Austrian Journal of Statistics
July 2023, Volume 52, 117–142.
http://www.ajs.or.at/
doi:10.17713/ajs.v52i4.1503

Accelerated Iterated Filtering

Dao Nguyen
University of Mississippi

Abstract

Simulation-based inferences have attracted much attention in recent years, as the direct computation of the likelihood function in many real-world problems is difficult or even impossible. Iterated filtering (Ionides, Bretó, and King 2006; Ionides, Bhadra, Atchadé, and King 2011) enables maximization of likelihood function via model perturbations and approximation of the gradient of log-likelihood through sequential Monte Carlo filtering. By an application of Stein's identity, Doucet, Jacob, and Rubenthaler (2013) developed a second-order approximation of the gradient of log-likelihood using sequential Monte Carlo smoothing. Based on these gradient approximations, we develop a new algorithm for maximizing the likelihood using the Nesterov accelerated gradient. We adopt the accelerated inexact gradient algorithm (Ghadimi and Lan 2016) to iterated filtering framework, relaxing the unbiased gradient approximation condition. We devise a perturbation policy for iterated filtering, allowing the new algorithm to converge at an optimal rate for both concave and non-concave log-likelihood functions. It is comparable to the recently developed Bayes map iterated filtering approach and outperforms the original iterated filtering approach.

Keywords: accelerated iterated filtering, sequential Monte Carlo, partially observed Markov process model, parameter estimation.

1. Introduction

Partially observed Markov process (POMP) models (also called state-space models) have been used as a powerful tool in modeling time series in many disciplines, including ecology, econometrics, engineering, and statistics. However, making inferences on POMP models can be challenging because of the presence of incomplete measurements and possibly weakly identifiable parameters. Standard methods for inference (e.g., maximum likelihood) with restrictive assumptions of linear Gaussian models often fail to produce satisfactory results when the assumptions are violated. Simulation-based inferences, also called plug-and-play (Bretó, He, Ionides, and King 2009; He, Ionides, and King 2010), likelihood-free (Sisson, Fan, and Tanaka 2007; Yıldırım, Singh, Dean, and Jasra 2015), or equation-free inferences (Kevrekidis, Gear, and Hummer 2004), are a class of algorithms where inferences only access the dynamic model through simulations. This class of inference algorithms is attractive because it enables routine parameter inferences in general POMP models, even in the case of intractable likelihoods. As a result, the last decade has seen a great increase in the use of simulation-based inferences where numerical approximations are primarily based on Monte Carlo sampling (Ionides *et al.*

2006; Toni, Welch, Strelkowa, Ipsen, and Stumpf 2009; Andrieu, Doucet, and Holenstein 2010; Wood 2010; Chopin, Jacob, and Papaspiliopoulos 2013; Ionides, Nguyen, Atchadé, Stoev, and King 2015). A simulation-based inference can either be described as Bayesian or frequentist, based on posterior distributions or likelihoods. Depending on how the information is exploited, Bayesian inferences can be further categorized into full information such as particle Markov chain Monte Carlo (PMCMC) (Andrieu *et al.* 2010) or partial information such as approximate Bayesian computation (ABC) (Sisson *et al.* 2007). Similarly, frequentist inferences can also be classified as full information such as iterated filtering (Ionides *et al.* 2015) or feature-based such as nonlinear forecasting (Ellner, Bailey, Bobashev, Gallant, Grenfell, and Nychka 1998), and synthetic likelihood (Wood 2010). This paper deals with full information, frequentist, and simulation-based inferences.

Iterated filtering (Ionides *et al.* 2006), the first algorithm in this category, enables maximization of likelihood function via model perturbations and approximation of the gradient of log-likelihood through sequential Monte Carlo filtering. Since then, several variations of the original algorithm have been developed. Lindström, Ionides, Frydendall, and Madsen (2012) extended iterated filtering to improve numerical performance. Doucet *et al.* (2013) expanded it to include general latent variable models and to use sequential Monte Carlo smoothing to compute both the gradient and the Hessian with very attractive theoretical properties. Ionides *et al.* (2015) generalized Lindström *et al.* (2012)'s approach and combined the idea with data cloning (Lele, Dennis, and Lutscher 2007), developed a Bayes map iterated filtering with an entirely different theoretical approach. Nguyen and Ionides (2017) revisited the approach of Doucet *et al.* (2013), using different perturbation noises and exploiting these derivatives of log-likelihood to improve on convergence rate. However, the Hessian approximation is often computationally expensive and the inaccuracies are exacerbated when combined with gradient estimations, decreasing the convergence rate, especially in the iterated filtering framework. This paper chooses an alternative, maximizing the likelihood by using accelerated gradient approaches. Thus, the proposed approach inherits a higher convergence rate from an accelerated gradient family, enhancing performance without the expensive computation of the Hessian.

The key contributions of this paper are three-fold. First, we show that the accelerated biased stochastic gradient algorithm still converges at an optimal rate for some chosen step sizes and bias sequences. In particular, we assume a rather weak condition that is often satisfied for many gradient approximations from Monte Carlo sampling. Second, we develop an efficient perturbation policy for iterated filtering, which ensures higher convergence rates than the original iterated filtering, both in concave and non-concave log-likelihood. Third, the proposed algorithm offers good numerical performance on some benchmark models.

The paper is organized as follows. In the next section, we introduce notations and recall some background of gradient approximation in an iterated filtering framework. In Section 3, we relax unbiased approximation condition, allowing the algorithm to converge with the optimal rate under the biased approximation of the gradients. Based on the biased assumption, a perturbation policy is derived for an iterated filtering framework. We validate the proposed algorithm by a toy example and a challenging inference problem of fitting malaria models to time series data in Section 4, showing substantial improvement for our methods over the original iterated filtering approach and showing that the algorithm is comparable to the recently developed Bayes map iterated filtering. We conclude in Section 5 with suggestions for the future work to be extended. The proofs and some additional illustrations are postponed to the Appendix.

2. Background of gradient approximation of iterated filtering

Let $\{X(t),\ t \in \mathbb{T}\}$ be a Markov process where $X(t)$ takes values in a measurable space $\mathcal{X}$. The time index set, $\mathbb{T} \subset \mathbb{R}$, may be an interval or a discrete set and it contains a

finite subset $t_1 < t_2 \cdots < t_N$ at which $X(t)$ is observed, along with an initial time $t_0 < t_1$. Specifically, we write $X_{0:N} = (X_0, \ldots, X_N) = (X(t_0), \ldots, X(t_N))$. Hereafter, for any generic sequence $\{X_n\}$, we shall use $X_{i:j}$ to denote $(X_i, X_{i+1}, \ldots, X_j)$. The distribution of $X_{0:N}$ is characterized by the initial density $X_0 \sim \mu(x_0;\theta)$ and the conditional density of X_n given X_{n-1}, written as $f_n(x_n|x_{n-1};\theta)$ for $1 \le n \le N$. Here, θ is an unknown parameter in $\Theta \subset \mathbb{R}^d$. The process $\{X_n\}$ is observed only through another process $\{Y_n, n = 1, \ldots, N\}$ taking values in a measurable space $\mathcal{Y}$. The observations are assumed to be conditionally independent given $X_{0:n}$, and their probability density is of the form

$$p_{Y_n|Y_{1:n-1},\ X_{0:n}}(y_n|y_{1:n-1},\ x_{0:n};\theta) = g_n(y_n|x_n;\theta),$$

for $1 \le n \le N$. We assume that $X_{0:N}$ and $Y_{1:N}$ have a joint density of $p_{X_{0:N},\ Y_{1:N}}(x_{0:N},\ y_{1:N};\theta)$ on $\mathcal{X}^{N+1} \times \mathcal{Y}^N$. The data are a sequence of observations by $y^*_{1:N} = (y^*_1, \ldots, y^*_N) \in \mathcal{Y}^N$, considered as fixed. We write the log likelihood function of the data for the POMP model as $\ell(\theta)$, given by

$$\begin{aligned} \ell(\theta) &= \log p_{Y_{1:N}}(y^*_{1:N};\theta) \\ &= \log \int \mu(x_0;\theta) \prod_{n=1}^{N} f_n(x_n|x_{n-1};\theta)\ g_n(y^*_n|x_n;\theta)\ dx_{0:N}. \end{aligned}$$

We work with the maximum likelihood estimator (MLE), which is $\hat{\theta} = \arg\max \ell(\theta)$. This MLE problem often uses the first order stochastic approximation (Kushner and Clark 1978), which involves a Monte Carlo approximation to a difference equation, $\theta_m = \theta_{m-1} + \gamma_m \nabla \ell(\theta_{m-1})$, where $\theta_0 \in \Theta$ is an arbitrary initial estimate of the parameter space and $\{\gamma_m\}_{m\ge1}$ is a sequence of step sizes with $\sum_{m\ge1} \gamma_m = \infty$ and $\sum_{m\ge1} \gamma_m^2 < \infty$. Under regularity conditions, the algorithm converges to a local maximum of $\ell(\theta)$. The term $\nabla\ell(\theta)$, also called the score function, is shorthand for the $\mathbb{R}^d$-valued vector of partial derivatives, $\nabla\ell(\theta) = \frac{\partial \ell(\theta)}{\partial \theta}$.

Sequential Monte Carlo (SMC) approaches have previously been developed to estimate the score function (Poyiadjis, Doucet, and Singh 2011; Nemeth, Fearnhead, and Mihaylova 2013; Dahlin, Lindsten, and Schön 2015). However, under a simulation-based inference setting, these approaches are not applicable, since the simulation of the derivatives of the dynamic of the model is also required. As a result, we follow the approach of Nguyen and Ionides (2017), using their Theorem 5 to approximate the score function. To be self-contained, we rewrite their assumptions and Theorem 5. Interested readers are encouraged to read their paper for full details. A POMP model is a specific latent variable model with $X = X_{0:N}$ and $Y = Y_{1:N}$. A perturbed POMP model is defined to have a similar construction to a perturbed latent variable model (Nguyen and Ionides 2017) with $\breve{X} = \breve{X}_{0:N}$, $\breve{Y} = \breve{Y}_{1:N}$ and $\breve{\Theta} = \breve{\Theta}_{0:N}$. Let $Z_0, \ldots, Z_N$ be $N+1$ independent draws from a density κ, Nguyen and Ionides (2017) introduced $N+2$ perturbation parameters, τ and $\tau_0, \ldots, \tau_N$, and constructed a process $\breve{\Theta}_{0:N}$ by setting $\breve{\Theta}_n = \theta + \tau \sum_{i=0}^{n} \tau_i Z_i$ for $0 \le n \le N$. They designed a perturbed parameter log-likelihood function

$$\breve{\ell}(\breve{\vartheta}_{0:N}) = \log p_{\breve{Y}_{1:N}|\breve{\Theta}_{0:N}}(y^*_{1:N}|\breve{\vartheta}_{0:N};\theta,\tau,\tau_{0:N}). \tag{1}$$

so that the log-likelihood of the unperturbed model is

$$\ell(\theta) = \breve{\ell}(\breve{\vartheta}^{[N+1]}),$$

where

$$\breve{\vartheta}^{[N+1]} = (\theta, \theta, \ldots, \theta) \in \mathbb{R}^{d(N+1)}.$$

For the perturbed likelihood, they assume

Assumption 1. *$\breve{\ell}$ is four times continuously differentiable. For all $\theta \in \mathbb{R}^d$, there exist $\xi > 0$, $D > 0$ and δ defined as in Assumption 3, such that for all $0 < \eta_4 < \delta$ and $u_{0:N} \in \mathbb{R}^{d(N+1)}$,*

$$\breve{\mathcal{L}}(\breve{\vartheta}^{[N+1]} + u_{0:N}) \le D e^{\xi \sum_{n=1}^{N} |u_n|^{\eta_4}},$$

where $\breve{\mathcal{L}}(\breve{\vartheta}_{0:N}) = \exp\{\breve{\ell}(\breve{\vartheta}_{0:N})\}$ is the perturbed likelihood.

