Inference for dynamic and latent variable models via iterated, perturbed Bayes maps

Edward L. Ionides [*], **Dao Nguyen** [*], **Yves Atchadé** [*], **Stilian Stoev** [*] and **Aaron A. King** [*]

[*]University of Michigan, Ann Arbor, Michigan

Submitted to Proceedings of the National Academy of Sciences of the United States of America

Iterated filtering algorithms are stochastic optimization procedures for latent variable models that recursively combine parameter perturbations with latent variable reconstruction. Previously, theoretical support for these algorithms has been based on the use of conditional moments of perturbed parameters to approximate derivatives of the log likelihood function. Here, a new theoretical approach is introduced based on the convergence of an iterated Bayes map. A new algorithm supported by this theory displays substantial numerical improvement on the computational challenge of inferring parameters of a partially observed Markov process.

sequential Monte Carlo | particle filter | maximum likelihood | Markov process

Abbreviations: POMP, partially observed Markov process ; MLE, maximum likelihood estimate; CLT, central limit theorem

An iterated filtering algorithm was originally proposed for maximum likelihood inference on partially observed Markov process (POMP) models by Ionides et al [1]. Variations on the original algorithm have been proposed to extend it to general latent-variable models [2] and to improve numerical performance [3, 4]. In this paper, we study a new iterated filtering algorithm which generalizes the data cloning method [5, 6] and is therefore also related to other Monte Carlo methods for likelihood-based inference [7, 8, 9]. Data cloning methodology is based on the observation that iterating a Bayes map converges to a point mass at the maximum likelihood estimate. Combining such iterations with perturbations of model parameters improves the numerical stability of data cloning and provides a foundation for stable algorithms in which the Bayes map is numerically approximated by sequential Monte Carlo computations.

We investigate convergence of a sequential Monte Carlo implementation of an iterated filtering algorithm which combines data cloning, in the sense of Lele et al [5], with the stochastic parameter perturbations used by the iterated filtering algorithm of [1]. Lindström et al [4] proposed a similar algorithm, termed fast iterated filtering, but the theoretical support for that algorithm involved unproved conjectures. We present convergence results for our algorithm, which we call IF2. Empirically, it can dramatically out-perform the previous iterated filtering algorithm of [1], which we refer to as IF1. Though IF1 and IF2 both involve recursively filtering through the data, the theoretical justification and practical implementations of these algorithms are fundamentally different. IF1 approximates the Fisher score function, whereas IF2 implements an iterated Bayes map. IF1 has been used in applications for which no other computationally feasible algorithm for statistically efficient, likelihood-based inference was known [10, 11, 12, 13, 14, 15]. The extra capabilities offered by IF2 open up further possibilities for drawing inferences about nonlinear partially observed stochastic dynamic models from time series data.

Iterated filtering algorithms implemented using basic sequential Monte Carlo techniques have the property that they do not need to evaluate the transition density of the latent Markov process. Algorithms with this property have been called plug-and-play [16, 12]. Various other plug-and-play methods for POMP models have been recently proposed [17, 18, 19, 20], due largely to the convenience of this property in scientific applications.

An algorithm and related questions

A general POMP model consists of an unobserved stochastic process $\{X(t), t \ge t_0\}$ with observations $Y_1, \ldots, Y_N$ made at times $t_1, \ldots, t_N$. We suppose that $X(t)$ takes values in $\mathbb{X} \subset \mathbb{R}^{\dim(\mathbb{X})}$, Y_n takes values in $\mathbb{Y} \subset \mathbb{R}^{\dim(\mathbb{Y})}$, and there is an unknown parameter θ taking values in $\Theta \subset \mathbb{R}^{\dim(\Theta)}$. We adopt notation $y_{m:n} = y_m, y_{m+1}, \ldots, y_n$ for integers $m \le n$, so we write the collection of observations as $Y_{1:N}$. Writing $X_n = X(t_n)$, the joint density of $X_{0:N}$ and $Y_{1:N}$ is assumed to exist, and the Markovian property of $X_{0:N}$ together with the conditional independence of the observation process means that this joint density can be written as

$$f_{X_{0:N},Y_{1:N}}(x_{0:N}, y_{1:N}\,;\theta) = f_{X_0}(x_0\,;\theta)\prod_{n=1}^{N} f_{X_n|X_{n-1}}(x_n \,|\, x_{n-1}\,;\theta) f_{Y_n|X_n}(y_n \,|\, x_n\,;\theta).$$

The data consist of a sequence of observations, $y^*_{1:N}$. We write $f_{Y_{1:N}}(y_{1:N}\,;\theta)$ for the marginal density of $Y_{1:N}$, and the likelihood function is defined to be $\ell(\theta) = f_{Y_{1:N}}(y^*_{1:N}\,;\theta)$. We look for a maximum likelihood estimate (MLE), i.e., a value $\hat\theta$ maximizing $\ell(\theta)$. The IF2 algorithm defined below provides

Significance

Many scientific challenges involve the study of stochastic dynamic systems about which only noisy or incomplete measurements are available. Inference for partially observed Markov process models provides a framework for formulating and answering questions about these systems. Except when the system is small, or approximately linear and Gaussian, state-of-the-art statistical methods are required to make efficient use of available data. Evaluation of the likelihood for a partially observed Markov process model can be formulated as a filtering problem. Iterated filtering algorithms carry out repeated Monte Carlo filtering operations to maximize the likelihood. We develop a new theoretical framework for iterated filtering and construct a new algorithm that dramatically out-performs previous approaches on a challenging inference problem in disease ecology.

