{V}_n^P/\tau^2$. We look to apply Theorem 7 (presented in Section A.2) with $\widetilde{f} = g$, $\check{X}_n = (\check{X}_n, \check{\Theta}_n)$, $\check{Y}_n = \check{Y}_n$, $\widetilde{\theta} = (\theta^{[n+1]}, \sigma, \tau)$, J particles, and

$$\phi(\check{X}_n, \check{\Theta}_n) = (\check{\Theta}_n - \theta)/\tau.$$

Using the notation from (44), we have $u_n = \phi_n^F - \phi_{n-1}^F$ and $\widetilde{u}_n = \widetilde{\phi}_n^F - \widetilde{\phi}_{n-1}^F$. By assumption, $\kappa(u)$ is supported on some set $\{u : |u| < C_{17}\}$ from which we derive the bound $|\phi(\check{X}_n, \check{\Theta}_n)| \leq C_{17}(1 + n\sigma/\tau)$. Theorem 7 then provides for the existence of a C_{18} and C_{19} such that

$$(32) \qquad\qquad E_{MC}[|\widetilde{u}_n - u_n|^2] \leq C_{18}/J,$$

$$(33) \qquad\qquad |E_{MC}[\widetilde{u}_n - u_n]| \leq C_{19}/J.$$

The explicit bounds in (47) of Theorem 7, together with (A4) and (A5), assure us that $C_{18} = C_{18}(\theta, \sigma, \tau)$ and $C_{19} = C_{19}(\theta, \sigma, \tau)$ can be chosen so that (32) and

(33) hold uniformly over $(\theta, \sigma, \tau) \in K_7 \times K_8$. The same argument applied to $v_n = V_n^P/\tau^2$ and $\widetilde{v}_n = \widetilde{V}_n^P/\tau^2$ gives

$$(34) \qquad |E_{MC}[\widetilde{v}_n - v_n]| \le C_{20}/J, \qquad E_{MC}[|\widetilde{v}_n - v_n|^2] \le C_{21}/J$$

uniformly over $(\theta, \sigma, \tau) \in K_7 \times K_8$. We now proceed to carry out a Taylor series expansion:

$$(35) \qquad \widetilde{v}_n^{-1}\widetilde{u}_n = v_n^{-1}u_n + v_n^{-1}(\widetilde{u}_n - u_n)v_n^{-1}(\widetilde{v}_n - v_n)v_n^{-1}u_n + R_3,$$

where $|R_3| < C_{22}(|\widetilde{u}_n - u_n|^2 + |\widetilde{v}_n - v_n|^2)$ for some constant C_{22}. The existence of such a C_{22} is guaranteed since the determinant of v_n is bounded away from zero. Taking expectations of both sides of (35) and applying (32)–(34) gives

$$(36) \qquad |E_{MC}[\widetilde{v}_n^{-1}\widetilde{u}_n] - v_n^{-1}u_n| \le C_{23}/J$$

for some constant $C_{23} < \infty$. Another Taylor series expansion,

$$\widetilde{v}_n^{-1}\widetilde{u}_n = v_n^{-1}u_n + R_4$$

with $|R_4| < C_{24}(|\widetilde{u}_n - u_n| + |\widetilde{v}_n - v_n|)$ implies

$$(37) \qquad \mathrm{Var}_{MC}(\widetilde{v}_n^{-1}\widetilde{u}_n) \le C_{25}/J.$$

Rewriting (36) and (37), defining $\check{\theta}_{n,m}^F = \check{\theta}_n^F(\theta, \sigma_m, \tau_m)$ and $\check{V}_{n,m}^P = \check{V}_n^P(\theta, \sigma_m, \tau_m)$, we deduce that

$$(38) \quad \tau_m J_m |E_{MC}[(\widetilde{V}_{n,m}^P)^{-1}(\widetilde{\theta}_{n,m}^F - \widetilde{\theta}_{n-1,m}^F)] - (\check{V}_{n,m}^P)^{-1}(\check{\theta}_{n,m}^F - \check{\theta}_{n-1,m}^F)| \le C_{23}$$

and

$$(39) \qquad \tau_m^2 J_m \, \mathrm{Var}_{MC}[(\widetilde{V}_{n,m}^P)^{-1}(\widetilde{\theta}_{n,m}^F - \widetilde{\theta}_{n-1,m}^F)] \le C_{25}.$$

Combining (38) with Theorem 3, and summing over n, leads to (30). Summing (39) over n justifies (30). □

Theorem 4 suggests that a Monte Carlo method which leans on Theorem 3 will require a sequence of Monte Carlo sample sizes, J_m, which increases faster than τ_m^{-1}. Even with $\tau_m J_m \to \infty$, we see from (31) that the estimated derivative in (30) may have increasing Monte Carlo variability as $m \to \infty$. Theorem 5 gives an example of a stochastic approximation procedure, defined by the recursive sequence $\hat{\theta}_m$ in (40), that makes use of the Monte Carlo estimates studied in Theorem 4. Because each step of this recursion involves an application of the filtering procedure in Algorithm 2, we call (40) an iterated filtering algorithm. The rate assumptions in Theorem 5 are satisfied, for example, by $a_m = m^{-1}$, $\tau_m^2 = m^{-1}$, $\sigma_m^2 = m^{-(1+\delta)}$ and $J_m = m^{(\delta+1/2)}$ for $\delta > 0$.