Assumption 2. *$\mathcal{K}$ is a set of symmetric probability density kernels for which each $\kappa \in \mathcal{K}$ is associated with a non-singular and finite covariance matrix, Ψ. There exists $C_6 < \infty$ such that for any integer $k \geq 1, 1 \leq i_1, \ldots, i_k \leq d$ and $\beta_1, \ldots, \beta_k \geq 1$ and any $\kappa \in \mathcal{K}$,*

$$\int \left| u_{i_1}^{\beta_1} u_{i_2}^{\beta_2} \cdots u_{i_k}^{\beta_k} \right| \kappa(u)\, du \leq C_6.$$

Assumption 3. *There exist $\gamma,\ \delta,\ M > 0$, such that for all $u \in \mathbb{R}^d$ and all $\kappa \in \mathcal{K}$,*

$$|u| > M \Rightarrow \kappa(u) < e^{-\gamma |u|^{\delta}}.$$

Assumption 4. *κ is mesokurtic, meaning that $\int u_i^4 \kappa(u) du = 3\sigma_i^4$.*

Lemma 1. *(Theorem 5 of Nguyen and Ionides (2017)). Suppose Assumptions 1, 2, 3 and 4 hold. In addition, assume that $\tau_n = O(\tau^2)$ for all $n = 1 \ldots N$. It follows that*

$$\left| \nabla \ell(\theta) - \frac{1}{N+1} \tau^{-2} \tau_0^{-2} \Psi^{-1} \sum_{n=0}^{N} \left\{ \breve{\mathbb{E}} \left(\breve{\Theta}_n - \theta | \breve{Y}_{1:N} = y_{1:N}^* \right) \right\} \right| = O(\tau^2). \tag{2}$$

where Ψ is the non-singular covariance matrix associated to κ.

An example of a density kernel satisfying the assumptions 1-4 is the Gaussian kernel. However, the family of distributions that satisfy these conditions is strictly larger than the Gaussian distribution (Doucet *et al.* 2013). Lemma 1 is useful for our approach because we can approximate the gradient of the log-likelihood of the extended model to the second order of τ which, later on, will fit well with our accelerated simulation-based setup.

3. Proposed accelerated iterated filtering

In the MLE problem, it is possible to use an accelerated gradient method in place of a naive stochastic approximation to improve the convergence rate of the estimations. One issue with the accelerated gradient approach is that it is not clear how the technique can be used in situations where both the likelihood and the gradient are intractable. These sorts of issues are common in scientific applications of state-space models (Poyiadjis *et al.* 2011; Nemeth *et al.* 2013; Dahlin *et al.* 2015). In addition, an approximation of the gradient is often biased (Kiefer and Wolfowitz 1952; Ionides *et al.* 2011; Doucet *et al.* 2013), making the application of the accelerated gradient method less straightforward. In this section, we first show that, under a chosen biased control policy, an accelerated gradient algorithm still converges at an optimal rate. Then, we apply it to the iterated filtering framework with a specified perturbation policy.

Let us denote $\{\epsilon_k\}$ the sequences of the errors in the gradient approximations of the log-likelihood and suppose the following assumption:

Assumption 5. *The function $\ell : \Theta \to \mathbb{R}$ is differentiable, bounded from above and has a L-Lipschitz-continuous gradient, i.e. $L > 0$ and for all $\theta,\ \vartheta \in \Theta$, $\|\nabla \ell(\theta) - \nabla \ell(\vartheta)\| \leq L \|\theta - \vartheta\|$, where $\nabla \ell$ denotes the gradient of ℓ. The function ℓ attains its maximum value ℓ^* at a certain $\theta^* \in \Theta$.*

In the assumption, Θ represents a subset of a finite-dimensional Euclidean space equipped with Euclidean norm $\|\cdot\|$ and inner product $\langle \cdot, \cdot \rangle$. It can be shown that (e.g. in Nesterov (2005)) Assumption 5 is equivalent to

$$|\ell(\vartheta) - \ell(\theta) - \langle \nabla \ell(\theta), \vartheta - \theta \rangle| \leq \frac{L}{2} \|\vartheta - \theta\|^2,\ \forall \theta,\ \vartheta \in \Theta. \tag{3}$$

It is well-known that the gradient ascent method converges for a general non-concave optimization problem but it does not achieve the optimal rate of convergence (in terms of the

functional optimality gap) when $\ell(\cdot)$ is concave (Ghadimi and Lan 2016). In contrast, the accelerated gradient method in Nesterov (2013) is optimal for solving concave optimization problems but does not necessarily converge for solving non-concave optimization problems. Ghadimi and Lan (2016) proposed a modified accelerated gradient method that can converge on both concave and non-concave optimization problems. However, they assumed an unbiased estimation of the gradient, which is not satisfactory for our simulation-based inference. Below, we adapt the approach of Ghadimi and Lan (2016) to accelerate a gradient ascent method while accounting for bias gradient. That is, we allow bias in gradient approximation. By properly specifying the biased control policy, we prove that it not only converges but also exhibits the optimal rate of convergence for both concave and non-concave log-likelihoods.

Algorithm 1 Accelerated Biased Gradient (ABG)

Input:

$\theta_0 \in \Theta$.
$\{\beta_k > 0\}$, $\{\lambda_k > 0\}$
$\{\alpha_k\} \in (0, 1)$ for $k > 1$ and $\alpha_1 = 1$.

1: $\theta_0^{ag} = \theta_0$. ▷ Initialize, ag stands for accelerated gradient
2: **for** k in $1...N$ **do**
3:

$$\theta_k^{ga} = (1 - \alpha_k)\theta_{k-1}^{ag} + \alpha_k \theta_{k-1} \tag{4}$$

▷ ga stands for gradient ascent
4:

$$\theta_k = \theta_{k-1} + \lambda_k \left(\widehat{\nabla \ell(\theta_k^{ga})}\right) \tag{5}$$

5:

$$\theta_k^{ag} = \theta_k^{ga} + \beta_k \left(\widehat{\nabla \ell(\theta_k^{ga})}\right) \tag{6}$$

▷ where $\widehat{\nabla \ell(\theta_k^{ga})}$ is an estimation of $\nabla \ell(\theta_k^{ga})$ with error ϵ_k.
6: **end for**

Along with Assumption 5, a biased control condition is also assumed for Algorithm 1.

Assumption 6. *Θ is a compact set. Let ϵ_k be the bias of an estimation $\widehat{\nabla \ell(\theta_k^{ga})}$. There exists an $A < \infty$ such that $\sum_{k=1}^{N} \lambda_k \|\epsilon_k\| < A$ for any $N \in \mathbb{N}$.*

These conditions, which are often satisfied by many simulation-based approximations, for example, in (Ionides *et al.* 2006, 2011), original iterated filtering assumes Θ is a compact set and the perturbation is a geometric series. Given these conditions, we have the following result.

Theorem 1. *(Extension of Theorem 1 of Ghadimi and Lan (2016)).*
Suppose Assumptions 5 and 6 hold. In addition, let $\{\theta_k,\ \theta_k^{ag}\}$, $k \geq 1$ be computed by Algorithm 1.
a) If sequences $\{\alpha_k\}, \{\beta_k\}, \{\lambda_k\}$ and $\{\Gamma_k\}$ satisfy

$$\Gamma_k := \begin{cases} 1 & k = 1 \\ (1 - \alpha_k)\Gamma_{k-1} & k \geq 2 \end{cases}, \tag{7}$$

$$C_k := 1 - L\lambda_k - \frac{L(\lambda_k - \beta_k)^2}{2\lambda_k \alpha_k \Gamma_k} \left(\sum_{\tau=k}^{N} \Gamma_\tau\right) > 0, \ \textit{for } 1 \leq k \leq N, \tag{8}$$

then for any $N \geq 1$, we have for some $B < \infty$,

$$\min_{k=1,\ldots,N} \left\|\nabla \ell(\theta_k^{ga}) + \epsilon_k\right\|^2 \leq \frac{\ell^* - \ell(\theta_0) + B}{\sum_{k=1}^{N} \lambda_k C_k}. \tag{9}$$

b) Suppose that $\ell(\cdot)$ is concave. If sequences $\{\alpha_k\}, \{\beta_k\}, \{\lambda_k\}$ and $\{\Gamma_k\}$ satisfy

$$\alpha_k \lambda_k \leq \beta_k < \frac{1}{L}, \tag{10}$$

$$\frac{\alpha_1}{\lambda_1 \Gamma_1} \geq \frac{\alpha_2}{\lambda_2 \Gamma_2} \geq \ldots, \tag{11}$$

then for any $N \geq 1$, we have

$$\min_{k=1,\ldots,N} \left\|\nabla \ell(\theta_k^{ga}) + \epsilon_k\right\|^2 \leq 2 \frac{\frac{\|\theta^* - \theta_0\|^2}{2\lambda_1} + \sum_{k=1}^N \Gamma_k^{-1} \left[\beta_k \|\epsilon_k\| \left\|\nabla \ell(\theta_k^{ga}) + \epsilon_k\right\| + \alpha_k \|\epsilon_k\| \|\theta_{k-1} - \theta_0\|\right]}{\sum_{k=1}^N \Gamma_k^{-1} \beta_k (1 - L\beta_k)}, \tag{12}$$

$$\ell(\theta^*) - \ell(\theta_N^{ag}) \leq \Gamma_N \left[\frac{\|\theta_0 - \theta^*\|^2}{2\lambda_1} + \sum_{k=1}^N \Gamma_k^{-1} \left[\beta_k \|\epsilon_k\| \left\|\nabla \ell(\theta_k^{ga}) + \epsilon_k\right\| + \alpha_k \|\epsilon_k\| \|\theta_{k-1} - \theta_0\|\right]\right]. \tag{13}$$

Various options are available for selecting $\{\alpha_k\}$, $\{\beta_k\}$, $\{\lambda_k\}$, $\{\Gamma_k\}$. By controlling the error ϵ_k, we can provide some of these selections below which ensure the optimal convergence rate of the ABG algorithm for both concave and non-concave problems.

Theorem 2. *Suppose Assumptions 5 and 6 hold. In addition, suppose that $\{\beta_k\}$ in the accelerated biased gradient method is set to $\beta_k = \frac{1}{2L}$, $\Gamma_k = \frac{1}{k^{1+\delta}}$, $\alpha_k = 1 - \frac{(k-1)^{1+\delta}}{k^{1+\delta}}$ for some $\delta > 0$. a) If sequence $\{\lambda_k\}$ satisfies*

$$\lambda_k \in \left[\beta_k, (1 + \frac{1}{k})\beta_k\right], \text{ for } \forall k \geq 1, \tag{14}$$

then for any $N \geq 1$, we have

$$\min_{k=1,\ldots,N} \|\nabla \ell(\theta_k^{ga}) + \epsilon_k\|^2 \leq O\left(\frac{1}{N}\right). \tag{15}$$

Suppose that $\epsilon_k = O\left(\tau^2\right) \leq O(\frac{1}{k})$, then the ABG method can find a solution $\bar{\theta}$ such that $\left\|\nabla \ell(\bar{\theta})\right\|^2 \leq \epsilon$ in at most $O(1/\epsilon^2)$ iterations.
b) Suppose that $\ell(\cdot)$ is concave and $\epsilon_k = O\left(\tau^2\right) \leq O(\frac{1}{k^{2+\delta+\delta_1}})$ for some $\delta, \delta_1 > 0$. If $\{\lambda_k\}$ satisfies

$$\lambda_k = \left(k^{1+\delta} - (k-1)^{1+\delta}\right) \forall k \geq 1, \tag{16}$$

then for any $N \geq 1$, we have

$$\min_{k=1,\ldots,N} \left\|\nabla \ell(\theta_k^{ga}) + \epsilon_k\right\|^2 \leq O\left(\frac{1}{N^{2+\delta}}\right), \tag{17}$$

$$\ell(\theta^*) - \ell(\theta_N^{ag}) \leq O\left(\frac{1}{N^{1+\delta}}\right), \tag{18}$$

and the ABG method can find a solution $\bar{\theta}$ such that $\left\|\nabla \ell(\bar{\theta})\right\|^2 \leq \epsilon$ in $O\left(1/\epsilon^{\frac{1}{2+\delta}}\right)$ at most.