Reserved for Publication Footnotes

Algorithm IF2. Iterated filtering

input:
Simulator for $f_{X_0}(x_0\,;\theta)$
Simulator for $f_{X_n|X_{n-1}}(x_n\,|\,x_{n-1}\,;\theta)$, n in $1:N$
Evaluator for $f_{Y_n|X_n}(y_n\,|\,x_n\,;\theta)$, n in $1:N$
Data, $y^*_{1:N}$
Number of iterations, M
Number of particles, J
Initial parameter swarm, $\{\Theta^0_j,\ j \text{ in } 1:J\}$
Perturbation density, $h_n(\theta\,|\,\varphi\,;\sigma)$, n in $1:N$
Perturbation sequence, $\sigma_{1:M}$

output: Final parameter swarm, $\{\Theta^M_j,\ j \text{ in } 1:J\}$

For m in $1:M$
  $\Theta^{F,m}_{0,j} \sim h_0(\theta\,|\,\Theta^{m-1}_j\,;\sigma_m)$ for j in $1:J$
  $X^{F,m}_{0,j} \sim f_{X_0}(x_0;\Theta^{F,m}_{0,j})$ for j in $1:J$
  For n in $1:N$
    $\Theta^{P,m}_{n,j} \sim h_n(\theta\,|\,\Theta^{F,m}_{n-1,j},\sigma_m)$ for j in $1:J$
    $X^{P,m}_{n,j} \sim f_{X_n|X_{n-1}}(x_n\,|\,X^{F,m}_{n-1,j};\Theta^{P,m}_j)$ for j in $1:J$
    $w^m_{n,j} = f_{Y_n|X_n}(y^*_n\,|\,X^{P,m}_{n,j};\Theta^{P,m}_{n,j})$ for j in $1:J$
    Draw $k_{1:J}$ with $\mathbb{P}(k_j=i) = w^m_{n,i}\Big/\sum_{u=1}^J w^m_{n,u}$
    $\Theta^{F,m}_{n,j} = \Theta^{P,m}_{n,k_j}$ and $X^{F,m}_{n,j} = X^{P,m}_{n,k_j}$ for j in $1:J$
  End For
  Set $\Theta^m_j = \Theta^{F,m}_{N,j}$ for j in $1:J$
End For

a plug-and-play Monte Carlo approach to obtaining $\hat\theta$. A simplification of IF2 arises when $N=1$, in which case iterated filtering is called iterated importance sampling [2] (SI, Sec. S2). Algorithms similar to IF2 with a single iteration ($M=1$) have been proposed in the context of Bayesian inference [21, 22] (SI, Sec. S6). When $M=1$ and $h_n(\theta\,|\,\varphi\,;\sigma)$ degenerates to a point mass at φ, the IF2 algorithm becomes a standard particle filter [23, 24]. In the IF2 algorithm description, $\Theta^{F,m}_{n,j}$ and $X^{F,m}_{n,j}$ are the jth particles at time n in the Monte Carlo representation of the mth iteration of a filtering recursion. The filtering recursion is coupled with a prediction recursion, represented by $\Theta^{P,m}_{n,j}$ and $X^{P,m}_{n,j}$. The resampling indices $k_{1:J}$ in IF2 are taken to be a multinomial draw for our theoretical analysis, but systematic resampling is preferable in practice [23]. A natural choice of $h_n(\theta\,|\,\varphi\,;\sigma)$ is a multivariate normal density with mean φ and variance $\sigma^2\Sigma$ for some covariance matrix Σ, but in general h_n could be any conditional density parameterized by σ. Combining the perturbations over all the time points, we define

$$h(\theta_{0:N}\,|\,\varphi\,;\sigma) = h_0(\theta_0\,|\,\varphi\,;\sigma)\prod_{n=1}^N h_n(\theta_n\,|\,\theta_{n-1}\,;\sigma).$$

We define an extended likelihood function on Θ^{N+1} by

$$\breve\ell(\theta_{0:N}) \;=\; \int\!\!\cdots\!\!\int d x_0\ldots d x_N\Big\{f_{X_0}(x_o\,;\theta_0)\times \prod_{n=1}^N f_{X_n|X_{n-1}}(x_n\,|\,x_{n-1}\,;\theta_n) f_{Y_n|X_n}(y^*_n\,|\,x_n\,;\theta_n)\Big\}.$$

Each iteration of IF2 is a Monte Carlo approximation to a map

$$T_\sigma f(\theta_N) = \frac{\int \breve\ell(\theta_{0:N})h(\theta_{0:N}|\varphi\,;\sigma)f(\varphi)\,d\varphi\,d\theta_{0:N-1}}{\int \breve\ell(\theta_{0:N})h(\theta_{0:N}|\varphi\,;\sigma)f(\varphi)\,d\varphi\,d\theta_{0:N}}, \qquad [1]$$

with f and $T_\sigma f$ approximating the initial and final density of the parameter swarm. For our theoretical analysis, we consider the case when the standard deviation of the parameter perturbations is held fixed at $\sigma_m=\sigma>0$ for $m=1,\ldots,M$. In this case, IF2 is a Monte Carlo approximation to $T^M_\sigma f(\theta)$. We call the fixed σ version of IF2 *homogeneous* iterated filtering since each iteration implements the same map. For any fixed σ, one cannot expect a procedure such as IF2 to converge to a point mass at the MLE. However, for fixed but small σ, we show that IF2 does approximately maximize the likelihood, with an error that shrinks to zero in a limit as $\sigma\to 0$ and $M\to\infty$. An immediate motivation for studying the homogeneous case is simplicity; it turns out that even with this simplifying assumption the theoretical analysis is not entirely straightforward. Moreover, the homogeneous analysis gives at least as much insight as an asymptotic analysis into the practical properties of IF2, when σ_m decreases down to some positive level $\sigma>0$ but never completes the asymptotic limit $\sigma_m\to 0$. Iterated filtering algorithms have been primarily developed in the context of making progress on complex models for which successfully achieving and validating global likelihood optimization is challenging. In such situations, it is advisable to run multiple searches and continue each search up to the limits of available computation [25]. If no single search can reliably locate the global maximum, a theory assuring convergence to a neighborhood of the maximum is as relevant as a theory assuring convergence to the maximum itself in a practically unattainable limit.