THEOREM 5. *Let $\{a_m\}$, $\{\sigma_m\}$, $\{\tau_m\}$ and $\{J_m\}$ be positive sequences with $\tau_m \to 0$, $\sigma_m \tau_m^{-1} \to 0$, $J_m \tau_m \to \infty$, $a_m \to 0$, $\sum_m a_m = \infty$ and $\sum_m a_m^2 J_m^{-1} \tau_m^{-2} < \infty$. Specify a recursive sequence of parameter estimates $\{\hat{\theta}_m\}$ by*

$$(40) \qquad \hat{\theta}_{m+1} = \hat{\theta}_m + a_m \sum_{n=1}^{N} (\widetilde{V}_{n,m}^P)^{-1} (\widetilde{\theta}_{n,m}^F - \widetilde{\theta}_{n-1,m}^F),$$

where $\widetilde{\theta}_{n,m}^F = \widetilde{\theta}_n^F(\hat{\theta}_m, \sigma_m, \tau_m, J_m)$ and $\widetilde{V}_{n,m}^P = \widetilde{V}_{n,m}^P(\hat{\theta}_m, \sigma_m, \tau_m, J_m)$ are defined in (29) *via an application of Algorithm* 2. *Assuming* (A4), (B1) *and* (B2), $\lim_{m \to \infty} \hat{\theta}_m = \hat{\theta}$ *with probability one.*

PROOF.   Theorem 5 follows directly from a general stochastic approximation result, Theorem 6 of Section A.1, applied to $\ell_N(\theta)$. Conditions (B4) and (B5) of Theorem 6 hold from Theorem 4 and the remaining assumptions of Theorem 6 hold by hypothesis.   □

4. Discussion.   One alternative approach to likelihood maximization for POMP models involves plugging the (log) likelihood estimate from a particle filter directly into a general-purpose stochastic optimization algorithm such as Simultaneous Perturbation Stochastic Approximation (SPSA), Kiefer–Wolfowitz or stochastic Nelder–Mead [Spall (2003)]. An advantage of iterated filtering, and other methods based on particle filtering with parameter perturbations [Kitagawa (1998); Liu and West (2001)], is that the many thousands of particles are simultaneously exploring the parameter space and evaluating an approximation to the likelihood. When the data are a long time series, the perturbed parameters can make substantial progress toward plausible parameter values in the course of one filtering operation. From the point of view of a general-purpose stochastic optimization algorithm, carrying out one filtering operation (which can be a significant computational burden in many practical situations) yields only one function evaluation of the likelihood.

The practical applicability of particle filters may be explained by their numerical stability on models possessing a mixing property [e.g., Crisan and Doucet (2002)]. The sequential Monte Carlo analysis in Theorem 4 did not address the convergence of iterated filtering under mixing assumptions as the number of observations, N, increases. We therefore studied experimentally the numerical stability of the Monte Carlo estimate of the derivative of the log likelihood in equation (30). The role of mixing arises regardless of the dimension of the state space, the dimension of the parameter space, the nonlinearity of the system, or the non-Gaussianity of the system. This suggests that a simple linear Gaussian example may be representative of behavior on more complex models. Specifically, we considered a POMP model defined by a scalar Markov process $X_n = \theta X_{n-1} + \varepsilon_n$, with $X_0 = 0$, and a scalar observation process $Y_n = X_n + \eta_n$. Here, $\{\varepsilon_n\}$ and $\{\eta_n\}$ were taken to be sequences of independent Gaussian random variables having zero mean and unit variance. We
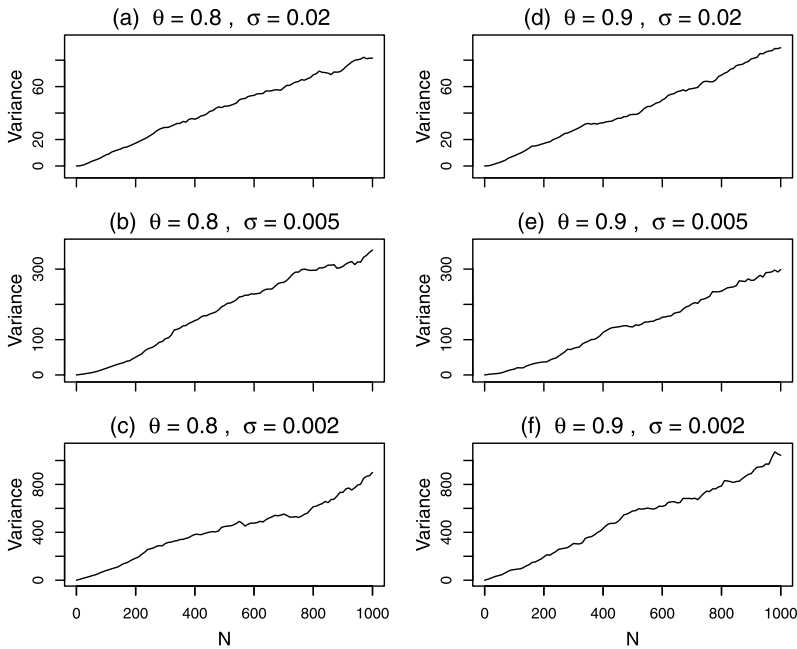
FIG. 1. *Monte Carlo variance of the derivative approximation in* (30) *for varying values of* θ *and* σ.

fixed the true parameter value as $\theta^* = 0.8$ and we evaluated $\nabla \ell_N(\theta)$ at $\theta = \theta^*$ and $\theta = 0.9$ using a Kalman filter (followed by a finite difference derivative computation) and via the sequential Monte Carlo approximation in (30) using $J = 1,000$ particles. We investigated $\sigma \in \{0.002, 0.005, 0.02\}$, chosen to include a small value where Monte Carlo variance dominates, a large value where bias dominates, and an intermediate value; we then fixed $\tau = 20\sigma$.