We add a few remarks about the extension results obtained in Theorem 2, along with some guidelines for a perturbation policy to ensure Assumption 6 holds. First, if the problem is concave, choosing more aggressive step sizes $\{\lambda_k\}$ in (16) for the ABG method yields the optimal rate of convergence in (18). The result has been demonstrated by Nesterov (2005) and Ghadimi and Lan (2016) but only for the accelerated unbiased gradient methods. It is

also worth emphasizing that the ABG method can find a solution $\bar{\theta}$ such that $\left\|\nabla \ell(\bar{\theta})\right\|^2 \leq \epsilon$ in at most $O(1/\epsilon^{1/(2+\delta)})$ iterations with $\{\lambda_k\}$ is of order $O(k/L)$. To make the Assumption 6 held in this case, the perturbation sequence $\{\tau_k\} = O(1/k^b)$ for $b > 1$. For general non-concave problems, $\{\lambda_k\}$ is of order $O(1/L)$. In this situation, the perturbation sequence $\{\tau_k\}$ would only need to be of order $O(1/k^b$ for $b > 0.5$. Since we can control τ_k, we select the perturbation sequence $\{\tau_k\} = O(1/k^b)$ for $b > 1$ so that it works in both cases. The value δ is optimal at 1 for the convergence rate (see Appendix A). However, it may not be optimal for controlling the noises, so an analysis of the biases would be of great interest but it is beyond the scope of this paper.

Now, we are ready to present the pseudo-code of the proposed algorithm as in Algorithm 2.

Algorithm 2 Accelerated Iterated Filtering (AIF)

Input:

Starting parameter, $\theta_0 = \theta_0^{ag}$, sequences, $\alpha_n, \beta_n, \lambda_n, \Gamma_n$
simulator for $f_{X_0}(x_0|\theta)$, $f_{X_n|X_{n-1}}(x_n|x_{n-1}|\theta)$, evaluator for $g_{Y_n|X_n}(y_n|x_n|\theta)$
data, $y^*_{1:N}$, labels designating initial value parameters (IVPs), $I \subset \{1, \ldots, p\}$, initial scale multiplier, $C > 0$, number of particles, J, number of iterations, M, cooling rate, $0 < a < 1$, perturbation scales, $\sigma_{1:p}$

Output:

Maximum likelihood estimate θ_{MLE}

1: $\theta_0^{ga} = \theta_0$ ▷ Initialize
2: **for** m in 1...M **do**
3: $[\Theta^F_{0,j}]_i \sim \mathcal{N}\left([\theta_0^{ga}]_i, (Ca^{m-1}\sigma_i)^2\right)$ for i in 1..p, j in 1...J.
4: simulate $X^F_{0,j} \sim f_{X_0}\left(\cdot; \Theta^F_{0,j}\right)$ for j in 1..J. ▷ Initialize states
5: $\theta_m^{ga} = (1-\alpha_m)\theta_{m-1}^{ag} + \alpha_m \theta_{m-1}$.
6: **for** n in 1...N **do**
7: $\left[\Theta^P_{n,j}\right]_i \sim \mathcal{N}\left(\left[\theta_m^{ga}\right]_i, (a^{m-1}\sigma_i)^2\right)$ for $i \notin I$, j in $1:J$. ▷ Perturb
8: $X^P_{n,j} \sim f_n\left(x_n|X^F_{n-1,j}; \Theta^P_{n,j}\right)$ for j in $1:J$. ▷ Simulate prediction particles
9: $w(n,j) = g_n(y^*_n|X^P_{n,j}; \Theta^P_{n,j})$ for j in $1:J$. ▷ Evaluate weights
10: $\breve{w}(n,j) = w(n,j)/\sum_{u=1}^J w(n,u)$. ▷ Normalize weights
11: $k_{1:J}$ with $P\{k_u = j\} = \breve{w}(n,j)$. ▷ Apply systematic resampling to select indices
12: $X^F_{n,j} = X^P_{n,k_j}$ and $\Theta^F_{n,j} = \Theta^P_{n,k_j}$ for j in $1:J$. ▷ Resample particles
13: $\bar{\theta}_n = \sum_{j=1}^J \breve{w}(n,j)\Theta^F_{n,j}$
14: **end for**
15: $S_m = C^{-2}a^{-2(m-1)}\Psi^{-1}\sum_{n=1}^N \left[\left(\bar{\theta}_n - \theta_m^{ga}\right)\right]/(N+1)$ ▷ Update Parameters
16: $\left[\theta_m\right]_i = \theta_{m-1} + \lambda_{m-1}\left[S_m\right]_i$ for $i \notin I$.
17: $\left[\theta_m^{ag}\right]_i = \theta_m^{ga} + \beta_{m-1}\left[S_m\right]_i$ for $i \notin I$.
18: $\left[\theta_m\right]_i = \frac{1}{J}\sum_{j=1}^J \left[\Theta^F_j\right]_i$ for $i \in I$.
19: **end for**

It should be mentioned that this algorithm follows the general framework of iterated filtering family (Ionides *et al.* 2006, 2015), where inputs of the algorithm are the number of iterations M, the number of particles J, perturbation scales $\sigma_{1:p}$ simulator and evaluator at each time point. In addition, we also need four sequences $\{\alpha_k\}$, $\{\beta_k\}$, $\{\lambda_k\}$ and $\{\Gamma_k\}$. Here we use $\{\beta_k\} = 1/(2L)$, $\lambda_k = \beta_k + \beta_k/k$, $\alpha_k = 2/(k+1)$ for $k > 0$ and $\Gamma_k = 2/(k(k+1))$. It is worth noting that S_m in line 15 of Algorithm 2 is the estimated gradient $\widehat{\nabla \ell(\theta_k^{ga})}$ while the perturbation $\{\tau_k\}$ are absorbed into the constant C and cooling rate a. The cooling rate a are often selected in accordance with iterated filtering algorithm such that $a = 0.95^{0.02}$ while constant C is an initial scaling parameter chosen by the user. Other standard notation inherits from iterated filtering algorithm such as X^F denotes state filtering and X^P denotes state prediction. The initial value parameters (IVPs) in Algorithm 2 are the initial values of

the state variables at time 0, which can be treated as unknown parameters. The IVPs are, however, special parameters since they affect the dynamics only at time 0 and have no benefit to perturbing them at other times (Ionides *et al.* 2015).

4. Numerical examples

To measure the performance of the proposed inference algorithm, we evaluate our accelerated iterated filtering (AIF) against existing simulation-based approaches on some benchmark examples. In particular, we compare our algorithm to the original iterated filtering (IF1) (Ionides *et al.* 2006), the tuned iterated filtering algorithm (TIF) (**?**), and the recently developed Bayes map iterated filtering (IF2) (Ionides *et al.* 2015). It has been shown that the second-order iterated smoothing (IS2) Nguyen and Ionides (2017) is comparable to the Bayes map iterated filtering, while the Particle Markov chain Monte Carlo (PMCMC) Andrieu *et al.* (2010) belongs to the Bayesian approach, so we leave them out. We make use of the well-tested and maintained code of R package pomp (King, Nguyen, and Ionides 2016). Specifically, models are coded using C snippet declarations (King *et al.* 2016). The new algorithm is written in R package is2, which inherits user-friendly interfaces of R and efficient inference of pomp (King *et al.* 2016). In all the simulation-based approaches mentioned above, the sequential Monte Carlo algorithm (SMC) is used, implemented by a bootstrap filter (Gordon, Salmond, and Smith 1993). Experiments were carried out on 32 cores Intel Xeon E5-2680 2.7 Ghz with 256 GB memory. For fair comparisons, the perturbation configuration and initial starting point are the same for every inference method. Scripts for reproducing our results are available in a public GitHub repository at `https://github.com/nxdao2000/AIFcomparisons`.

4.1. Linear toy example

For a computationally convenient setting, simple models can be used to test the basic features of inference algorithms. As a first step, we consider a bivariate linear Gaussian model. We chose this model so that the Monte Carlo simulations can be verified using Kalman filters. In this example, alternative approaches such as expected maximization (EM) or Markov chain Monte Carlo (MCMC) algorithms would also be practical, but they are not simulation-based and generally do not scale well to large dynamic models, so we do not include them here. The model is given by the state space forms: $X_n|X_{n-1} = x_{n-1} \sim \mathcal{N}(\alpha x_{n-1}, \sigma^\top \sigma)$, $Y_n|X_n = x_n \sim \mathcal{N}(x_n, I_2)$ where α, σ are 2×2 matrices and I_2 is 2×2 identity matrix. The following parameters are used to simulate the data:

$$\alpha = \begin{bmatrix} \alpha_1 & \alpha_2 \\ \alpha_3 & \alpha_4 \end{bmatrix} = \begin{bmatrix} 0.8 & -0.5 \\ 0.3 & 0.9 \end{bmatrix}, \ \sigma = \begin{bmatrix} 3 & 0 \\ -0.5 & 2 \end{bmatrix},$$

with the initial starting point X_0 is set to $(-3, 4)$ and the number of time points N is set to 100. For each method mentioned above, we estimate parameters α_2 and α_3 for this model using $J = 1000$ particles and run our estimation for $M = 25$ iterations. We start the initial search uniformly on a rectangular region $[-1, 1] \times [-1, 1]$.

Table 1 displays the results of estimating parameters α_2 and α_3 for the bivariate linear Gaussian model using IF1, IF2, TIF, and AIF algorithms. The last row shows the exact MLE computed from the Kalman filter and the first row shows the true value of the parameter. The first two columns present the estimated values of the parameters while the next two columns display the log likelihood, $\hat{\ell}$, computed by SMC with the number of particles 10000 and its standard error, respectively. The exact log likelihood, ℓ, and the time(s) are given in the two rightmost columns.

Table 1: Summary results of different algorithms for the bivariate linear Gaussian model

Algorithms	α_2	α_3	$\hat{\ell}$	s.e.	ℓ	time(s)
Truth	-0.5000	0.3000	-478.9700	0.08	-478.99	
IF1	-0.5340	0.2860	-480.0700	0.09	-480.11	26.37
IF2	-0.4800	0.3370	-479.8300	0.09	-479.84	25.89
TIF	-0.5600	0.2740	-479.0900	0.08	-479.10	30.97
AIF	-0.5430	0.3090	-479.0900	0.09	-479.01	31.41
Exact MLE	-0.5280	0.2860	-478.7300	0.26	-478.79	