The map T_σ can be expressed as a composition of a parameter perturbation with a Bayes map that multiplies by the likelihood and renormalizes. Iteration of the Bayes map alone has a central limit theorem (CLT) [5] which forms the theoretical basis for the data cloning methodology of [5, 6]. Repetitions of the parameter perturbation may also be expected to follow a CLT. One might therefore imagine that the composition of these two operations also has a Gaussian limit. This is not generally true, since the rescaling involved in the perturbation CLT prevents the Bayes map CLT from applying (SI, Sec. S4). Our agenda is to seek conditions guaranteeing the following:

(A1) For every fixed $\sigma>0$, $\lim_{m\to\infty} T^m_\sigma f = f_\sigma$ exists.
(A2) When J and M become large, IF2 numerically approximates f_σ.
(A3) As the noise intensity becomes small, $\lim_{\sigma\to 0} f_\sigma$ approaches a point mass at the MLE, if it exists.

Stability of filtering problems and uniform convergence of sequential Monte Carlo numerical approximations are closely related, and so A1 and A2 are studied together in Theorem 1. Each iteration of IF2 involves standard sequential Monte Carlo filtering techniques applied to an extended model where latent variable space is augmented to include a time-varying parameter. Indeed, all M iterations together can be represented as a filtering problem for this extended POMP model on M replications of the data. The proof of Theorem 1 therefore leans on existing results. The novel issue of A3 is then addressed in Theorem 2.

Convergence of IF2

First, we set up some notation. Let $\{\breve\Theta^m_{0:N}, m=1,2,\ldots\}$ be a Markov chain taking values in Θ^{N+1} such that $\breve\Theta^1_{0:N}$ has

density $\int h(\theta_{0:N} \,|\, \varphi\,; \sigma) f(\varphi)\, d\varphi$, and $\breve{\Theta}^m_{0:N}$ has conditional density $h(\theta_{0:N} \,|\, \varphi_N\,; \sigma)$ given $\breve{\Theta}^{m-1}_{0:N} = \varphi_{0:N}$ for $m \geq 2$. Suppose that $\{\breve{\Theta}^m_{0:N}, m \geq 1\}$ is constructed on the canonical probability space $\Omega = \{(\theta^1_{0:N}, \theta^2_{0:N}, \dots)\}$ with $\theta^m_{0:N} = \breve{\Theta}^m_{0:N}(\vartheta)$ for $\vartheta = (\theta^1_{0:N}, \theta^2_{0:N}, \dots) \in \Omega$. Let $\{\mathcal{F}_m\}$ be the corresponding Borel filtration. To consider a time-rescaled limit of $\{\breve{\Theta}^m_{0:N}, m = 1, 2, \dots\}$ as $\sigma \to 0$, let $\{W_\sigma(t), t \geq 0\}$ be a continuous-time, right-continuous, piecewise constant process defined at its points of discontinuity by $W_\sigma(k\sigma^2) = \breve{\Theta}^{k+1}_N$ when k is a nonnegative integer. Let $\{\breve{Z}^m_{0:N}, m = 1, 2, \dots\}$ be the filtered process defined such that, for any event $E \in \mathcal{F}_M$,

$$\mathbb{P}_{\breve{Z}}(E) = \frac{\mathbb{E}_{\breve{\Theta}}[\breve{\ell}_{1:M} I_E]}{\mathbb{E}_{\breve{\Theta}}[\breve{\ell}_{1:M}]}, \qquad \textbf{[2]}$$

where I_E is the indicator function for event E and

$$\breve{\ell}_{1:M}(\vartheta) = \prod_{m=1}^{M} \breve{\ell}(\theta^m_{0:N}).$$

In **[2]**, $\mathbb{P}_{\breve{Z}}(E)$ denotes probability under the law of $\{\breve{Z}^m_n\}$, and $\mathbb{E}_{\breve{\Theta}}$ denotes expectation under the law of $\{\breve{\Theta}^m_n\}$. The process $\{\breve{Z}^m_n\}$ is constructed so that $\breve{Z}^m_N$ has density $T^m f$. We make the following assumptions.

(B1) $\{W_\sigma(t), 0 \leq t \leq 1\}$ converges weakly as $\sigma \to 0$ to a diffusion $\{W(t), 0 \leq t \leq 1\}$, in the space of right-continuous functions with left limits equipped with the uniform convergence topology. For any open set $A \subset \Theta$ with positive Lebesgue measure and $\epsilon > 0$, there is a $\delta(A, \epsilon) > 0$ such that $\mathbb{P}\big[W(t) \in A \text{ for all } \epsilon \leq t \leq 1 \,|\, W(0)\big] > \delta$.

(B2) For some $t_0(\sigma)$ and $\sigma_0 > 0$, $W_\sigma(t)$ has a positive density on Θ, uniformly over the distribution of $W(0)$ for all $t > t_0$ and $\sigma < \sigma_0$.

(B3) $\ell(\theta)$ is continuous in a neighborhood $\{\theta : \ell(\theta) > \lambda_1\}$ for some $\lambda_1 < \sup_\varphi \ell(\varphi)$.

(B4) There is an $\epsilon > 0$ with $\epsilon^{-1} > f_{Y_n|X_n}(y^*_n \,|\, x_n, \theta) > \epsilon$ for all $1 \leq n \leq N$, $x_n \in \mathbb{X}$ and $\theta \in \Theta$.

(B5) There is a C_1 such that $h_n(\theta \,|\, \varphi\,; \sigma) = 0$ when $|\theta - \varphi| > C_1\sigma$, for all σ.

(B6) There is a C_2 such that $\sup_{1 \leq n \leq N} |\theta_n - \theta_{n-1}| < C_1\,\sigma$ implies $|\breve{\ell}(\theta_{0:N}) - \ell(\theta_N)| < C_2\,\sigma$, for all σ and all n.

Conditions B1 and B2 hold when $h_n(\theta \,|\, \varphi\,; \sigma)$ corresponds to a reflected Gaussian random walk and $\{W(t)\}$ is a reflected Brownian motion (SI, Sec. S8). More generally, $h_n(\theta \,|\, \varphi\,; \sigma)$ is a location-scale family with mean φ away from a boundary, then $\{W(t)\}$ will behave like Brownian motion in the interior of Θ. B4 follows if $\mathbb{X}$ is compact and $f_{Y_n|X_n}(y^*_n \,|\, x_n\,; \theta)$ is positive and continuous as a function of θ and x_n. B5 can be guaranteed by construction. B3 and B6 are undemanding regularity conditions on the likelihood and extended likelihood. A formalization of A1 and A2 can now be stated as follows.