Figures 1 and 2 show how the Monte Carlo variance and the bias vary with N for each value of σ. These quantities were evaluated from 100 realizations of the model, with 5 replications of the filtering operation per realization, via standard unbiased estimators. We see from Figure 1 that the Monte Carlo variance increases approximately linearly with N. This numerical stability is a substantial improvement on the exponential bound guaranteed by Theorem 7. The ordinate values in Figure 1 show that, as anticipated from Theorem 4, the variance increases as σ decreases. Figure 2 shows that the bias diminishes as σ decreases and is small when θ is close to θ^*. When θ is distant from θ^*, the perturbed parameter values migrate toward θ^* during the course of the filtering operation, as shown in Figure 3(b). Once the perturbed parameters have arrived in the vicinity of θ^*, the sum in (30) approximates the derivative of the log likelihood at θ^* rather than at θ. Figure 3(a) demonstrates the resulting bias in the estimate of $\nabla \ell_N(\theta)$. However, this bias may be helpful, rather than problematic, for the convergence of the iterated filtering algorithm. The update in (40) is a weighted average of the filtered means
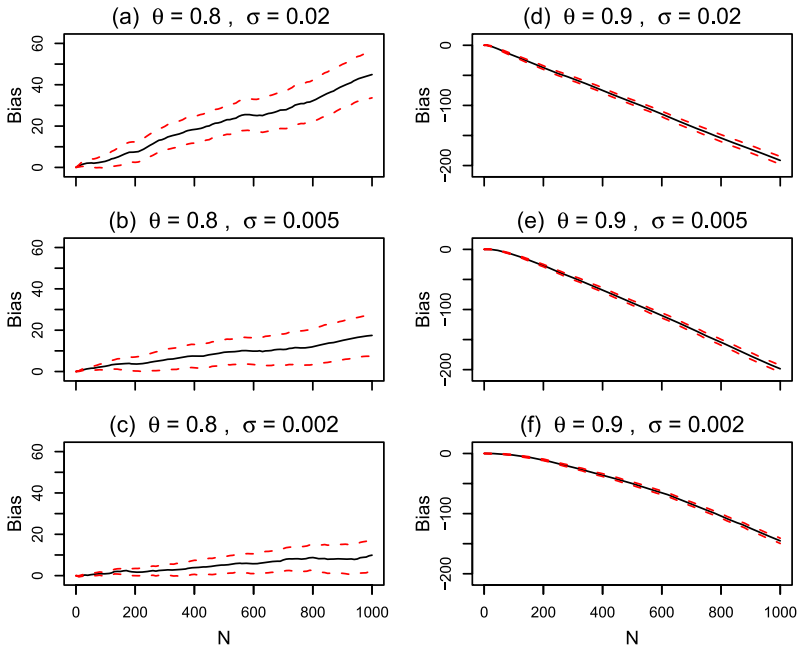
FIG. 2.   *Bias of the derivative approximation in* (30) *for varying values of θ and σ. Dashed lines show pointwise 95% Monte Carlo confidence intervals.*

of the perturbed parameters. Heuristically, if the perturbed parameters successfully locate a neighborhood of θ^* then this will help to generate a good update for the iterated filtering algorithm. The utility of perturbed parameter values to identify a neighborhood of θ^*, in addition to estimating a derivative, does not play a role in
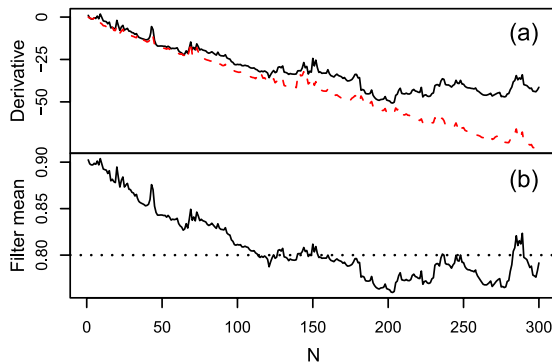

FIG. 3.   *One realization from the simulation study, with* $\theta = 0.9$ *and* $\sigma = 0.005$. (a) *The estimate of* $\nabla \ell_N(\theta)$ *using* (30) (*solid line*) *and calculated directly* (*dashed line*). (b) *The filter mean* $\tilde{\theta}_N^F$ (*solid line*) *approaching the vicinity of the true parameter value* $\theta^* = 0.8$ (*dotted line*).

our asymptotic justification of iterated filtering. However, it may contribute to the nonasymptotic properties of the method at early iterations.

APPENDIX: SOME STANDARD RESULTS ON SEQUENTIAL MONTE CARLO AND STOCHASTIC APPROXIMATION THEORY

We state some basic theorems that we use to prove Theorems 2, 4 and 5, both for completeness and because we require minor modifications of the standard results. Our goal is not to employ the most recent results available in these research areas, but rather to show that some fundamental and well-known results from both areas can be combined with our Theorems 1 and 3 to synthesize a new theoretical understanding of iterated filtering and iterated importance sampling.

A.1. A version of a standard stochastic approximation theorem. We present, as Theorem 6, a special case of Theorem 2.3.1 of Kushner and Clark (1978). For variations and developments on this result, we refer the reader to Kushner and Yin (2003), Spall (2003), Andrieu, Moulines and Priouret (2005) and Maryak and Chin (2008). In particular, Theorem 2.3.1 of Kushner and Clark (1978) is similar to Theorem 4.1 of Spall (2003) and to Theorem 2.1 of Kushner and Yin (2003).

THEOREM 6. *Let $\ell(\theta)$ be a continuously differentiable function $\mathbb{R}^p \to \mathbb{R}$ and let $\{D_m(\theta), m \geq 1\}$ be a sequence of independent Monte Carlo estimators of the vector of partial derivatives $\nabla \ell(\theta)$. Define a sequence $\{\hat{\theta}_m\}$ recursively by $\hat{\theta}_{m+1} = \hat{\theta}_m + a_m D_m(\hat{\theta}_m)$. Assume* (B1) *and* (B2) *of Section 2 together with the following conditions*:

(B3) $a_m > 0, a_m \to 0, \sum_m a_m = \infty$.
(B4) $\sum_m a_m^2 \sup_{|\theta| < r} \mathrm{Var}_{MC}(D_m(\theta)) < \infty$ for every $r > 0$.
(B5) $\lim_{m \to \infty} \sup_{|\theta| < r} |E_{MC}[D_m(\theta)] - \nabla \ell(\theta)| = 0$ for every $r > 0$.

Then $\hat{\theta}_m$ converges to $\hat{\theta} = \arg \max \ell(\theta)$ with probability one.