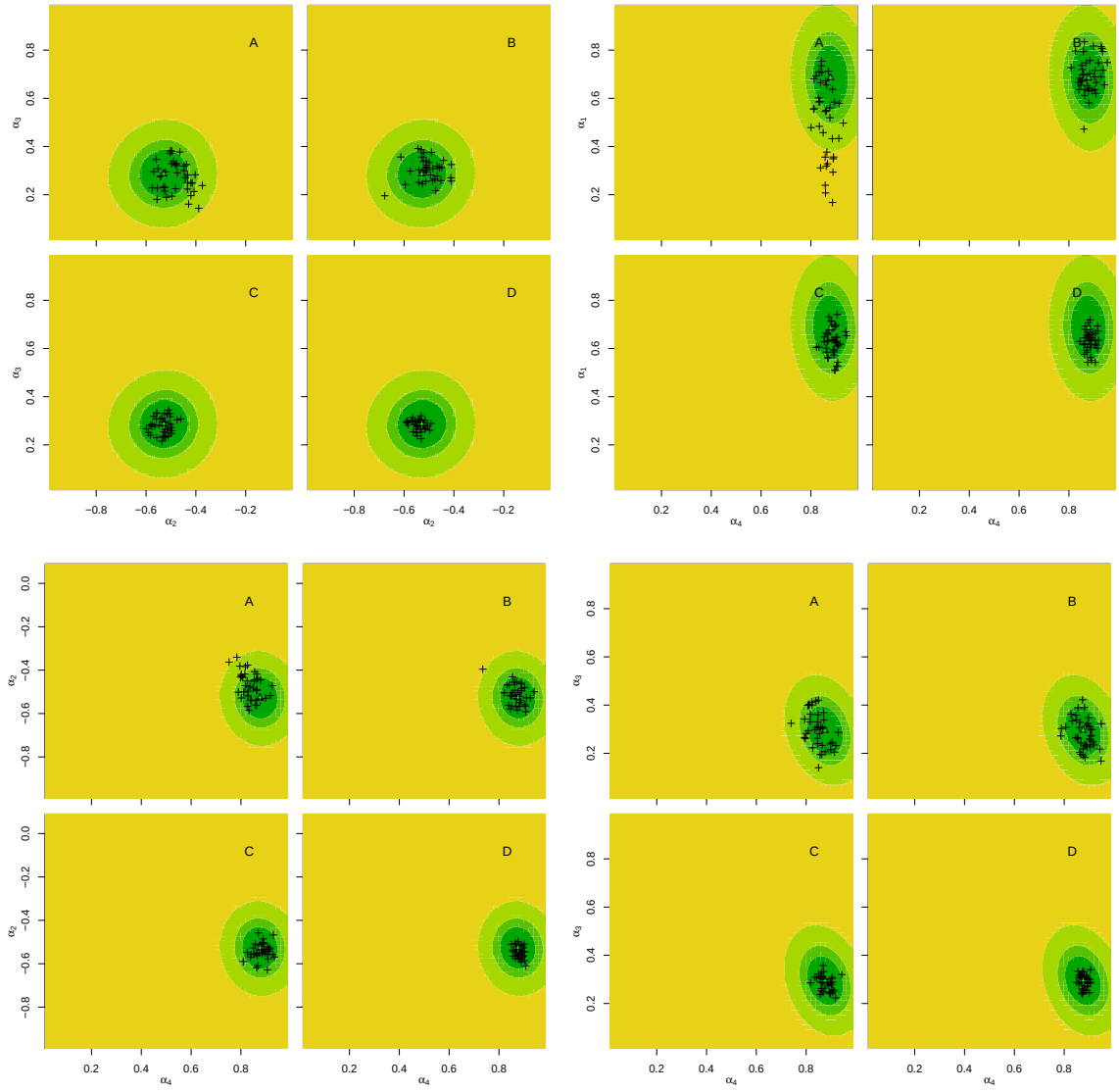

Figure 1: Comparison of different estimators. The likelihood surface for the bivariate linear Gaussian model, with the location of the MLE is marked with a green cross. The black crosses show final points from 40 Monte Carlo replications of the estimators: (A) Original iterated filtering method; (B) Bayes map iterated filtering method; (C) Tuned iterated filtering method; (D) Accelerated iterated filtering method; Each method was started uniformly over the rectangle region $[-1, 1]\times[-1, 1]$ with $M = 25$ iterations, $J = 1000$ particles, and a random walk standard deviation decreasing from 0.02 geometrically to 0.011 for both estimated parameters. (TopLeft) The likelihood surface for α_2, α_3; (TopRight) The likelihood surface for α_1, α_4; (BottomLeft) The likelihood surface for α_2, α_4;(BottomRight) The likelihood surface for α_3, α_4.

We estimate parameters using the highest estimated likelihood between 20 independent runs, evaluating their likelihood and standard error based on 20 replications to reduce the Monte Carlo error in the likelihood evaluation employing the particle filter. Since the particle filter produces an unbiased estimate of the likelihood, we average the likelihoods and calculate the standard errors. As noted by King *et al.* (2016), an ideal likelihood-ratio 95% confidence set is expected to be within qchisq $(0.95, \text{df} = 2)/2 = 2.99$ of the exact MLE. As seen from the table, it is the case in this example for all methods, but accelerated iterated filtering appears to be the best with the smallest error of about 0.22 log units from the exact MLE. By using AIF, the results have higher estimated likelihoods compared to other approaches, indicating a higher empirical convergence rate.

To see how the final MLEs clustered around the true MLE, we only use 40 Monte Carlo replications for this toy example. As can be observed from Fig. 1, most of the replications clustered near the true MLE for the AIF approach, while none of them stayed in a lower likelihood region. As shown in this figure, AIF appears to be the most effective method compared to other methods considered for this test. Given additional computational resources, we also checked how the results of each method were compared. Specifically, we set $M = 100$ iterations and $J = 10000$ particles, with the random walk standard deviation decreasing geometrically from 0.02 down to 0.0018 for each method. Based on 200 Monte Carlo replications, Fig. 2 can be viewed as a statistical summary of this example. In this situation, we confirm that AIF is the most effective among other IF1, IF2, and TIF, while all methods have comparable computational demands for a given M and J.

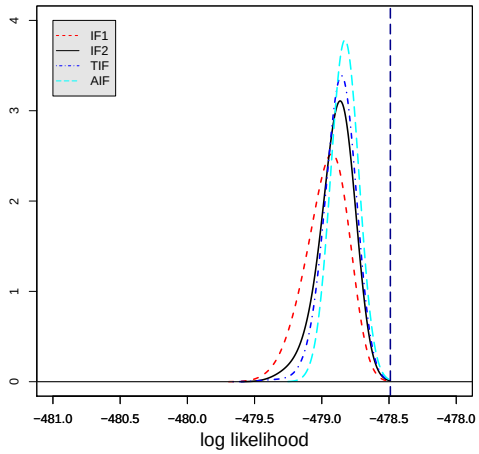

Figure 2: Comparison of estimators for the bivariate linear Gaussian model, showing the densities of the maximized log-likelihood computed by the IF1, IF2, TIF and AIF methods using $M = 100$ iterations and $J = 10000$ particles. The parameters α_2 and α_3 were estimated, started from 200 randomly uniform initial values over a rectangular region $[-1, 1] \times [-1, 1]$.

Moreover, the results also imply that AIF is robust to initial starting guesses. Algorithmically, AIF has similar computational costs to the first-order approaches IF1 and IF2. Additional overheads for estimating the score function make the computation time of AIF a bit larger compared to that of IF1 and IF2. However, with complex models and a large enough number of particles, the overheads become negligible and the computation time of AIF will be similar to other first-order approaches. The fact that it has a higher convergence rate with the comparable computational complexity of the first-order implies that it is a very promising algorithm.

4.2. Nonlinear toy examples

Besides linear models, some nonlinear models are also examined to test the capacity of the proposed algorithm. However, unlike linear models, nonlinear models generally can not be verified using Kalman filters. To overcome this difficult problem, we use particle filter with a large number of particles to evaluate the likelihood at the ground truth points. Specifically, we use the number of particles $J = 20000$ to evaluate the likelihood surface. Here, to illustrate the behavior of algorithms, and on the other hand, to simplify the computational procedures, we only focus on a few parameters of interest while fixing the rest of the parameters. We start from 40 randomly uniform initial value over the region of interests and run different inference methods using $M = 100$ iterations. The purpose is to show that, with a sufficiently large computational resource, every method will converge to the MLEs. However, when resources are limited, which frequently occurs in the big data regime, only a few iterations are executed or not many number of particles are used, the proposed algorithm outperforms its counterparts. This will be demonstrated in the following subsections.

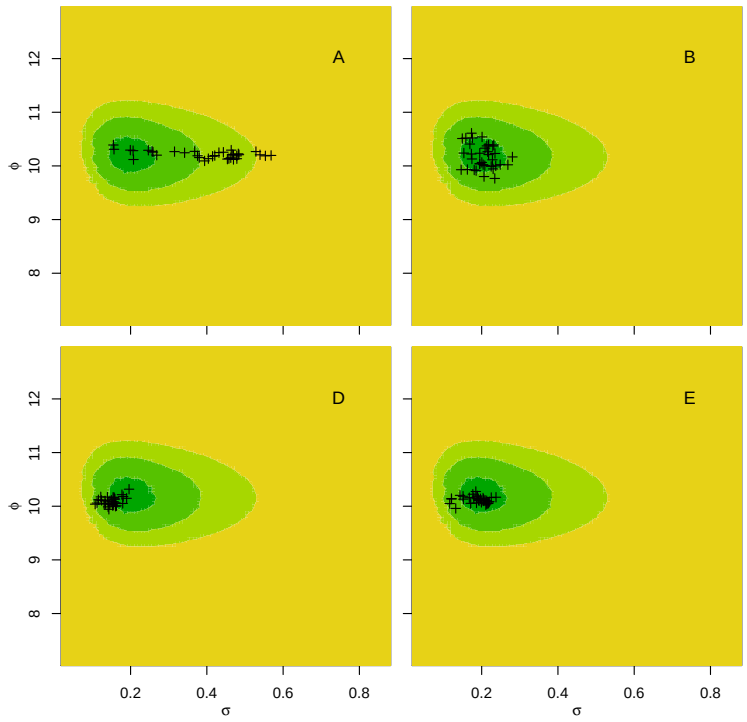

Figure 3: Comparison of MLEs computed by the IF1, IF2, TIF and AIF methods for the Ricker model, using $M = 100$ iterations and $J = 1000$ particles. The parameters σ and ϕ was estimated, started from 40 randomly uniform initial values over a interval $[0, 1] \times [7, 13]$.

Ricker model

We consider Ricker model, which was introduced by (Ricker 1954) in the context of stock and recruitment fisheries. In this model, the state process is

$$N_{t+1} = rN_t \exp(-cN_t + e_t),$$

where the e_t are i.i.d. normal random deviates with zero mean and variance σ^2. The observed variables y_t are modeled by Poisson(ϕN_t) distribution. We are interested in estimating the parameters σ and ϕ started from 200 randomly uniform initial value over the region $[0, 1] \times [7, 13]$ using $J = 1000$ particles. As can be seen from Figure 3, after 100 iterations, accelerated iterated filtering seems to be the best algorithm in approaching the MLE in terms of root mean square errors.