Theorem 1. *Let T_σ be the map of* **[1]** *and Suppose B2 and B4. There is a unique probability density f_σ such that for any probability density f on Θ,*

$$\lim_{m \to \infty} \|T^m_\sigma f - f_\sigma\|_1 = 0, \qquad \textbf{[3]}$$

where $\|f\|_1$ is the L^1 norm of f. Let $\{\Theta^M_j,\ j = 1, \dots, J\}$ be the output of IF2, with $\sigma_m = \sigma > 0$. There is a finite constant $C > 0$ such that, for any function $\phi : \Theta \to \mathbb{R}$ and all M,

$$\mathbb{E}\left\{\Big|\frac{1}{J}\sum_{j=1}^{J} \phi(\Theta^M_j) - \int \phi(\theta) f_\sigma(\theta)\, d\theta\Big|\right\} \leq \frac{C\, \sup_\theta |\phi(\theta)|}{\sqrt{J}}. \qquad \textbf{[4]}$$

Proof. B2 and B4 imply that T^k_σ is mixing, in the sense of [26], for all sufficiently large k. The results of [26] are based on the contractive properties of mixing maps in the Hilbert projective metric. Although [26] stated their results in the case where T itself is mixing, the required geometric contraction in the Hilbert metric holds as long as T^k is mixing for all $K \leq k \leq 2K - 1$ for some $K \geq 1$ [27, Theorem 2.5.1]. Corollary 4.2 of [26] implies **[3]**, noting the equivalence of the Hilbert projective metric and the total variation norm shown in their Lemma 3.4. Then, Corollary 5.12 of [26] implies **[4]**, completing the proof of Theorem 1. A longer version of this proof is given in the supplement (Sec. S9). □

Results similar to Theorem 1 can be obtained using Dobrushin contraction techniques [28]. Results appropriate for non-compact spaces can be obtained using drift conditions on a potential function [29]. Now we move on to our formalization of A3:

Theorem 2. *Assume B1–B6. For $\lambda_2 < \sup_\varphi \ell(\varphi)$, $\lim_{\sigma \to 0} \int f_\sigma(\theta) 1_{\{\ell(\theta) < \lambda_2\}}\, d\theta = 0$.*

Proof. Let $\lambda_0 = \sup_\varphi \ell(\varphi)$ and $\lambda_3 = \inf_\varphi \ell(\varphi)$. From B4, $\infty > \lambda_0 > \lambda_3 > 0$. For positive constants ϵ_1, ϵ_2, η_1, η_2 and $\lambda_1 < \lambda_0$, define

$$\begin{aligned} e_1 &= (1 - \epsilon_1)\log(\lambda_0 + \epsilon_2) + \epsilon_1 \log(\lambda_2 + \epsilon_2), \\ e_2 &= (1 - \eta_1)\log(\lambda_1 - \eta_2) + \eta_1 \log(\lambda_3 - \eta_2). \end{aligned}$$

We can pick ϵ_1, ϵ_2, η_1, η_2 and λ_1 so that $e_1 < e_2$. Suppose that $\{\breve{\Theta}^m_n\}$ is initialized with the stationary distribution $f = f_\sigma$ identified in Theorem 1. Now, set M to be the greatest integer less than $1/\sigma^2$, and let F_1 be the event that $\{\Theta^m_N, m = 1, \dots, M\}$ spends at least a fraction of time ϵ_1 in $\{\theta : \ell(\theta) < \lambda_2\}$. Formally,

$$F_1 = \Big\{\vartheta \in \Omega : \frac{1}{M}\sum_{m=1}^{M} 1_{\{\ell(\theta^m_N) < \lambda_2\}} > \epsilon_1\Big\}.$$

We wish to show that $\mathbb{P}_{\breve{Z}}[F_1]$ is small for σ small. Let F_2 be the set of sample paths that spend at least a fraction of time $(1 - \eta_1)$ up to time M in $\{\theta : \ell(\theta) > \lambda_1\}$, i.e.,

$$F_2 = \Big\{\vartheta \in \Omega : \frac{1}{M}\sum_{m=1}^{M} 1_{\{\ell(\theta^m_N) > \lambda_1\}} > (1 - \eta_1)\Big\}.$$

Then, we calculate

$$\begin{aligned} \mathbb{P}_{\breve{Z}}[F_1] &= \frac{\mathbb{E}_{\breve{\Theta}}\big[\breve{\ell}_{1:M} 1_{F_1}\big]}{\mathbb{E}_{\breve{\Theta}}\big[\breve{\ell}_{1:M}\big]} \\ &\leq \frac{\mathbb{E}_{\breve{\Theta}}\big[\breve{\ell}_{1:M} 1_{F_1}\big]}{\mathbb{E}_{\breve{\Theta}}\big[\breve{\ell}_{1:M} 1_{F_2}\big]} \\ &\leq \frac{\mathbb{E}_{\breve{\Theta}}\Big[\prod_{m=1}^{M}\big\{\ell(\theta^m_N) + C_2\sigma\big\} 1_{F_1}\Big]}{\mathbb{E}_{\breve{\Theta}}\Big[\prod_{m=1}^{M}\big\{\ell(\theta^m_N) - C_2\sigma\big\} 1_{F_2}\Big]} \qquad \textbf{[5]} \\ &\leq \frac{\mathbb{E}_{\breve{\Theta}}\big[\exp\{Me_1\} 1_{F_1}\big]}{\mathbb{E}_{\breve{\Theta}}\big[\exp\{Me_2\} 1_{F_2}\big]} \qquad \textbf{[6]} \\ &= \exp\big\{(e_1 - e_2)M\big\}\frac{\mathbb{P}_{\breve{\Theta}}[F_1]}{\mathbb{P}_{\breve{\Theta}}[F_2]}. \qquad \textbf{[7]} \end{aligned}$$