PROOF. The most laborious step in deducing Theorem 6 from Kushner and Clark (1978) is to check that (B1)–(B5) imply that, for all $\varepsilon > 0$,

$$(41) \qquad \lim_{n \to \infty} \mathbb{P}\left[\sup_{j \geq 1} \left| \sum_{m=n}^{n+j} a_m \{D_m(\hat{\theta}_m) - E_{MC}[D_m(\hat{\theta}_m) \mid \hat{\theta}_m]\} \right| \geq \varepsilon \right] = 0,$$

which in turn implies condition A2.2.4″ and hence A2.2.4 of Kushner and Clark (1978). To show (41), we define $\xi_m = D_m(\hat{\theta}_m) - E_{MC}[D_m(\hat{\theta}_m) \mid \hat{\theta}_m]$ and

$$(42) \qquad \xi_m^k = \begin{cases} \xi_m, & \text{if } |\hat{\theta}_m| \leq k, \\ 0, & \text{if } |\hat{\theta}_m| > k. \end{cases}$$

Define processes $\{M_j^n = \sum_{m=n}^{n+j} a_m \xi_m, \, j \geq 0\}$ and $\{M_j^{n,k} = \sum_{m=n}^{n+j} a_m \xi_m^k, \, j \geq 0\}$ for each k and n. These processes are martingales with respect to the filtration defined by the Monte Carlo stochasticity. From the Doob–Kolmogorov martingale inequality [e.g., Grimmett and Stirzaker (1992)],

$$(43) \qquad \mathbb{P}\Big[\sup_j |M_j^{n,k}| \geq \varepsilon\Big] \leq \frac{1}{\varepsilon^2} \sum_{m=n}^{\infty} a_m^2 \sup_{|\theta| < k} \mathrm{Var}_{MC}(D_m(\theta)).$$

Define events $F_n = \{\sup_j |M_j^n| \geq \varepsilon\}$ and $F_{n,k} = \{\sup_j |M_j^{n,k}| \geq \varepsilon\}$. It follows from (B4) and (43) that $\lim_{n\to\infty} \mathbb{P}\{F_{n,k}\} = 0$ for each k. In light of the nondivergence assumed in (B2), this implies $\lim_{n\to\infty} \mathbb{P}\{F_n\} = 0$ which is exactly (41).

To expand on this final assertion, let $\Omega = \{\sup_m |\hat{\theta}_m| < \infty\}$ and $\Omega_k = \{\sup_m |\hat{\theta}_m| < k\}$. Assumption (B2) implies that $\mathbb{P}(\Omega) = 1$. Since the sequence of events $\{\Omega_k\}$ is increasing up to Ω, we have $\lim_{k\to\infty} \mathbb{P}(\Omega_k) = \mathbb{P}(\Omega) = 1$. Now observe that $\Omega_k \cap F_{n,j} = \Omega_k \cap F_n$ for all $j \geq k$, as there is no truncation of the sequence $\{\xi_m^j, m = 1, 2, \ldots\}$ for outcomes in Ω_k when $j \geq k$. Then,

$$\begin{aligned}
\lim_{n\to\infty} \mathbb{P}[F_n] &\leq \lim_{n\to\infty} \mathbb{P}[F_n \cap \Omega_k] + 1 - \mathbb{P}[\Omega_k] \\
&= \lim_{n\to\infty} \mathbb{P}[F_{n,k} \cap \Omega_k] + 1 - \mathbb{P}[\Omega_k] \\
&\leq \lim_{n\to\infty} \mathbb{P}[F_{n,k}] + 1 - \mathbb{P}[\Omega_k] \\
&= 1 - \mathbb{P}[\Omega_k].
\end{aligned}$$

Since k can be chosen to make $1 - \mathbb{P}[\Omega_k]$ arbitrarily small, it follows that $\lim_{n\to\infty} \mathbb{P}[F_n] = 0$. $\square$

A.2. Some standard results on sequential Monte Carlo and importance sampling. A general convergence result on sequential Monte Carlo combining results by Crisan and Doucet (2002) and Del Moral and Jacod (2001) is stated in our notation as Theorem 7 below. The theorem is stated for a POMP model with generic density $\tilde{f}$, parameter vector $\tilde{\theta}$, Markov process $\tilde{X}_{0:N}$ taking values in $\tilde{\mathfrak{X}}^{N+1}$, observation process $\tilde{Y}_{1:N}$ taking values in $\mathfrak{Y}^N$, and data $y_{1:N}^*$. For application to the unperturbed model one sets $\tilde{f} = f$, $\tilde{X}_n = X_n$, $\tilde{\mathfrak{X}} = \mathfrak{X}$, $\tilde{Y}_n = Y_n$ and $\tilde{\theta} = \theta$. For application to the stochastically perturbed model one sets $\tilde{f} = g$, $\tilde{X}_n = (\check{X}_n, \check{\Theta}_n)$, $\tilde{\mathfrak{X}} = \mathfrak{X} \times \mathbb{R}^p$, $\tilde{Y}_n = \check{Y}_n$ and $\tilde{\theta} = (\theta^{[N+1]}, \sigma, \tau)$. When applying Theorem 7 in the context of Theorem 4, the explicit expressions for the constants C_{26} and C_{27} are required to show that the bounds in (45) and (46) apply uniformly for a collection of models indexed by the approximation parameters $\{\tau_m\}$ and $\{\sigma_m\}$.