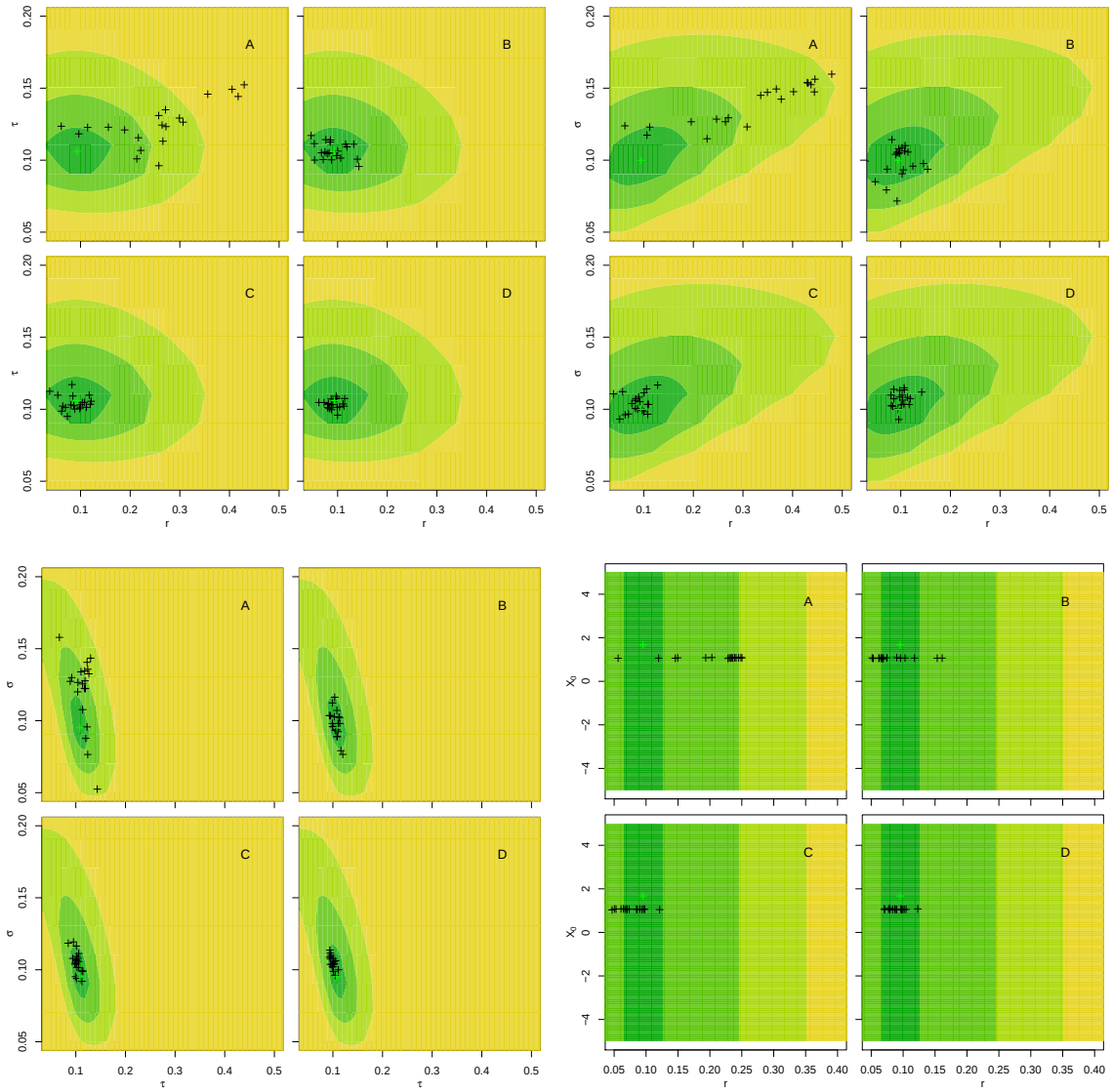

Figure 4: Comparison of different estimators. The likelihood surface for the Gompertz model, with the location of the MLE is marked with a green cross. The black crosses show final points from 40 Monte Carlo replications of the estimators: (A) Original iterated filtering method; (B) Bayes map iterated filtering method; (C) Tuned iterated filtering method; (D) Accelerated iterated filtering method; Each method was started uniformly over the rectangle region of intertest with $M = 100$ iterations, $J = 200$ particles, and a random walk standard deviation decreasing from 0.03 geometrically to 0.027 for both parameters. (TopLeft) The likelihood surface for τ, r; (TopRight) The likelihood surface for σ, r; (BottomLeft) The likelihood surface for σ, τ;(BottomRight) The likelihood surface for r, X_0.

Gompertz model

The Gompertz model, which is frequently used to model animal growth in biology, is also considered. Let the state process be $X_{t+1} = K^{1-S} X_t^S \epsilon_t$, where $S = e^{-r}$ and the ϵ_t are i.i.d. lognormal random deviates with variance σ^2. The observed variables Y_t are following lognormal($\log X_t$; τ) distribution. The parameters include the per-capita growth rate r, the carrying capacity K, the process noise standard deviation σ, the measurement error standard deviation τ, and the initial condition X_0. In this model, we want to estimate the measurement error standard deviation τ. Note that we use log-transform to remove the positive constraints for estimation purposes. We start the initial search uniformly on the (0,1) interval using $J = 200$ particles and run our estimation for $M = 100$ iterations.

After transforming, we can estimate the true likelihood using the Kalman filter in this relatively simple example. From the experiment, we observe similar patterns as in the Ricker model that, with $M = 100$ iterations, all inference methods converge well to the MLEs but with a small number of particles (e.g. $J = 200$), the proposed algorithm can converge faster to the MLEs compared to other approaches. This feature is desirable, especially in situations where computational resources are limited.

4.3. Malaria benchmark

The majority of dynamic systems in the real world are highly nonlinear, partially observed, or even weakly identifiable. To demonstrate the capabilities of accelerated iterated filtering for such situations, we apply it to evaluate the likelihood of a stochastic differential equation for malaria with relapse model in northwest India of Roy, Bouma, Ionides, Dhiman, and Pascual (2013). We chose this challenging model because it provides a rigorous performance benchmark for our verification. The model we consider splits up the study population of size $P(t)$ into different classes: susceptible individuals, $S(t)$, exposure $E(t)$, infected individuals, $I(t)$, dormant classes $H(t)$ and recovered individuals, $Q(t)$. The dormant class H in the model is further subdivided into three classes $H_1(t)$, $H_2(t)$, $H_3(t)$ to allow some flexibilities in representing relapse. The state process is written as

$$X(t) = \big(S(t), E(t), I(t), Q(t), H_1(t), H_2(t), H_3(t), \kappa(t), \mu_{SE}(t)\big),$$

where infected population enters dormancy transition rate is μ_{IH} and transition rates from stage H_1 to H_2, H_2 to H_3 and H_3 to Q are specified to be $3\mu_{HI}$. The model satisfies the following balance equation system

$$\begin{aligned}
dS/dt &= \delta P + \mathrm{d}P/dt + \mu_{IS} I + \mu_{QS} Q \\
&\quad + a\mu_{IH} I + b\mu_{EI} E - \mu_{SE}(t) S - \delta S, \\
dE/dt &= \mu_{SE}(t) S - \mu_{EI} E - \delta E, \\
dI/dt &= (1-b)\mu_{EI} E + 3\mu_{HI} H_n - (\mu_{IH} + \mu_{IS} + \mu_{IQ}) I - \delta I, \\
dH_1/dt &= (1-a)\mu_{IH} I - n\mu_{HI} H_1 - \delta H_1, \\
dH_2/dt &= 3\mu_{HI} H_1 - 3\mu_{HI} H_2 - \delta H_2, \\
dH_3/dt &= 3\mu_{HI} H_2 - 3\mu_{HI} H_3 - \delta H_3 \\
dQ/dt &= \mu_{IQ} I - \mu_{QS} Q - \delta Q.
\end{aligned}$$

The malaria pathogen reproduction within the mosquito vector also satisfies

$$\begin{aligned}
\mathrm{d}\kappa/\mathrm{d}t &= [\lambda(t) - \kappa(t)]/\tau_D, \\
\mathrm{d}\mu_{SE}/\mathrm{d}t &= [\kappa(t) - \mu_{SE}(t)]/\tau_D.
\end{aligned}$$

In this equation, $\lambda(t)$ is the latent force of infection and $\lambda(t)$, $\kappa(t)$ and $\mu_{SE}(t)$ is given by

$$\mu_{SE}(t) = \int_{-\infty}^{t} \gamma(t-s)\lambda(s)\mathrm{d}s, \tag{19}$$

with $\gamma(s) = \frac{(2/\tau_D)^2 s^{2-1}}{(2-1)!} \exp(-2s/\tau_D)$, a gamma distribution with shape parameter 2. Since the latent force of infection is constrained by rainfall covariate $R(t)$ and some Gamma white noise, from Roy *et al.* (2013) we have: m

$$\lambda(t) = \left(\frac{I + qQ}{P}\right) \times \exp\left\{\sum_{i=1}^{N_s} b_i s_i(t) + b_r R(t)\right\} \times \left[\frac{\mathrm{d}\Gamma(t)}{\mathrm{d}t}\right].$$

where q denotes a reduced infection risk from humans in the Q class. $\{s_i(t), i = 1, \ldots, N_s\}$ is a periodic cubic B-spline basis, and we set $N_s = 6$. $M_n = \rho \int_{t_{n-1}}^{t_n} [\mu_{EI} E(s) + 3\mu_{HI} H_3(s)] ds$

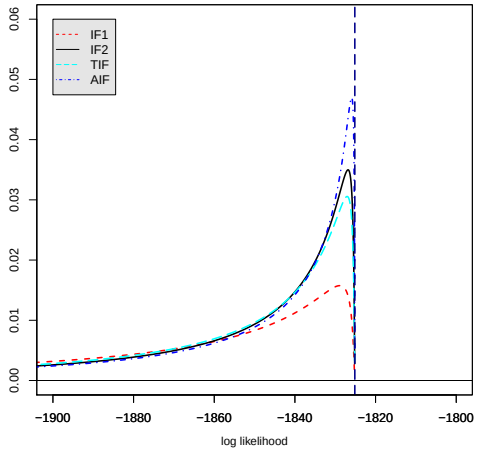

Figure 5: The density of the maximized log-likelihoods estimated by IF1, IF2, TIF and AIF for the malaria model when using J = 1000 and M = 50. The log-likelihood at a computed MLE is shown as a dashed vertical line.

is the cumulative number of cases observed from time t_{n-1} to time t_n and ρ is the mean age. To account for the under-reported fact, the observation model for Y_n is a negative binomial distribution with mean M_n and variance $M_n + M_n^2\sigma_{\text{obs}}^2$. In our case, we use an Euler-Maruyama scheme (Kloeden and Platen 1999) with a time step of 1/20 month to solve the coupled system of stochastic differential equations.

We carried out simulation-based inference via the original iterated filtering (IF1), the Bayes map iterated filtering (IF2), and the proposed accelerated iterated filtering (AIF). The inference goal used to assess all of these methods is to find high likelihood parameter values starting from randomly drawn values in a large hyper-rectangle. We provide this initial hyper-rectangle in the appendix. In the presence of possible multi-modality, weak identifiability, and considerable Monte Carlo error of this model, we start 100 random searches. The random walk standard deviation is initially set to 0.1 for estimated parameters while the cooling rate c is set to $0.1^{0.02} \approx 0.95$. These corresponding quantities for initial value parameters are 2 and $0.1^{0.02}$, respectively, but they are applied only at time zero. Our experiment runs on a cluster computer with $M = 50$ iterations and $J = 1000$ particles. Figure 5 shows the MLEs estimated by IF1, IF2, TIF and AIF with standard errors. With the higher mean and smaller variance of IF2, TIF, AIF estimation clearly demonstrates that they are considerably more effective than IF1. Note that the computational times for IF1, IF2, TIF and AIF are 44.27, 43.83, 52.14 and 52.35 minutes respectively, confirming that accelerated iterated filtering has essentially the same computational cost as first-order methods IF1, IF2 for a given Monte Carlo sample size and the number of iterations.

Experimentation with more extensive computation ($M = 100$ and $J = 10^4$) in Figure 6 suggests that the performance improvement of AIF over IF2 occurs primarily in simpler models, such as the toy example, or during earlier stages of optimization on complex models. We have had similar experiences with other complex models (Nguyen and Ionides 2017). Our interpretation is that the parameter interpolating involved in the parameter update rule for AIF can be inefficient when the likelihood surface contains non-linear ridges, whereas the IF2 algorithm does not carry out any interpolating in parameter space. In this hard problem, while IF1 reveal their limitations, we have shown that IF2, TIF and especially AIF can offer a substantial improvement. As shown in Fig 6, the proposed method has a clear advantage over IF1 and may be an alternative to IF2 and TIF in many applications. Note that IF2 is very efficient in climbing along the ridge of the likelihood, so a natural heuristic idea to further improve the method is hybridizing IF2 and AIF but we leave it for the future work.

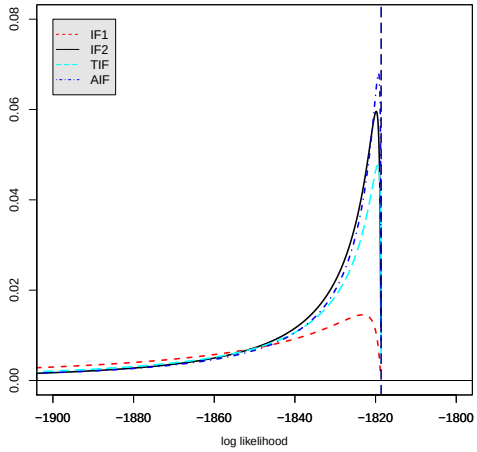

Figure 6: The density of the maximized log-likelihoods estimated by IF1, IF2, TIF and AIF for the malaria model when using $J = 10000$ and $M = 100$. The-log likelihood at a computed MLE is shown as a dashed vertical line.

5. Conclusion

In this paper, we have proposed an efficient simulation-based algorithm using an accelerated biased gradient approach. We have shown that choosing a suitable perturbation sequence results in an algorithm that leads to some advances including statistical efficiency and computational efficiency. Only standard gradient conditions are used and a more systematic approach could be generalized using the state-of-the-art algorithm in the optimization literature such as proximal theory. The convergence rates are also explicitly stated. From a theoretical point of view, it could be an interesting perspective.

From a practical point of view, we have provided an efficient framework, applicable to a general class of nonlinear, non-Gaussian POMP models, especially suitable for infectious disease modeling and control feedback systems. There are many such systems, which could be well-treated by the proposed framework. We also provide an open-source software R package, IS2, which could be useful for the community to further explore in this direction. In certain models, this novel approach may be of interest compared to the other approaches in this area.