We used B5 and B6 to arrive at **[5]**, then to get to **[6]** we have taken σ small enough that $C_2\sigma < \epsilon_2$ and $C_2\sigma < \eta_2$. From B3, $\{\theta : \ell(\theta) > \lambda_1\}$ is an open set, and B1 therefore ensures each of the probabilities $\mathbb{P}_{\Theta_{1:M}}[F_1]$ and $\mathbb{P}_{\Theta_{1:M}}[F_2]$ in **[7]** tends

to a positive limit as $\sigma \to 0$ given by the probability under the limiting distribution $\{W(t)\}$ (SI, Lemma S1). The term $\exp\{(e_1-e_2)M\}$ tends to zero as $\sigma \to 0$ since, by construction, $M \to \infty$ and $e_1 < e_2$. Setting $L = \{\theta : \ell(\theta) \le \lambda_2\}$, and noting that $\{\check{Z}_N^m, m = 1, 2, \dots\}$ is constructed to have stationary marginal density f_σ, we have

$$\begin{aligned}\int_L f_\sigma(\theta)\,d\theta &= \frac{1}{M}\sum_{m=1}^{M}\Big\{\mathbb{P}_{\check{Z}}\big[\check{Z}_N^m \in L \,|\, F_1\big]\,\mathbb{P}_{\check{Z}}[F_1] + \\ &\qquad \mathbb{P}_{\check{Z}}\big[\check{Z}_N^m \in L \,|\, F_1^c\big]\,\mathbb{P}_{\check{Z}}[F_1^c]\Big\}, \\ &\le \epsilon_1 + \mathbb{P}_{\check{Z}}[F_1],\end{aligned}$$

which can be made arbitrarily small by picking ϵ_1 small and σ small, completing the proof. □

Demonstration of IF2 with nonconvex superlevel sets

Theorems 1 and 2 do not involve any Taylor series expansions, which are basic in the justification of IF1 [2]. This might suggest that IF2 can be effective on likelihood functions without good low-order polynomial approximations. In practice, this can be seen by comparing IF2 with IF1 on a simple two-dimensional toy example ($\dim(\Theta) = \dim(\mathbb{X}) = \dim(\mathbb{Y}) = 2$) in which the superlevel sets $\{\theta : \ell(\theta) > \lambda\}$ are connected but not convex. We also compare with particle Markov chain Monte Carlo (PMCMC) implemented as the PMMH algorithm of [17]. The justification of PMCMC also does not depend on Taylor series expansions, but PMCMC is computationally expensive compared to iterated filtering [30]. Our toy example has a constant and non-random latent process, $X_n = \big(\exp\{\theta_1\}, \theta_2 \exp\{\theta_1\}\big)$ for $n = 1, \dots, N$. The known measurement model is

$$f_{Y_n|X_n}(y \,|\, x\,; \theta) \sim \text{Normal}\left[x, \begin{pmatrix} 100 & 0 \\ 0 & 1 \end{pmatrix}\right],$$

This example was designed so that a nonlinear combination of the parameters is well identified whereas each parameter is marginally weakly identified. For the truth, we took $\theta = (1, 1)$. We supposed that θ_1 is suspected to fall in the interval $[-2, 2]$ and θ_2 is expected in $[0, 10]$. We used a uniform distribution on this rectangle to specify the prior for PMCMC and to generate random starting points for all the algorithms. We set $N = 100$ observations, and we used a Monte Carlo sample size of $J = 100$ particles. For IF1 and IF2, we employed $M = 100$ filtering iterations, with initial random walk standard deviation 0.1 decreasing geometrically down to 0.01. For PMCMC, we used 10^4 filtering iterations with random walk standard deviation 0.1, awarding PMCMC 100 times the computational resources offered to IF1 and IF2. Independent, normally distributed parameter perturbations were used for IF1, IF2 and PMCMC. The random walk standard deviation for PMCMC is not immediately comparable to that for IF1 and IF2, since the latter add the noise at each observation time whereas the former adds it only between filtering iterations. All three methods could have their parameters fine-tuned, or be modified in other ways to take advantage of the structure of this particular problem. However, this example demonstrates a feature that makes tuning algorithms tricky: the nonlinear ridge along contours of constant $\theta_2 \exp(\theta_1)$ becomes increasingly steep as θ_1 increases, so no single global estimate of the second derivative of the likelihood is appropriate. Reparameterization can linearize the ridge in this toy example, but in practical problems with much larger parameter spaces one does not always know how to find appropriate reparameterizations, and a single reparameterization may not be appropriate throughout the parameter space.

Fig. 1 compares the the performance of the three methods, based on 30 Monte Carlo replications. These replications investigate the likelihood and posterior distribution for a single draw from our toy model, since our interest is in the Monte Carlo behavior for a given dataset. For this simulated dataset, the MLE is $\theta = (1.20, 0.81)$, shown as a green triangle in Fig. 1, panels A, B and C. In this toy example, the posterior distribution can also be computed directly by numerical integration. In Fig. 1A, we see that IF1 performs poorly on this challenge. None of the 30 replications approach the MLE. The linear combination of perturbed parameters involved in the IF1 update formula can all too easily knock the search off a nonlinear ridge. Fig. 1B shows that IF2 performs well on this test, with almost all the Monte Carlo replications clustering in the region of highest likelihood. Fig. 1C shows the end points of the PMCMC replications, which are nicely spread around the region of high posterior probability. However, Fig. 1D shows that mixing of the PMCMC Markov chains was problematic.

Application to a cholera model

Highly nonlinear, partially observed, stochastic dynamic systems are ubiquitous in the study of biological processes. The physical scale of the systems vary widely from molecular biology [31] to population ecology and epidemiology [32], but POMP models arise naturally at all scales. In the face of biological complexity, it is necessary to determine which scientific aspects of a system are critical for the investigation. Giving consideration to a range of potential mechanisms, and their interactions, may require working with highly parameterized models. Limitations in the available data may result in some combinations of parameters being weakly identifiable. Despite this, other combinations of parameters may be adequately identifiable and give rise to some interesting statistical inferences. To demonstrate the capabilities of IF2 for such analyses, we fit a model for cholera epidemics in historic Bengal developed by King et al [10]. The model, the data, and the implementations of IF1 and IF2 used below are all contained in the open source R package `pomp` [33]. The code generating the results in this article is provided as supplementary data.