THEOREM 7 [Crisan and Doucet (2002); Del Moral and Jacod (2001)]. *Let $\tilde{f}$ be a generic density for a POMP model having parameter vector $\tilde{\theta}$, unobserved Markov process $\tilde{X}_{0:N}$, observation process $\tilde{Y}_{1:N}$ and data $y_{1:N}^*$. Define $\tilde{X}_{n,j}^F$ via*

applying Algorithm 2 with J particles. Assume that $\tilde{f}_{\tilde{Y}_n|\tilde{X}_n}(y_n^* \mid \tilde{x}_n; \tilde{\theta})$ is bounded as a function of $\tilde{x}_n$. For any $\phi : \tilde{\mathfrak{X}} \to \mathbb{R}$, denote the filtered mean of $\phi(\tilde{X}_n)$ and its Monte Carlo estimate by

$$(44) \quad \phi_n^F = \int \phi(\tilde{x}_n) \tilde{f}_{\tilde{X}_n|\tilde{Y}_{1:n}}(\tilde{x}_n \mid y_{1:n}^*; \tilde{\theta}) \, d\tilde{x}_n, \qquad \widetilde{\phi}_n^F = \frac{1}{J} \sum_{j=1}^J \phi(\tilde{X}_{n,j}^F).$$

There are constants C_{26} and C_{27}, independent of J, such that

$$(45) \qquad E_{MC}[(\widetilde{\phi}_n^F - \phi_n^F)^2] \leq \frac{C_{26} \sup_{\tilde{x}} |\phi(\tilde{x})|^2}{J},$$

$$(46) \qquad |E_{MC}[\widetilde{\phi}_n^F - \phi_n^F]| \leq \frac{C_{27} \sup_{\tilde{x}} |\phi(\tilde{x})|}{J}.$$

Specifically, C_{26} and C_{27} can be written as linear functions of 1 and $\eta_{n,1}, \ldots, \eta_{n,n}$ defined as

$$(47) \qquad \eta_{n,i} = \prod_{k=n-i+1}^n \left(\frac{\sup_{\tilde{x}_k} \tilde{f}_{\tilde{Y}_k|\tilde{X}_k}(y_k^* \mid \tilde{x}_k; \tilde{\theta})}{\tilde{f}_{\tilde{Y}_k|\tilde{Y}_{1:k-1}}(y_k^* \mid y_{1:k-1}^*; \tilde{\theta})} \right)^2.$$

PROOF. Theorem 2 of Crisan and Doucet (2002) derived (45), and here we start by focusing on the assertion that the constant C_{26} in equation (45) can be written as a linear function of 1 and the quantities $\eta_{n,1}, \ldots, \eta_{n,n}$ defined in (47). This was not explicitly mentioned by Crisan and Doucet (2002) but is a direct consequence of their argument. Crisan and Doucet [(2002), Section V] constructed the following recursion, for which $c_{n|n}$ is the constant C_{26} in equation (45). For $n = 1, \ldots, N$ and $c_{0|0} = 0$, define

$$(48) \qquad c_{n|n} = (\sqrt{C} + \sqrt{\bar{c}_{n|n}})^2,$$

$$(49) \qquad \bar{c}_{n|n} = 4 c_{n|n-1} \left(\frac{\|\tilde{f}_{\tilde{Y}_n|\tilde{X}_n}\|}{\tilde{f}_{\tilde{Y}_n|\tilde{Y}_{1:n-1}}(y_n^* \mid y_{1:n-1}^*; \tilde{\theta})} \right)^2,$$

$$(50) \qquad c_{n|n-1} = (1 + \sqrt{c_{n-1|n-1}})^2,$$

where $\|\tilde{f}_{\tilde{Y}_n|\tilde{X}_n}\| = \sup_{\tilde{x}_n} \tilde{f}_{\tilde{Y}_n|\tilde{X}_n}(y_n^* \mid \tilde{x}_n; \tilde{\theta})$. Here, C is a constant that depends on the resampling procedure but not on the number of particles J. Now, (48)–(50) can be reformulated by routine algebra as

$$(51) \qquad c_{n|n} \leq K_1 + K_2 \bar{c}_{n|n},$$

$$(52) \qquad \bar{c}_{n|n} \leq K_3 q_n c_{n|n-1},$$

$$(53) \qquad c_{n|n-1} \leq K_4 + K_5 c_{n-1|n-1},$$

where $q_n = \|\tilde{f}_{\tilde{Y}_n|\tilde{X}_n}\|^2[\tilde{f}_{\tilde{Y}_n|\tilde{Y}_{1:n-1}}(y_n^* \mid y_{1:n-1}^*; \tilde{\theta})]^{-2}$ and $K_1, \ldots, K_5$ are constants which do not depend on $\tilde{f}$, $\tilde{\theta}$, $y_{1:N}^*$ or J. Putting (52) and (53) into (51),

$$ c_{n|n} \leq K_1 + K_2 K_3 q_n c_{n|n-1} $$

(54)

$$ \leq K_1 + K_2 K_3 K_4 q_n + K_2 K_3 K_5 q_n c_{n-1|n-1}. $$

Since $\eta_{n,i} = q_n \eta_{n-1,i}$ for $i < n$, and $\eta_{n,n} = q_n$, the required assertion follows from (54).

To show (46), we introduce the unnormalized filtered mean ϕ_n^U and its Monte Carlo estimate $\tilde{\phi}_n^U$, defined by

$$ (55) \quad \phi_n^U = \phi_n^F \prod_{k=1}^n \tilde{f}_{Y_k|Y_{1:k-1}}(y_k^* \mid y_{1:k-1}^*; \tilde{\theta}), \quad \tilde{\phi}_n^U = \tilde{\phi}_n^F \prod_{k=1}^n \frac{1}{J} \sum_{j=1}^J w(k, j), $$

where $w(k, j)$ is computed in step 4 of Algorithm 2 when evaluating $\tilde{\phi}_n^F$. Then, Del Moral and Jacod (2001) showed that

$$ (56) \qquad E_{MC}[\tilde{\phi}_n^U] = \phi_n^U, $$

$$ (57) \quad E_{MC}[(\tilde{\phi}_n^U - \phi_n^U)^2] \leq \frac{(n + 1) \sup_{\tilde{x}} |\phi(\tilde{x})|^2}{J} \prod_{k=1}^n \left(\sup_{\tilde{x}_k} \tilde{f}_{\tilde{Y}_k|\tilde{X}_k}(y_k^* \mid \tilde{x}_k; \tilde{\theta}) \right)^2. $$

We now follow an approach of Del Moral and Jacod [(2001), equation 3.3.14], by defining the unit function $1(\tilde{x}) = 1$ and observing that $\phi_n^F = \phi_n^U/1_n^U$ and $\tilde{\phi}_n^F = \tilde{\phi}_n^U/\tilde{1}_n^U$. Then (56) implies the identity

$$ (58) \qquad E_{MC}[\tilde{\phi}_n^F - \phi_n^F] = E_{MC}\left[(\tilde{\phi}_n^F - \phi_n^F)\left(1 - \frac{\tilde{1}_n^U}{1_n^U}\right) \right]. $$