In principle, different simulation-based inference methods can be hybridized to build on the strongest features of the multiple algorithms. Our results could also be applied to develop other simulation-based methodologies which can take advantage of the optimal convergence rate of accelerated methods. For example, it may be possible to use our approach to help design efficient proposal distributions for particle Markov chain Monte Carlo algorithms. Applying this approach to methodologies like Approximate Bayesian Computation (ABC) (Beaumont, Zhang, and Balding 2002; Sisson *et al.* 2007; Toni *et al.* 2009), Liu-West Particle Filter (LW-PF) (**?**), and Particle Markov chain Monte Carlo (PMCMC) (Andrieu *et al.* 2010) with different sampler schemes, such as forward-backward particle filter (Huys and Paninski 2006), forward smoothing (Del Moral, Doucet, and Singh 2010), or forward filter-backward smoothing (Doucet, Godsill, and Andrieu 2000), are foreseeable extensions.

A. Proofs

We first need a simple technical result (see Lemma 1 of Ghadimi and Lan (2016)). We provide it here for completeness.

Lemma 2 (Lemma 1 of Ghadimi and Lan (2016))**.** .

Assume sequences $\{\alpha_k\} \in (0,1)$ *for* $k > 1$ *and* $\alpha_1 = 1$ *and sequences* $\{a_k\}$, $\{\eta_k\}$ *satisfy*

$$a_k \leq (1 - \alpha_k)a_{k-1} + \eta_k, \ k = 1, 2, \ldots \tag{20}$$

If we define a positive sequence $\{\Gamma_k\}$ *as in* (7) *then for any* $k \geq 1$, *we have*

$$a_k \leq \Gamma_k \sum_{i=1}^{k} (\eta_i / \Gamma_i).$$

Proof. Since $\alpha_1 = 1$ and $\Gamma_1 = 1$, from (7) we have

$$a_1 \leq \eta_1$$

or

$$\frac{a_1}{\Gamma_1} \leq \frac{\eta_1}{\Gamma_1}.$$

Since $\Gamma_k > 0$ for every $k > 1$, dividing both sides of (20) by Γ_k,

$$\frac{a_k}{\Gamma_k} \leq \frac{(1-\alpha_k)a_{k-1} + \eta_k}{\Gamma_k} = \frac{a_{k-1}}{\Gamma_{k-1}} + \frac{\eta_k}{\Gamma_k}, \ \forall k \geq 2.$$

Summing up the above inequalities and rearranging the terms, the conclusion follows. □

Lemma 3.

$$\sum_{\tau=1}^{k} \frac{\alpha_\tau}{\Gamma_\tau} = \frac{1}{\Gamma_k}. \tag{21}$$

Proof. We have

$$\sum_{\tau=1}^{k} \frac{\alpha_\tau}{\Gamma_\tau} = \frac{\alpha_1}{\Gamma_1} + \sum_{\tau=2}^{k} \frac{1}{\Gamma_\tau}(1 - (1 - \alpha_\tau))$$

$$= \frac{1}{\Gamma_1} + \sum_{\tau=2}^{k} (\frac{1}{\Gamma_\tau} - \frac{1}{\Gamma_{\tau-1}}) = \frac{1}{\Gamma_k}.$$

□

A.1. Proof of Theorem 1

Proof. The proof follows closely to the proof of theorem 1 of Ghadimi and Lan (2016) except we consider bias estimate of the gradient. We first prove part a.

By (3) and (5), we have

$$-\ell(\theta_k) \leq -\ell(\theta_{k-1}) + \langle -\nabla\ell(\theta_{k-1}), \ \theta_k - \theta_{k-1}\rangle + \frac{L}{2}\|\theta_k - \theta_{k-1}\|^2$$

$$= -\ell(\theta_{k-1}) + \left\langle \left(-\nabla\ell(\theta_{k-1}) + \nabla\ell(\theta_k^{ga}) + \epsilon_k\right) - \left(\nabla\ell(\theta_k^{ga}) + \epsilon_k\right), \ \lambda_k \left(\nabla\ell(\theta_k^{ga}) + \epsilon_k\right)\right\rangle$$

$$+ \frac{L\lambda_k^2}{2}\|\nabla\ell(\theta_k^{ga}) + \epsilon_k\|^2$$

$$= -\ell(\theta_{k-1}) - \lambda_k(1 - \frac{L\lambda_k}{2})\|\nabla\ell\,(\theta_k^{ga}) + \epsilon_k\|^2 + \lambda_k \left\langle \left(-\nabla\ell(\theta_{k-1}) + \nabla\ell(\theta_k^{ga}) + \epsilon_k\right), \ \left(\nabla\ell(\theta_k^{ga}) + \epsilon_k\right)\right\rangle$$

$$\leq -\ell(\theta_{k-1}) - \lambda_k(1 - \frac{L\lambda_k}{2})\|\nabla\ell\,(\theta_k^{ga}) + \epsilon_k\|^2 + \lambda_k \left(\| - \nabla\ell(\theta_{k-1}) + \nabla\ell(\theta_k^{ga})\| + \|\epsilon_k\|\right) \cdot \|\nabla\ell(\theta_k^{ga}) + \epsilon_k\|,$$

$$\leq -\ell(\theta_{k-1}) - \lambda_k(1 - \frac{L\lambda_k}{2})\|\nabla\ell\,(\theta_k^{ga}) + \epsilon_k\|^2 + \lambda_k \left(L\|\theta_{k-1} - \theta_k^{ga}\| + \|\epsilon_k\|\right) \cdot \|\nabla\ell(\theta_k^{ga}) + \epsilon_k\|,$$

$$= -\ell(\theta_{k-1}) - \lambda_k(1-\frac{L\lambda_k}{2})\|\nabla\ell\left(\theta_k^{ga}\right)+\epsilon_k\|^2+\lambda_k\left(L(1-\alpha_k)\|\theta_{k-1}^{ag}-\theta_{k-1}\|+\|\epsilon_k\|\right)\cdot\|\nabla\ell(\theta_k^{ga})+\epsilon_k\|,$$

$$= -\ell(\theta_{k-1}) - \lambda_k\left(1-\frac{L\lambda_k}{2}\right)\left\|\nabla\ell(\theta_k^{ga})+\epsilon_k\right\|^2$$

$$+L(1-\alpha_k)\lambda_k\left\|\nabla\ell(\theta_k^{ga})+\epsilon_k\right\|\cdot\left\|\theta_{k-1}^{ag}-\theta_{k-1}\right\|+\lambda_k\left\|\epsilon_k\right\|\cdot\left\|\nabla\ell(\theta_k^{ga})+\epsilon_k\right\|$$

$$\leq -\ell\left(\theta_{k-1}\right) - \lambda_k\left(1-\frac{L\lambda_k}{2}\right)\left\|\nabla\ell(\theta_k^{ga})+\epsilon_k\right\|^2$$

$$+\frac{L\lambda_k^2}{2}\left\|\nabla\ell\left(\theta_k^{ga}\right)+\epsilon_k\right\|^2+\frac{L(1-\alpha_k)^2}{2}\left\|\theta_{k-1}^{ag}-\theta_{k-1}\right\|^2+\lambda_k\left\|\epsilon_k\right\|\cdot\left\|\nabla\ell(\theta_k^{ga})+\epsilon_k\right\|$$

$$= -\ell(\theta_{k-1}) - \lambda_k(1-L\lambda_k)\left\|\nabla\ell(\theta_k^{ga})+\epsilon_k\right\|^2$$

$$+\frac{L(1-\alpha_k)^2}{2}\left\|\theta_{k-1}^{ag}-\theta_{k-1}\right\|^2+\lambda_k\left\|\epsilon_k\right\|\cdot\left\|\nabla\ell(\theta_k^{ga})+\epsilon_k\right\| \tag{22}$$

The second inequality is from triangular inequality and the Cauchy-Schwarz inequality while the second inequality is due to the Lipschitz of gradient assumption 5 and the last equality coming from (4). We have the last inequality follows from $ab \leq (a^2+b^2)/2$. From (4), (5), and (6), it follows that

$$\theta_k^{ag}-\theta_k = (1-\alpha_k)\theta_{k-1}^{ag}+\alpha_k\theta_{k-1}-\beta_k\left(\nabla\ell(\theta_k^{ga})+\epsilon_k\right)-\left(\theta_{k-1}-\lambda_k\left(\nabla\ell(\theta_k^{ga})+\epsilon_k\right)\right)$$

$$= (1-\alpha_k)(\theta_{k-1}^{ag}-\theta_{k-1})+(\lambda_k-\beta_k)\left(\nabla\ell(\theta_k^{ga})+\epsilon_k\right).$$

Applying Lemma 2 where $\theta_k^{ag}-\theta_k := a_k$ and $\eta_k := (\lambda_k-\beta_k)\left(\nabla\ell(\theta_k^{ga})+\epsilon_k\right)$, we obtain

$$\theta_k^{ag}-\theta_k = \Gamma_k\sum_{\tau=1}^{k}\left(\frac{\lambda_\tau-\beta_\tau}{\Gamma_\tau}\right)\left(\nabla\ell(\theta_\tau^{ga})+\epsilon_\tau\right).$$

Since$\|\cdot\|^2$ is convex, using Jensen's inequality and Lemma 3 we have

$$\left\|\theta_k^{ag}-\theta_k\right\|^2 = \left\|\Gamma_k\sum_{\tau=1}^{k}\left(\frac{\lambda_\tau-\beta_\tau}{\Gamma_\tau}\right)\left(\nabla\ell\left(\theta_\tau^{ga}\right)+\epsilon_k\right)\right\|^2$$

$$= \left\|\Gamma_k\sum_{\tau=1}^{k}\frac{\alpha_\tau}{\Gamma_\tau}\left[\left(\frac{\lambda_\tau-\beta_\tau}{\alpha_\tau}\right)\left(\nabla\ell(\theta_\tau^{ga})+\epsilon_k\right)\right]\right\|^2$$

$$\leq \Gamma_k\sum_{\tau=1}^{k}\frac{\alpha_\tau}{\Gamma_\tau}\left\|\left(\frac{\lambda_\tau-\beta_\tau}{\alpha_\tau}\right)\left(\nabla\ell(\theta_\tau^{ga})+\epsilon_k\right)\right\|^2$$

$$= \Gamma_k\sum_{\tau=1}^{k}\frac{(\lambda_\tau-\beta_\tau)^2}{\Gamma_\tau\alpha_\tau}\left\|\nabla\ell(\theta_\tau^{ga})+\epsilon_\tau\right\|^2. \tag{23}$$