Cholera is a diarrheal disease caused by the bacterial pathogen *Vibrio cholerae*. Without appropriate medical treatment, severe infections can rapidly result in death by dehydration. Many questions regarding cholera transmission remain unresolved: what is the epidemiological role of free-living environmental vibrio? how important are mild and asymptomatic infections for the transmission dynamics? how long does protective immunity last following infection? The model we consider splits up the study population of $P(t)$ individuals into those who are susceptible, $S(t)$, infected, $I(t)$, and recovered, $R(t)$. $P(t)$ is assumed known from census data. To allow flexibility in representing immunity, $R(t)$ is subdivided into $R_1(t), \dots, R_k(t)$, where we take $k = 3$. Cumulative cholera mortality in each month is tracked with a variable $M(t)$ that resets to zero at the beginning of each observation period. The state process, $\{X(t) = (S(t), I(t), R_1(t), \dots, R_k(t), M(t)), t \ge t_0\}$ follows a stochastic differential equation,

$$\begin{aligned} dS &= \big\{k\epsilon R_k + \delta(S-H) - \lambda(t)S\big\}dt dP - (\sigma SI/P)\,dB, \\ dI &= \big\{\lambda(t)S - (m+\delta+\gamma)I\big\}dt + (\sigma SI/P)\,dB, \\ dR_1 &= \big\{\gamma I - (k\epsilon+\delta)R_1\big\}dt, \\ &\vdots \\ dR_k &= \big\{k\epsilon R_{k-1} - (k\epsilon+\delta)R_k\big\}dt, \end{aligned}$$

driven by a Brownian motion $\{B(t)\}$. Nonlinearity arises through the force of infection, $\lambda(t)$, specified as

$$\lambda(t) = \bar{\beta}\exp\Big\{\beta_{\text{trend}}(t-t_0)+\textstyle\sum_{j=1}^{N_s}\beta_j s_j(t)\Big\}(I/P)+ \bar{\omega}\exp\Big\{\textstyle\sum_{j=1}^{N_s}\omega_j s_j(t)\Big\},$$

where $\{s_j(t), j = 1,\dots,N_s\}$ is a periodic cubic B-spline basis; $\{\beta_j, j = 1,\dots,N_s\}$ model seasonality of transmission; $\{\omega_j, j = 1,\dots,N_s\}$ model seasonality of the environmental reservoir; $\bar{\omega}$ and $\bar{\beta}$ are scaling constants set to $\bar{\omega} = \bar{\beta} = 1\text{yr}^{-1}$, and we set $N_s = 6$. The data, consisting of monthly counts of cholera mortality, are modeled via $Y_n \sim \text{Normal}(M_n, \tau^2 M_n^2)$ for $M_n = \int_{t_{n-1}}^{t_n} m\, I(s)\, ds$.

The inference goal used to assess IF1 and IF2 is to find high-likelihood parameter values starting from randomly drawn starting values in a large hyper-rectangle (SI, Table S-1). A single search cannot necessarily be expected to reliably obtain the maximum of the likelihood, due to multi-modality, weak identifiability, and considerable Monte Carlo error in evaluating the likelihood. Multiple starts and restarts may be needed both for effective optimization and for assessing the evidence to validate effective optimization. However, optimization progress made on an initial search provides a concrete criterion to compare methodologies. Since IF1 and IF2 have essentially the same computational cost, for a given Monte Carlo sample size and number of iterations, shared fixed values of these algorithmic parameters provide an appropriate comparison.

Fig. 2 compares results for 100 searches with $J = 10^4$ particles and $M = 100$ iterations of the search. An initial Gaussian random walk standard deviation of 0.1 geometrically decreasing down to a final value of 0.01 was used for all parameters except S_0, I_0, $R_{1,0}$, $R_{2,0}$ and $R_{3,0}$. For those initial value parameters, the random walk standard deviation decreased geometrically from 0.2 down to 0.02, but these perturbations were applied only at time t_0. Since some starting points may lead both IF1 and IF2 to fail to approach the global maximum, Fig. 2 plots the likelihoods of parameter vectors output by IF1 and IF2 for each starting point. Fig. 2 shows that, on this problem, IF2 is considerably more effective than IF1. This maximization was considered challenging for IF1, and [10] required multiple restarts and refinements of the optimization procedure. Our implementation of PMCMC failed to converge on this inference problem (SI, Sec. S5), and we are not aware of any previous successful PMCMC solution for a comparable situation. For IF2, however, this situation appears routine. Some Monte Carlo replication is needed because searches occasionally fail to approach the global optimum, but replication is always appropriate for Monte Carlo optimization procedures.

A fair numerical comparison of methods is difficult. For example, it could hypothetically be the case that the algorithmic settings used here favor IF2. However, the settings used are those that were developed for IF1 by [10] and reflect considerable amounts of trial and error with that method. Likelihood-based inference for general partially observed nonlinear stochastic dynamic models was considered computationally unfeasible prior to the introduction of IF1, even in situations considerably simpler than the one investigated in this section [19]. We have shown that IF2 offers a substantial improvement on IF1, by demonstrating that it functions effectively on a problem at the limit of the capabilities of IF1.

Discussion

Theorems 1 and 2 assert convergence without giving insights into the rate of convergence. In the particular case of a quadratic log likelihood function and additive Gaussian parameter perturbations, $\lim_{M\to\infty} T_\sigma^M f$ is Gaussian, and explicit calculations are available (SI, Sec. S3). If $\log \ell(\theta)$ is close to quadratic and the parameter perturbation is close to additive Gaussian noise, then $\lim_{M\to\infty} T_\sigma^M f$ exists and is close to the limit for the approximating Gaussian system (SI, Sec. S3). These Gaussian and near-Gaussian situations also demonstrate that the compactness conditions for Theorem 2 are not always necessary. In the case $N = 1$, IF2 applies to the more general class of latent variable models. The latent variable model, extended to include a parameter vector that varies over iterations, nevertheless has the formal structure of a POMP in the context of the IF2 algorithm. Some simplifications arise when $N = 1$ (SI, Secs. S2, S3 and S4) but the proofs of Theorems 1 and 2 do not greatly change.