Applying the Cauchy–Schwarz inequality to (58), making use of (45) and (57), gives (46). □

We now give a corollary to Theorem 7 for a latent variable model $(\tilde{X}, \tilde{Y})$, as defined in Section 2, having generic density $\tilde{f}$, parameter vector $\tilde{\theta}$, unobserved variable $\tilde{X}$ taking values in $\tilde{\mathfrak{X}}$, observed variable $\tilde{Y}$ taking values in $\mathfrak{Y}$, and data y^*. Importance sampling for such a model is a special case of sequential Monte Carlo, with $N = 1$ and no resampling step. We present and prove a separate result, which takes advantage of the simplified situation, to make Section 2 and the proof of Theorem 2 self-contained. In the context of Theorem 2, one sets $\tilde{f} = g$, $\tilde{X} = (\check{X}, \check{\Theta})$, $\tilde{\mathfrak{X}} = \mathfrak{X} \times \mathbb{R}^p$, $\tilde{Y}_n = \check{Y}$ and $\tilde{\theta} = (\theta, \tau)$.

COROLLARY 8. *Let $\tilde{f}$ be a generic density for the latent variable model $(\tilde{X}, \tilde{Y})$ with parameter vector $\tilde{\theta}$ and data y^*. Let $\{\tilde{X}_j, j = 1, \ldots, J\}$ be J independent Monte Carlo draws from $\tilde{f}_{\tilde{X}}(\tilde{x}; \tilde{\theta})$ and let $w_j = \tilde{f}_{\tilde{Y}|\tilde{X}}(y^* \mid \tilde{X}_j; \tilde{\theta})$. Letting*

$\phi : \tilde{\mathfrak{X}} \to \mathbb{R}$ *be a bounded function, write the conditional expectation of $\phi(\tilde{X})$ and its importance sampling estimate as*

$$(59) \qquad \phi^C = \int \phi(\tilde{x}_n) \tilde{f}_{\tilde{X}|\tilde{Y}}(\tilde{x} \mid y^*; \tilde{\theta}) \, d\tilde{x}, \quad \widetilde{\phi}^C = \frac{\sum_{j=1}^{J} w_j \phi(\tilde{X}_j)}{\sum_{j=1}^{J} w_j}.$$

Assume that $\tilde{f}_{\tilde{Y}|\tilde{X}}(y^* \mid \tilde{x}; \tilde{\theta})$ is bounded as a function of $\tilde{x}$. Then,

$$(60) \qquad \mathrm{Var}_{MC}(\widetilde{\phi}^C) \le \frac{4 \sup_{\tilde{x}} |\phi(\tilde{x})|^2 \sup_{\tilde{x}} (\tilde{f}_{\tilde{Y}|\tilde{X}}(y^* \mid \tilde{x}; \tilde{\theta}))^2}{J(\tilde{f}_{\tilde{Y}}(y^*; \tilde{\theta}))^2},$$

$$(61) \qquad |E_{MC}[\widetilde{\phi}^C] - \phi^C| \le \frac{2 \sup_{\tilde{x}} |\phi(\tilde{x})| \sup_{\tilde{x}} (\tilde{f}_{\tilde{Y}|\tilde{X}}(y^* \mid \tilde{x}; \tilde{\theta}))^2}{J(\tilde{f}_{\tilde{Y}}(y^*; \tilde{\theta}))^2}.$$

PROOF.   We introduce normalized weights $\hat{w}_j = w_j / \tilde{f}_{\tilde{Y}}(y^*; \tilde{\theta})$ and a normalized importance sampling estimator $\hat{\phi}^C = \frac{1}{J} \sum_{j=1}^{J} \hat{w}_j \phi(\tilde{X}_j)$ to compare to the self-normalized estimator in (59). It is not hard to check that $E_{MC}[\hat{\phi}^C] = \phi^C$ and $\mathrm{Var}_{MC}(\hat{\phi}^C) \le \frac{1}{J}(\sup_{\tilde{x}} |\phi(\tilde{x})| \sup_{\tilde{x}} \tilde{f}_{\tilde{Y}|\tilde{X}}(y^* \mid \tilde{x}; \tilde{\theta}) / \tilde{f}_{\tilde{Y}}(y^*; \tilde{\theta}))^2$. Now, $\mathrm{Var}_{MC}(\widetilde{\phi}^C) \le 2\{\mathrm{Var}_{MC}(\widetilde{\phi}^C - \hat{\phi}^C) + \mathrm{Var}_{MC}(\hat{\phi}^C)\}$ and

$$\mathrm{Var}_{MC}(\widetilde{\phi}^C - \hat{\phi}^C) \le E_{MC}\left[\left(\frac{(1/J)\sum_{j=1}^{J} \hat{w}_j \phi(\tilde{X}_j)}{(1/J)\sum_{j=1}^{J} \hat{w}_j} - \frac{1}{J}\sum_{j=1}^{J} \hat{w}_j \phi(\tilde{X}_j)\right)^2\right]$$

$$= E_{MC}\left[\left(\frac{(1/J)\sum_{j=1}^{J} \hat{w}_j \phi(\tilde{X}_j)}{(1/J)\sum_{j=1}^{J} \hat{w}_j}\right)^2 \left(1 - \frac{1}{J}\sum_{j=1}^{J} \hat{w}_j\right)^2\right]$$

$$\le \sup_{\tilde{x}} |\phi(\tilde{x})|^2 E_{MC}\left[1 - \frac{1}{J}\sum_{j=1}^{J} \hat{w}_j\right]^2$$

$$\le \frac{\sup_{\tilde{x}} |\phi(\tilde{x})|^2 \sup_{\tilde{x}} (\tilde{f}_{\tilde{Y}|\tilde{X}}(y^* \mid \tilde{x}; \tilde{\theta}))^2}{J(\tilde{f}_{\tilde{Y}}(y^*; \tilde{\theta}))^2}.$$