Replacing the above bound in (22), and the fact that $\Gamma_k = \Gamma_{k-1}(1-\alpha_k)$ as in (7) and that $\alpha_k \in (0,\ 1]$ for all $k \geq 1$we obtain

$$-\ell(\theta_k) \leq -\ell(\theta_{k-1}) - \lambda_k(1-L\lambda_k)\left\|\nabla\ell(\theta_k^{ga})+\epsilon_k\right\|^2$$

$$+\frac{L\Gamma_{k-1}(1-\alpha_k)^2}{2}\sum_{\tau=1}^{k-1}\frac{(\lambda_\tau-\beta_\tau)^2}{\Gamma_\tau\alpha_\tau}\left\|\nabla\ell(\theta_\tau^{ga})+\epsilon_\tau\right\|^2+\lambda_k\left\|\epsilon_k\right\|\cdot\left\|\nabla\ell(\theta_k^{ga})+\epsilon_k\right\|$$

$$\leq -\ell(\theta_{k-1}) - \lambda_k(1-L\lambda_k)\left\|\nabla\ell(\theta_k^{ga})+\epsilon_k\right\|^2$$

$$+\frac{L\Gamma_k}{2}\sum_{\tau=1}^{k}\frac{(\lambda_\tau-\beta_\tau)^2}{\Gamma_\tau\alpha_\tau}\left\|\nabla\ell(\theta_\tau^{ga})+\epsilon_\tau\right\|^2+\lambda_k\left\|\epsilon_k\right\|\cdot\left\|\nabla\ell(\theta_k^{ga})+\epsilon_k\right\| \tag{24}$$

for every $k \geq 1$. Using the definition of C_k in (8) and summing up the above inequalities, we have

$$-\ell(\theta_N) \leq -\ell(\theta_0) - \sum_{k=1}^{N} \lambda_k (1 - L\lambda_k) \left\|\nabla \ell(\theta_k^{ga}) + \epsilon_k\right\|^2$$

$$+\frac{L}{2} \sum_{k=1}^{N} \Gamma_k \sum_{\tau=1}^{k} \frac{(\lambda_\tau - \beta_\tau)^2}{\Gamma_\tau \alpha_\tau} \left\|\nabla \ell(\theta_\tau^m) + \epsilon_k\right\|^2 + \sum_{k=1}^{N} \lambda_k \epsilon_k \cdot \left\|\nabla \ell(\theta_k^{ga}) + \epsilon_k\right\|$$

$$= -\ell(\theta_0) - \sum_{k=1}^{N} \lambda_k (1 - L\lambda_k) \left\|\nabla \ell(\theta_k^{ga}) + \epsilon_k\right\|^2$$

$$+\frac{L}{2} \sum_{k=1}^{N} \frac{(\lambda_k - \beta_k)^2}{\Gamma_k \alpha_k} (\sum_{\tau=k}^{N} \Gamma_\tau) \left\|\nabla \ell(\theta_k^{ga}) + \epsilon_k\right\|^2 + \sum_{k=1}^{N} \lambda_k \left\|\epsilon_k\right\| \cdot \left\|\nabla \ell(\theta_k^{ga}) + \epsilon_k\right\|$$

$$= -\ell(\theta_0) - \sum_{k=1}^{N} \lambda_k C_k \left\|\nabla \ell(\theta_k^{ga}) + \epsilon_k\right\|^2 + \sum_{k=1}^{N} \lambda_k \left\|\epsilon_k\right\| \cdot \left\|\nabla \ell(\theta_k^{ga}) + \epsilon_k\right\| \tag{25}$$

Rearranging the terms in the above inequality

$$\sum_{k=1}^{N} \lambda_k C_k \left\|\nabla \ell(\theta_k^{ga}) + \epsilon_k\right\|^2 \leq \ell(\theta_N) - \ell(\theta_0) + \sum_{k=1}^{N} \lambda_k \left\|\epsilon_k\right\| \cdot \left\|\nabla \ell(\theta_k^{ga}) + \epsilon_k\right\|$$

By assumption 6 that $\|\nabla \ell(\cdot)\|$ and $\sum_{k=1}^{N} \lambda_k \|\epsilon_k\|$ are bounded. Since $\ell(\theta_N) \leq \ell(\theta^*)$ and in view of the assumption that $C_k > 0$, we obtain for some constant B,

$$\min_{k=1,.N} \|\nabla \ell(\theta_k^{ga}) + \epsilon_k\|^2 \leq \frac{\ell(\theta^*) - \ell(\theta_0) + B}{\sum_{k=1}^{N} \lambda_k C_k}$$

which clearly implies (9).

We now prove part b).

First, from L-Lipschitz-continuous gradient property (6), we have

$$-\ell(\theta_k^{ag}) \leq -\ell(\theta_k^{ga}) + \left\langle \nabla \ell(\theta_k^{ga}),\ \theta_k^{ag} - \theta_k^{ga} \right\rangle + \frac{L}{2} \left\|\theta_k^{ag} - \theta_k^{ga}\right\|^2$$

$$\leq -\ell(\theta_k^{ga}) - \beta_k \left\|\nabla \ell(\theta_k^{ga}) + \epsilon_k\right\|^2 + \left\|\epsilon_k\right\| \beta_k \left\|\nabla \ell(\theta_k^{ga}) + \epsilon_k\right\| + \frac{L\beta_k^2}{2} \left\|\nabla \ell(\theta_k^{ga}) + \epsilon_k\right\|^2. \tag{26}$$

By the assumption that $\ell(\cdot)$ is concave and (4),

$$-\ell(\theta_k^{ga}) + \left[(1 - \alpha_k)\ell(\theta_{k-1}^{ag}) + \alpha_k \ell(\theta)\right]$$

$$= \alpha_k \left[-\ell(\theta_k^{ga}) + \ell(\theta)\right] + (1 - \alpha_k) \left[-\ell(\theta_k^{ga}) + \ell(\theta_{k-1}^{ag})\right]$$

$$\leq \alpha_k \left\langle \nabla \ell(\theta_k^{ga}),\ \theta_k^{ga} - \theta \right\rangle + (1 - \alpha_k) \left\langle \nabla \ell(\theta_k^{ga}),\ \theta_k^{ga} - \theta_{k-1}^{ag} \right\rangle$$

$$= \langle \nabla \ell(\theta_k^{ga}),\ \alpha_k(\theta_k^{ga} - \theta) + (1 - \alpha_k)(\theta_k^{ga} - \theta_{k-1}^{ag}) \rangle$$

$$= \alpha_k \left\langle \nabla \ell(\theta_k^{ga}),\ \theta_{k-1} - \theta \right\rangle. \tag{27}$$

From (5), we have

$$\|\theta_k - \theta\|^2 = \left\|\theta_{k-1} + \lambda_k \widehat{\nabla \ell(\theta_k^{ga})} - \theta\right\|^2$$

$$= \|\theta_{k-1} - \theta\|^2 - 2\lambda_k \langle \widehat{\nabla \ell(\theta_k^{ga})},\ \theta_{k-1} - \theta \rangle + \lambda_k^2 \left\|\widehat{\nabla \ell(\theta_k^{ga})}\right\|^2,$$

$$= \|\theta_{k-1} - \theta\|^2 - 2\lambda_k \langle \nabla \ell(\theta_k^{ga}) + \epsilon_k,\ \theta_{k-1} - \theta \rangle + \lambda_k^2 \left\|\nabla \ell(\theta_k^{ga}) + \epsilon_k\right\|^2,$$

which implies

$$\alpha_k \left\langle \nabla \ell(\theta_k^{ga}) + \epsilon_k,\ \theta_{k-1} - \theta \right\rangle = \frac{\alpha_k}{2\lambda_k} \left[\|\theta_{k-1} - \theta\|^2 - \|\theta_k - \theta\|^2 \right]$$

$$+ \frac{\alpha_k \lambda_k}{2} \left\| \nabla \ell(\theta_k^{ga}) + \epsilon_k \right\|^2.$$

Hence we obtain

$$\alpha_k \left\langle \nabla \ell(\theta_k^{ga}),\ \theta_{k-1} - \theta \right\rangle \leq \frac{\alpha_k}{2\lambda_k} \left[\|\theta_{k-1} - \theta\|^2 - \|\theta_k - \theta\|^2 \right]$$

$$+ \frac{\alpha_k \lambda_k}{2} \left\| \nabla \ell(\theta_k^{ga}) + \epsilon_k \right\|^2 + \alpha_k \|\epsilon_k\| \|\theta_{k-1} - \theta\| \tag{28}$$

Using the results of (26), (27), and (28), we get

$$-\ell(\theta_k^{ag}) \leq -(1-\alpha_k)\ell(\theta_{k-1}^{ag}) - \alpha_k \ell(\theta) + \frac{\alpha_k}{2\lambda_k} \left[\|\theta_{k-1} - \theta\|^2 - \|\theta_k - \theta\|^2 \right] + \alpha_k \|\epsilon_k\| \|\theta_{k-1} - \theta\|$$

$$-\beta_k (1 - \frac{L\beta_k}{2} - \frac{\alpha_k \lambda_k}{2\beta_k}) \left\| \nabla \ell(\theta_k^{ga}) + \epsilon_k \right\|^2 + \|\epsilon_k\| \beta_k \left\| \nabla \ell(\theta_k^{ga}) + \epsilon_k \right\|$$

$$\leq -(1 - \alpha_k)\ell(\theta_{k-1}^{ag}) - \alpha_k \ell(\theta) + \frac{\alpha_k}{2\lambda_k} \left[\|\theta_{k-1} - \theta\|^2 - \|\theta_k - \theta\|^2 \right]$$

$$-\frac{\beta_k}{2}(1 - L\beta_k) \left\| \nabla \ell(\theta_k^{ga}) + \epsilon_k \right\|^2 + \|\epsilon_k\| \beta_k \left\| \nabla \ell(\theta_k^{ga}) + \epsilon_k \right\| + \alpha_k \|\epsilon_k\| \|\theta_{k-1} - \theta\|, \tag{29}$$

where the last inequality follows from the assumption in (10). Subtracting $\ell(\theta)$ from both sides of the above inequality and using Lemma 1, we conclude that

$$-\ell(\theta_N^{ag}) + \ell(\theta) \leq \Gamma_N \left[\sum_{k=1}^{N} \frac{\alpha_k}{2\lambda_k \Gamma_k} \left[\|\theta_{k-1} - \theta\|^2 - \|\theta_k - \theta\|^2 \right] \right.$$

$$\left. - \sum_{k=1}^{N} \frac{\beta_k}{2\Gamma_k}(1 - L\beta_k) \left\| \nabla \ell(\theta_k^{ga}) + \epsilon_k \right\|^2 + \sum_{k=1}^{N} \frac{1}{\Gamma_k} \left[\|\epsilon_k\| \beta_k \left\| \nabla \ell(\theta_k^{ga}) + \epsilon_k \right\| + \alpha_k \|\epsilon_k\| \|\theta_{k-1} - \theta\| \right] \right]$$

$$\leq \Gamma_N \frac{\|\theta_0 - \theta\|^2}{2\lambda_1} - \Gamma_N \sum_{k=1}^{N} \frac{\beta_k}{2\Gamma_k} (1 - L\beta_k) \left\| \nabla \ell(\theta_k^{ga}) + \epsilon_k \right\|^2$$

$$+ \Gamma_N \sum_{k=1}^{N} \frac{1}{\Gamma_k} \left[\|\epsilon_k\| \beta_k \left\| \nabla \ell(\theta_k^{ga}) + \epsilon_k \right\| + \alpha_k \|\epsilon_k\| \|\theta_{k-1} - \theta\| \right] \tag{30}$$

for every $\theta \in \mathbb{R}^n$. By our construction (11) that sequence $\left\{ \frac{\alpha_k}{\lambda_k \Gamma_k} \right\}$is decreasing and the fact that $\alpha_1 = \Gamma_1 = 1$, we have

$$\sum_{k=1}^{N} \frac{\alpha_k}{\lambda_k \Gamma_k} \left[\|\theta_{k-1} - \theta\|^2 - \|\theta_k - \theta\|^2 \right] \leq \frac{\alpha_1 \|\theta_0 - \theta\|^2}{\lambda_1 \Gamma_1} = \frac{\|\theta_0 - \theta\|^2}{\lambda_1} \tag{31}$$

which immediately implies the last inequality of (30).