A variation on iterated filtering, making white noise perturbations to the parameter rather than random walk perturbations, has favorable asymptotic properties [3]. However, practical algorithms based on this theoretical insight have not yet been published. Our experience suggests that white noise perturbations can be effective in a neighborhood of the MLE, but fail to match the performance of IF2 for global optimization problems in complex models.

The main theoretical innovation of this paper is Theorem 2, which does not depend on the specific sequential Monte Carlo filter used in IF2. One could, for example, modify IF2 to use an ensemble Kalman filter [20, 34] or an unscented Kalman filter [35]. Or, one could take advantage of variations of sequential Monte Carlo that may improve the numerical performance [36]. However, basic sequential Monte Carlo is a general and widely used nonlinear filtering technique that provides a simple yet theoretically supported foundation for the IF2 algorithm. The numerical stability of sequential Monte Carlo for the extended POMP model constructed by IF2 is comparable, in our cholera example, to the model with fixed parameters (SI, Sec. S7).

ACKNOWLEDGMENTS. Funding was provided by National Science Foundation grants DMS-1308919 and DMS-1106695, National Institutes of Health grants 1-R01-AI101155, 1-U54-GM111274 and 1-U01-GM110712, and the RAPIDD program of DHS and NIH-FIC. We acknowledge constructive comments by two anonymous referees, the editor, and Joon Ha Park.

1. Ionides, E. L, Bretó, C, & King, A. A. (2006) Inference for nonlinear dynamical systems. Proc. Natl. Acad. Sci. USA 103, 18438–18443.
2. Ionides, E. L, Bhadra, A, Atchadé, Y, & King, A. A. (2011) Iterated filtering. Ann. Stat. 39, 1776–1802.
3. Doucet, A, Jacob, P. E, & Rubenthaler, S. (2013) Derivative-free estimation of the score vector and observed information matrix with application to state-space models. Arxiv, http://arxiv.org/abs/1304.5768.
4. Lindström, E, Ionides, E. L, Frydendall, J, & Madsen, H. (2012) Efficient iterated filtering. 16th IFAC Symposium on System Identification 16, 1785–1790.
5. Lele, S. R, Dennis, B, & Lutscher, F. (2007) Data cloning: easy maximum likelihood estimation for complex ecological models using Bayesian Markov chain Monte Carlo methods. Ecology Letters 10, 551–563.
6. Lele, S. R, Nadeem, K, & Schmuland, B. (2010) Estimability and likelihood inference for generalized linear mixed models using data cloning. J. Am. Stat. Assoc. 105, 1617–1625.
7. Doucet, A, Godsill, S. J, & Robert, C. P. (2002) Marginal maximum a posteriori estimation using Markov chain Monte Carlo. Statistics and Computing 12, 77–84.
8. Gaetan, C & Yao, J.-F. (2003) A multiple-imputation Metropolis version of the EM algorithm. Biometrika 90, 643–654.

9. Jacquier, E, Johannes, M, & Polson, N. (2007) MCMC maximum likelihood for latent state models. J. Econometrics 137, 615–640.
10. King, A. A, Ionides, E. L, Pascual, M, & Bouma, M. J. (2008) Inapparent infections and cholera dynamics. Nature 454, 877–880.
11. Laneri, K, Bhadra, A, Ionides, E. L, Bouma, M, Yadav, R, Dhiman, R, & Pascual, M. (2010) Forcing versus feedback: Epidemic malaria and monsoon rains in NW India. PLoS Comp. Biol. 6, e1000898.
12. He, D, Ionides, E. L, & King, A. A. (2010) Plug-and-play inference for disease dynamics: Measles in large and small towns as a case study. J. R. Soc. Interface 7, 271–283.
13. Blackwood, J. C, Cummings, D. A. T, Broutin, H, Iamsirithaworn, S, & Rohani, P. (2013) Deciphering the impacts of vaccination and immunity on pertussis epidemiology in Thailand. Proc. Natl. Acad. Sci. USA 110, 9595–9600.
14. Shrestha, S, Foxman, B, Weinberger, D. M, Steiner, C, Viboud, C, & Rohani, P. (2013) Identifying the interaction between influenza and pneumococcal pneumonia using incidence data. Science Transl. Med. 5, 191ra84.
15. Blake, I. M, Martin, R, Goel, A, Khetsuriani, N, Everts, J, Wolff, C, Wassilak, S, Aylward, R. B, & Grassly, N. C. (2014) The role of older children and adults in wild poliovirus transmission. Proc. Natl. Acad. Sci. USA 111, 10604–10609.
16. Bretó, C, He, D, Ionides, E. L, & King, A. A. (2009) Time series analysis via mechanistic models. Ann. Appl. Stat. 3, 319–348.
17. Andrieu, C, Doucet, A, & Holenstein, R. (2010) Particle Markov chain Monte Carlo methods. J. R. Stat. Soc. B 72, 269–342.
18. Toni, T, Welch, D, Strelkowa, N, Ipsen, A, & Stumpf, M. P. (2009) Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. J. R. Soc. Interface 6, 187–202.
19. Wood, S. N. (2010) Statistical inference for noisy nonlinear ecological dynamic systems. Nature 466, 1102–1104.
20. Shaman, J & Karspeck, A. (2012) Forecasting seasonal outbreaks of influenza. Proc. Natl. Acad. Sci. USA 109, 20425–20430.
21. Kitagawa, G. (1998) A self-organising state-space model. J. Am. Stat. Assoc. 93, 1203–1215.
22. Liu, J & West, M. (2001) in Sequential Monte Carlo Methods in Practice, eds. Doucet, A, de Freitas, N, & Gordon, N. J. (Springer, New York), pp. 197–224.
23. Arulampalam, M. S, Maskell, S, Gordon, N, & Clapp, T. (2002) A tutorial on particle filters for online nonlinear, non-Gaussian Bayesian tracking. IEEE Trans. Sig. Proc. 50, 174 – 188.
24. Doucet, A, de Freitas, N, & Gordon, N. J, eds. (2001) Sequential Monte Carlo Methods in Practice. (Springer, New York).
25. Ingber, L. (1993) Simulated annealing: Practice versus theory. Mathematical and Computer Modelling 18, 29–57.
26. Le Gland, F & Oudjane, N. (2004) Stability and uniform approximation of nonlinear filters using the Hilbert metric and application to particle filters. Ann. Appl. Prob. 14, 144–187.
27. Eveson, S. P. (1995) Hilbert's projective metric and the spectral properties of positive linear operators. Proc. Lond. Math. Soc. 3, 411–440.
28. Del Moral, P & Doucet, A. (2004) Particle motions in absorbing medium with hard and soft obstacles. Stochastic Analysis and Applications 22, 1175–1207.
29. Whiteley, N, Kantas, N, & Jasra, A. (2012) Linear variance bounds for particle approximations of time-homogeneous Feynman–Kac formulae. Stochastic Processes and their Applications 122, 1840–1865.
30. Bhadra, A. (2010) Discussion of 'particle Markov chain Monte Carlo methods' by C. Andrieu, A. Doucet and R. Holenstein. J. R. Stat. Soc. B 72, 314–315.
31. Wilkinson, D. J. (2012) Stochastic Modelling for Systems Biology. (Chapman & Hall, Boca Raton, FL).
32. Keeling, M & Rohani, P. (2009) Modeling Infectious Diseases in Humans and Animals. (Princeton Univ. Press, Princeton, NJ).
33. King, A. A, Ionides, E. L, Bretó, C. M, Ellner, S, & Kendall, B. (2009) pomp: Statistical inference for partially observed markov processes. R package, available at `http://cran.r-project.org/web/packages/pomp`.
34. Yang, W, Karspeck, A, & Shaman, J. (2014) Comparison of filtering methods for the modeling and retrospective forecasting of influenza epidemics. PLoS Comp. Biol. 10, e1003583.
35. Julier, S & Uhlmann, J. (2004) Unscented filtering and nonlinear estimation. Proc. IEEE 92, 401–422.
36. Cappé, O, Godsill, S, & Moulines, E. (2007) An overview of existing methods and recent advances in sequential Monte Carlo. Proc. IEEE 95, 899–924.