This demonstrates (60). To show (61), we write

$$|E_{MC}[\widetilde{\phi}^C] - \phi^C| = |E_{MC}[\widetilde{\phi}^C - \hat{\phi}^C]|$$

$$= \left| E_{MC}\left[(\widetilde{\phi}^C - E_{MC}[\widetilde{\phi}^C])\left(1 - \frac{1}{J}\sum_{j=1}^{J} \hat{w}_j\right)\right] \right|$$

$$\le \sqrt{\mathrm{Var}_{MC}(\widetilde{\phi}^C) \mathrm{Var}_{MC}\left(1 - \frac{1}{J}\sum_{j=1}^{J} \hat{w}_j\right)}$$

$$\leq \frac{2 \sup_{\tilde{x}} |\phi(\tilde{x})| \sup_{\tilde{x}} (\tilde{f}_{\tilde{Y}|\tilde{X}}(y^* \mid \tilde{x}; \tilde{\theta}))^2}{J(\tilde{f}_{\tilde{Y}}(y^*; \tilde{\theta}))^2}. \qquad \square$$

REFERENCES

ANDERSON, J. L. and COLLINS, N. (2007). Scalable implementations of ensemble filter algorithms for data assimilation. *Journal of Atmospheric and Oceanic Technology* **24** 1452–1463.

ANDRIEU, C., DOUCET, A. and HOLENSTEIN, R. (2010). Particle Markov chain Monte Carlo methods. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **72** 269–342.

ANDRIEU, C., MOULINES, É. and PRIOURET, P. (2005). Stability of stochastic approximation under verifiable conditions. *SIAM J. Control Optim.* **44** 283–312. MR2177157

ARULAMPALAM, M. S., MASKELL, S., GORDON, N. and CLAPP, T. (2002). A tutorial on particle filters for online nonlinear, non-Gaussian Bayesian tracking. *IEEE Trans. Signal Process.* **50** 174–188.

BAILEY, N. T. J. (1955). Some problems in the statistical analysis of epidemic data. *J. Roy. Statist. Soc. Ser. B* **17** 35–58; Discussion 58–68. MR0073090

BARTLETT, M. S. (1960). *Stochastic Population Models in Ecology and Epidemiology*. Methuen, London. MR0118550

BEYER, H.-G. (2001). *The Theory of Evolution Strategies*. Springer, Berlin. MR1930756

BJØRNSTAD, O. N. and GRENFELL, B. T. (2001). Noisy clockwork: Time series analysis of population fluctuations in animals. *Science* **293** 638–643.

BRETÓ, C., HE, D., IONIDES, E. L. and KING, A. A. (2009). Time series analysis via mechanistic models. *Ann. Appl. Stat.* **3** 319–348. MR2668710

CAPPÉ, O., GODSILL, S. and MOULINES, E. (2007). An overview of existing methods and recent advances in sequential Monte Carlo. *Proceedings of the IEEE* **95** 899–924.

CAPPÉ, O., MOULINES, E. and RYDÉN, T. (2005). *Inference in Hidden Markov Models*. Springer, New York. MR2159833

CAUCHEMEZ, S. and FERGUSON, N. M. (2008). Likelihood-based estimation of continuous-time epidemic models from time-series data: Application to measles transmission in London. *Journal of the Royal Society Interface* **5** 885–897.

CELEUX, G., MARIN, J.-M. and ROBERT, C. P. (2006). Iterated importance sampling in missing data problems. *Comput. Statist. Data Anal.* **50** 3386–3404. MR2236856

CHOPIN, N. (2004). Central limit theorem for sequential Monte Carlo methods and its application to Bayesian inference. *Ann. Statist.* **32** 2385–2411. MR2153989

CRISAN, D. and DOUCET, A. (2002). A survey of convergence results on particle filtering methods for practitioners. *IEEE Trans. Signal Process.* **50** 736–746. MR1895071

DEL MORAL, P. and JACOD, J. (2001). Interacting particle filtering with discrete observations. In *Sequential Monte Carlo Methods in Practice* (A. Doucet, N. de Freitas and N. J. Gordon, eds.) 43–75. Springer, New York. MR1847786

DIGGLE, P. J. and GRATTON, R. J. (1984). Monte Carlo methods of inference for implicit statistical models. *J. Roy. Statist. Soc. Ser. B* **46** 193–227. MR0781880

DURBIN, J. and KOOPMAN, S. J. (2001). *Time Series Analysis by State Space Methods*. *Oxford Statist. Sci. Ser.* **24**. Oxford Univ. Press, Oxford. MR1856951

ERGÜN, A., BARBIERI, R., EDEN, U. T., WILSON, M. A. and BROWN, E. N. (2007). Construction of point process adaptive filter algorithms for neural systems using sequential Monte Carlo methods. *IEEE Trans. Biomed. Eng.* **54** 419–428.

FERNÁNDEZ-VILLAVERDE, J. and RUBIO-RAMÍREZ, J. F. (2007). Estimating macroeconomic models: A likelihood approach. *Rev. Econom. Stud.* **74** 1059–1087. MR2353620

FERRARI, M. J., GRAIS, R. F., BHARTI, N., CONLAN, A. J. K., BJORNSTAD, O. N., WOLF-SON, L. J., GUERIN, P. J., DJIBO, A. and GRENFELL, B. T. (2008). The dynamics of measles in sub-Saharan Africa. *Nature* **451** 679–684.

GODSILL, S., VERMAAK, J., NG, W. and LI, J. (2007). Models and algorithms for tracking of maneuvering objects using variable rate particle filters. *Proceedings of the IEEE* **95** 925–952.

GRENFELL, B. T., BJORNSTAD, O. N. and FINKENSTÄDT, B. F. (2002). Dynamics of measles epidemics: Scaling noise, determinism, and predictability with the TSIR model. *Ecological Monographs* **72** 185–202.