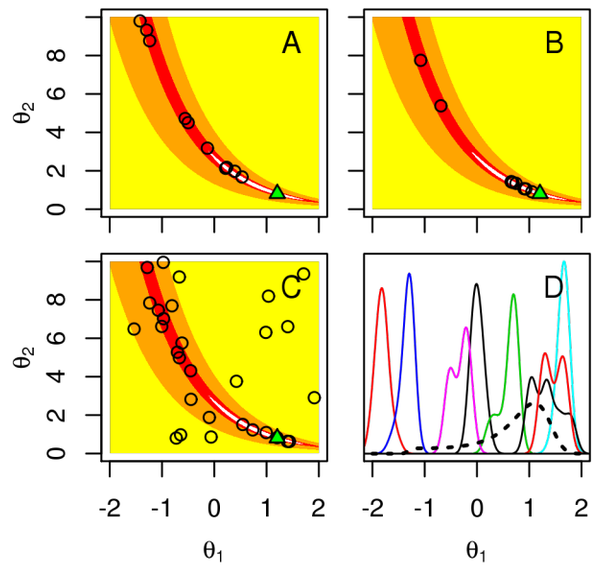

Fig. 1. Results for the simulation study of the toy example. A. IF1 point estimates from 30 replications (circles) and the MLE (green triangle). The region of parameter space with likelihood within 3 log units of the maximum (white), with 10 log units (red), within 100 log units (orange) and lower (yellow). B. IF2 point estimates from 30 replications (circles) with the same algorithmic settings as IF1. C. Final parameter value of 30 PMCMC chains (circles). D. kernel density estimates of the posterior for θ_1 for the first 8 of these 30 PMCMC chains (solid lines), with the true posterior distribution (dotted black line).

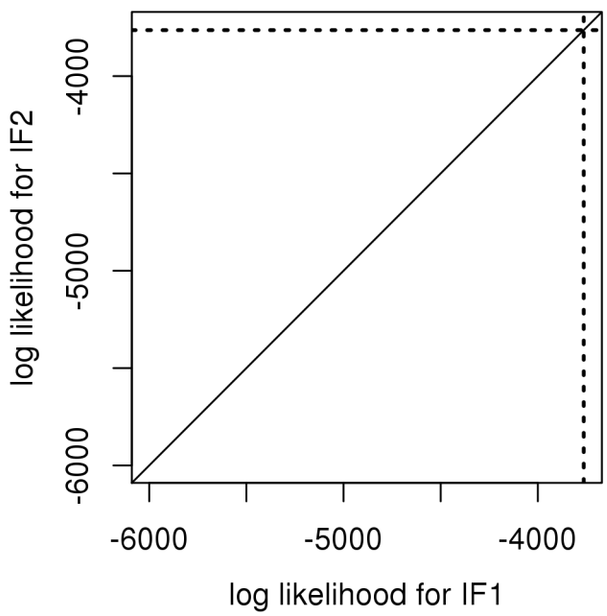

Fig. 2. Comparison of IF1 and IF2 on the cholera model. Points are the log likelihood of the parameter vector output by IF1 and IF2, both started at a uniform draw from a large hyper-rectangle (SI, Table S-1). Likelihoods were evaluated as the median of 10 particle filter replications (i.e., IF2 applied with $M = 1$ and $\sigma_1 = 0$) each with $J = 2 \times 10^4$ particles. 17 poorly performing searches are off the scale of this plot (15 due to the IF1 estimate, 2 due to the IF2 estimate). Dotted lines show the maximum log likelihood reported by [10].