GRIMMETT, G. R. and STIRZAKER, D. R. (1992). *Probability and Random Processes*, 2nd ed. The Clarendon Press/Oxford Univ. Press, New York. MR1199812

HE, D., IONIDES, E. L. and KING, A. A. (2010). Plug-and-play inference for disease dynamics: Measles in large and small towns as a case study. *Journal of the Royal Society Interface* **7** 271–283.

IONIDES, E. L., BRETÓ, C. and KING, A. A. (2006). Inference for nonlinear dynamical systems. *Proc. Natl. Acad. Sci. USA* **103** 18438–18443.

JENSEN, J. L. and PETERSEN, N. V. (1999). Asymptotic normality of the maximum likelihood estimator in state space models. *Ann. Statist.* **27** 514–535. MR1714719

JOHANNES, M., POLSON, N. and STROUD, J. (2009). Optimal filtering of jump diffusions: Extracting latent states from asset prices. *Review of Financial Studies* **22** 2759–2799.

JOHANSEN, A. M., DOUCET, A. and DAVY, M. (2008). Particle methods for maximum likelihood estimation in latent variable models. *Stat. Comput.* **18** 47–57. MR2416438

KEELING, M. and ROSS, J. (2008). On methods for studying stochastic disease dynamics. *Journal of the Royal Society Interface* **5** 171–181.

KENDALL, B. E., BRIGGS, C. J., MURDOCH, W. W., TURCHIN, P., ELLNER, S. P., MCCAULEY, E., NISBET, R. M. and WOOD, S. N. (1999). Why do populations cycle? A synthesis of statistical and mechanistic modeling approaches. *Ecology* **80** 1789–1805.

KIEFER, J. and WOLFOWITZ, J. (1952). Stochastic estimation of the maximum of a regression function. *Ann. Math. Statist.* **23** 462–466. MR0050243

KING, A. A., IONIDES, E. L., PASCUAL, M. and BOUMA, M. J. (2008). Inapparent infections and cholera dynamics. *Nature* **454** 877–880.

KITAGAWA, G. (1998). A self-organising state-space model. *J. Amer. Statist. Assoc.* **93** 1203–1215.

KUSHNER, H. J. and CLARK, D. S. (1978). *Stochastic Approximation Methods for Constrained and Unconstrained Systems*. *Appl. Math. Sci.* **26**. Springer, New York. MR0499560

KUSHNER, H. J. and YIN, G. G. (2003). *Stochastic Approximation and Recursive Algorithms and Applications*, 2nd ed. *Applications of Mathematics*: *Stochastic Modelling and Applied Probability* **35**. Springer, New York. MR1993642

LANERI, K., BHADRA, A., IONIDES, E. L., BOUMA, M., YADAV, R., DHIMAN, R. and PASCUAL, M. (2010). Forcing versus feedback: Epidemic malaria and monsoon rains in NW India. *PLoS Comput. Biol.* **6** e1000898.

LIU, J. and WEST, M. (2001). Combined parameter and state estimation in simulation-based filtering. In *Sequential Monte Carlo Methods in Practice* (A. Doucet, N. de Freitas and N. J. Gordon, eds.) 197–223. Springer, New York. MR1847793

MARYAK, J. L. and CHIN, D. C. (2008). Global random optimization by simultaneous perturbation stochastic approximation. *IEEE Trans. Automat. Control* **53** 780–783. MR2401029

MORTON, A. and FINKENSTÄDT, B. F. (2005). Discrete time modelling of disease incidence time series by using Markov chain Monte Carlo methods. *J. Roy. Statist. Soc. Ser. C* **54** 575–594. MR2137255

NEWMAN, K. B., FERNÁNDEZ, C., THOMAS, L. and BUCKLAND, S. T. (2009). Monte Carlo inference for state-space models of wild animal populations. *Biometrics* **65** 572–583. MR2751482

QIAN, Z. and SHAPIRO, A. (2006). Simulation-based approach to estimation of latent variable models. *Comput. Statist. Data Anal.* **51** 1243–1259. MR2297520

REUMAN, D. C., DESHARNAIS, R. A., COSTANTINO, R. F., AHMAD, O. S. and COHEN, J. E. (2006). Power spectra reveal the influence of stochasticity on nonlinear population dynamics. *Proc. Natl. Acad. Sci. USA* **103** 18860–18865.

ROGERS, L. C. G. and WILLIAMS, D. (1994). *Diffusions, Markov Processes, and Martingales. Vol. 1: Foundations*, 2nd ed. Wiley, Chichester. MR1331599

SPALL, J. C. (2003). *Introduction to Stochastic Search and Optimization: Estimation, Simulation, and Control*. Wiley-Interscience, Hoboken, NJ. MR1968388

TONI, T., WELCH, D., STRELKOWA, N., IPSEN, A. and STUMPF, M. P. (2008). Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. *Journal of the Royal Society Interface* **6** 187–202.

WEST, M. and HARRISON, J. (1997). *Bayesian Forecasting and Dynamic Models*, 2nd ed. Springer, New York. MR1482232

E. L. IONIDES
DEPARTMENT OF STATISTICS
UNIVERSITY OF MICHIGAN
ANN ARBOR, MICHIGAN 48109
USA
AND
FOGARTY INTERNATIONAL CENTER
NATIONAL INSTITUTES OF HEALTH
BETHESDA, MARYLAND 20892
USA
E-MAIL: ionides@umich.edu

A.BHADRA
Y. ATCHADÉ
DEPARTMENT OF STATISTICS
UNIVERSITY OF MICHIGAN
ANN ARBOR, MICHIGAN
USA
E-MAIL: tatar@umich.edu
        yvesa@umich.edu

A. KING
DEPARTMENT OF ECOLOGY AND
    EVOLUTIONARY BIOLOGY
UNIVERSITY OF MICHIGAN
ANN ARBOR, MICHIGAN
USA
AND
FOGARTY INTERNATIONAL CENTER
NATIONAL INSTITUTES OF HEALTH
BETHESDA, MARYLAND 20892
USA
E-MAIL: kingaa@umich.edu