Hence, we can conclude (13) from the above inequality and the assumption in (10):

$$\ell(\theta^*) - \ell(\theta_N^{ag}) \leq \Gamma_N \left[\frac{\|\theta_0 - \theta^*\|^2}{2\lambda_1} + \sum_{k=1}^{N} \Gamma_k^{-1} \left[\|\epsilon_k\| \beta_k \left\| \nabla \ell(\theta_k^{ga}) + \epsilon_k \right\| + \alpha_k \|\epsilon_k\| \|\theta_{k-1} - \theta\| \right] \right]$$

Finally, noting the fact that $\ell(\theta_N^{ag}) \leq \ell(\theta^*)$, substitute $\theta := \theta^*$, re-arranging the terms in (30) we obtain

$$\sum_{k=1}^{N} \frac{\beta_k}{2\Gamma_k}(1 - L\beta_k) \left\| \nabla \ell(\theta_k^{ga}) + \epsilon_k \right\|^2 \quad k = 1, \ldots, N$$

$$\leq \frac{\|\theta^* - \theta_0\|^2}{2\lambda_1} + \sum_{k=1}^{N} \frac{1}{\Gamma_k} \left[\|\epsilon_k\| \beta_k \left\| \nabla \ell(\theta_k^{ga}) + \epsilon_k \right\| + \alpha_k \|\epsilon_k\| \|\theta_{k-1} - \theta\| \right],$$

or

$$\min_{k=1,.N} \|\nabla \ell(\theta_k^{ga}) + \epsilon_k\|^2 \leq 2 \frac{\frac{\|\theta^* - \theta_0\|^2}{2\lambda_1} + \sum_{k=1}^{N} \frac{1}{\Gamma_k} \left[\|\epsilon_k\| \beta_k \left\| \nabla \ell(\theta_k^{ga}) + \epsilon_k \right\| + \alpha_k \|\epsilon_k\| \|\theta_{k-1} - \theta\| \right]}{\sum_{k=1}^{N} \Gamma_k^{-1} \beta_k (1 - L\beta_k)}$$

which together with (10), clearly imply (12). □

A.2. Proof of Theorem 2

Proof. We first prove part a). Note that by choosing

$$\beta_k = \frac{1}{2L}$$

$$\Gamma_k = \frac{1}{k^{1+\delta}}, \tag{32}$$

which implies that for sufficient large k

$$\sum_{\tau=k}^{N} \Gamma_\tau = \sum_{\tau=k}^{N} \frac{1}{\tau^{1+\delta}} = O(\frac{1}{k^\delta})$$

We also have

$$1 - \alpha_k = \frac{(k-1)^{1+\delta}}{k^{1+\delta}} \tag{33}$$

for every $k > 1$, or $\alpha_k = \frac{\left(k^{1+\delta} - (k-1)^{1+\delta}\right)}{k^{1+\delta}} = O(\frac{(1+\delta)k^\delta}{k^{1+\delta}}) = O(\frac{1}{k})$. If we choose λ_k such that $\lambda_k - \beta_k = o(k^{-1})$ then

$$\frac{(\lambda_k - \beta_k)^2}{2\alpha_k \Gamma_k \lambda_k} (\sum_{\tau=k}^{N} \Gamma_\tau) = \frac{o(k^{-2})}{k^{-1} k^{-(1+\delta)}} \frac{1}{k^\delta} = o(1)$$

so for sufficiently large k we have

$$C_k = 1 - L[\lambda_k + \frac{(\lambda_k - \beta_k)^2}{2\alpha_k \Gamma_k \lambda_k} (\sum_{\tau=k}^{N} \Gamma_\tau)] > \frac{1}{4}$$

Hence, it can also be seen from (9) that for some positive bounded constant B_2,

$$\min_{k=1,.N} \left\| \nabla \ell(\theta_k^{ga}) + \epsilon_k \right\|^2 \leq \frac{\ell^* - \ell(\theta_0) + B}{N B_2} = O(\frac{1}{N}),$$

which concludes the first part of the proof. Since $\|\epsilon_k\| = O\left(\tau^2\right) \leq O(\frac{1}{k})$, we have $\nabla \ell(\theta_k^{ga})$ converge to 0 at the rate of

$$\min \left\{ O(\frac{1}{\sqrt{N}}), O\left(\|\epsilon_k\|\right) \right\} = O(\frac{1}{\sqrt{N}}),$$

which gives us the desired result.

We now show part b). Let $\lambda_k = \left(k^{1+\delta} - (k-1)^{1+\delta}\right) c$ for some constant c then

$$\frac{\alpha_1}{\lambda_1 \Gamma_1} = \frac{\alpha_2}{\lambda_2 \Gamma_2} = \cdots = \frac{\alpha_k}{\lambda_k \Gamma_k}.$$

Observe that

$$\alpha_k\lambda_k = \frac{c^2\left(k^{1+\delta} - (k-1)^{1+\delta}\right)^2}{k^{1+\delta}} = \frac{c^2\,(1+\delta)^2\,O(k^{2\delta})}{k^{1+\delta}} \to 0$$

for $\delta < 1$ so $\frac{(1+\delta)^2 k^{2\delta}}{k^{1+\delta}} < \beta_k = \frac{1}{2L}$ for sufficient large k, which implies that conditions (10) and (11) hold. Moreover, it can also be easily seen from (22) that

$$\min_{k=1,.N}\|\nabla\ell(\theta_k^{ga}) + \epsilon_k\|^2 \leq \frac{\frac{\|\theta^* - \theta_0\|^2}{2\lambda_1} + C\sum_{k=1}^N \Gamma_k^{-1}\left[\|\epsilon_k\| + O(\frac{1}{k})\,\|\epsilon_k\|\right]}{\sum_{k=1}^N \Gamma_k^{-1}} = O(N^{-2-\delta}).$$

The last equality is due to the fact that $\sum_{k=1}^N \Gamma_k^{-1} = \sum_{k=1}^N k^{(1+\delta)} = O(N^{2+\delta})$. Combining the above relation with (9), and since $\|\epsilon_k\| = O\left(\tau^2\right) \leq O(\frac{1}{k^{2+\delta+\delta_1}})$ for some $\delta_1 > 0$, we have $\nabla\ell(\theta_k^{ga})$ converge to 0 at the rate of $O\left(\sqrt{\frac{1}{N^{2+\delta}}}\right)$.

Since $\alpha_k\lambda_k < \beta_k = \frac{1}{2L}$, we have $\delta \leq 1$ which implies that the best convergence rate is $O(N^{-3})$. □

B. Additional results with panel model

Motivated by testing the new algorithm on a large and more complex model, we use dynamic variation in sexual contact rates model of Romero-Severson, Volz, Koopman, Leitner, and Ionides (2015) following closely the setup of (Ionides 2018) as a case study. Specifically, let $X_i(t)$ be a latent rate of making contacts of a specific type for each individual, and y_{ij} be the number of reported contacts for individual i between time t_{j-1} and t_j, where $i = 1, \ldots, 882$ and $j = 1, \ldots, 4$. Let C_{ij} be the expected number of contacts for individual i in reporting interval j, taking into account the decline in reported contacts, we have

$$C_{ij} = \alpha^{j-1}\int_{t_{j-1}}^{t_j} X_i(t)\,dt,$$

where α is a decline fraction. In addition, to account for the higher variance of the data, Ionides (2018) replaced the traditional Poisson distribution with the negative binomially distribution (Bretó *et al.* 2009), assuming

$$y_{ij} \sim \text{NegBin}\,(C_{ij}, D_i)\,.$$

When the dispersion D_i becomes large, negative binomially distribution becomes Poisson distribution with the same mean and variance C_{ij} in the limit. When the dispersion D_i is small, it can model the increased variance compared to the Poisson distribution for individual contacts. At the start of each episode, $X_i(t)$ is drawn from a Gamma distribution with mean μ_X and variance σ_X. To account for autocorrelation between measurements on an individual over time observed in the data, Ionides (2018) supposes that individual i has behavioral episodes within which $X_i(t)$ is constant, but the individual enters new behavioral episodes at a rate R_i.

$$X_i(t) \sim \text{Gamma}(\mu_X, \sigma_X),$$

but does not result in autocorrelation between measurements over time for an individual. Hence, Ionides (2018) supposes that individual i has behavioral episodes within which $X_i(t)$ is constant, but the individual enters new behavioral episodes at a rate R_i. Finally, D_i and R_i are also drawn from Gamma distributions,

$$D_i \sim \text{Gamma}(\mu_D, \sigma_D),$$

$$R_i \sim \text{Gamma}(\mu_R, \sigma_R),$$

where σ_X, σ_D and σ_R control individual-level differences in behavioral parameters, covering a wide range of sexual contact patterns.

The distinction between the effects of the rate at which the new behavioral episodes begin, R_i, and the dispersion parameter, D_i, is subtle since both model within-individual variability. As noted by Ionides (2018), R_i and D_i both model within-individual variability and identify them from data depending on the high variance in the number of reported contacts. We get the following results using IF1, IF2, and AIF to solve this empirical question (Table 2), which include a few parameters and the estimated likelihood. As seen from this large model, both IF2 and AIF can reach the ideal likelihood-ratio 95% confidence set of MLE ($-9552.01 \pm qchisq(0.95, df = 6)$) while IF1 is just barely outside it. This reconfirms our earlier comparison between IF1, IF2, and AIF.

Table 2: Summary results of fitting dynamic variation in sexual contact rates model using IF1, IF2, AIF with number of particle $J = 10000$ and number of iteration $M = 50$

Algorithms	μ_D	μ_R	α	$\hat{\ell}$	s.e.	time(s)
IF1	3.8399	0.0415	0.8995	-9558.3904	1.6769	4441.292
IF2	3.0624	0.0400	0.8972	-9554.1762	1.7467	4519.276
AIF	3.0112	0.0426	0.9169	-9554.0951	1.1100	4551.132

C. Parameters definitions and starting ranges for the malaria model

Table S-4. Parameters for the malaria $SEIH^3Q$ model.

Symbol	Definition	Units	θ_{low}	θ_{high}
μ_{EI} $(*)$	E $\rightarrow$ I transition rate	yr^{-1}	24	24
μ_{IH}	I $\rightarrow$ H transition rate	yr^{-1}	1.00	5.00
μ_{HI}	H $\rightarrow$ I transition rate	yr^{-1}	1.00	5.00
μ_{IS}	I $\rightarrow$ S transition rate	yr^{-1}	0.5	2.00
μ_{IQ}	I $\rightarrow$ Q transition rate	yr^{-1}	1.00	2.00
μ_{QS}	Q $\rightarrow$ S transition rate	yr^{-1}	10.00	20.00
q $(*)$	relative infectivity of Q class	–	0.001	0.001
τ	mean lag for mosquitoes	month	0.10	0.50
ρ	case reporting fraction	–	0.001	0.01
σ_{pro}	s.d. of dynamic noise	$\text{yr}^{0.5}$	0.1	0.5
σ_{obs}	s.d. of measurement noise	–	0.1	0.5
b_r	coefficient of rainfall covariate	–	0.5	0.9
S_0	initial fraction in S class	–	0	1
E_0	initial fraction in E class	–	0	1
I_0	initial fraction in I class	–	0	1
$H_{i,0}$	initial fraction in H_i class	–	0	1
Q_0	initial fraction in Q class	–	0	1
κ_0	initial value, $\kappa(t_0)$	–	0.1	0.5
$\mu_{SE,0}$	initial value, $\mu_{SE}(t_0)$	–	0.1	0.5
b_1	1^{st} spline coefficient	–	-5	5
b_2	2^{nd} spline coefficient	–	-5	5
b_3	3^{rd} spline coefficient	–	-5	5
b_4	4^{th} spline coefficient	–	-5	5
b_5	5^{th} spline coefficient	–	-5	5
b_6	6^{th} spline coefficient	–	-5	5
$1/\delta$ $(*)$	mean human life span	yr	0.02	0.02

We follow definitions as in Roy *et al.* (2013). θ_{low} and θ_{high} are the lower and upper bounds for a hyper-rectangle used to generate starting points for the search. Parameters labeled with $(*)$ were set at fixed values. Non-negative parameters were logarithmically transformed for optimization.

Acknowledgements

This research is funded in part by the University of Mississippi Summer Grant.

References

Andrieu C, Doucet A, Holenstein R (2010). "Particle Markov chain Monte Carlo methods." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **72**(3), 269–342. URL https://doi.org/10.1111/j.1467-9868.2009.00736.x.

Beaumont MA, Zhang W, Balding DJ (2002). "Approximate Bayesian computation in population genetics." *Genetics*, **162**(4), 2025–2035. URL https://doi.org/10.1093/genetics/162.4.2025.

Bretó C, He D, Ionides EL, King AA (2009). "Time series analysis via mechanistic models." *Annals of Applied Statistics*, **3**, 319–348. URL https://doi.org/10.1214/08-AOAS201.

Chopin N, Jacob PE, Papaspiliopoulos O (2013). "SMC^2: an efficient algorithm for sequential analysis of state space models." *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, **75**(3), 397–426. ISSN 1467-9868. URL http://dx.doi.org/10.1111/j.1467-9868.2012.01046.x.

Dahlin J, Lindsten F, Schön TB (2015). "Particle Metropolis-Hastings using gradient and Hessian information." *Statistics and Computing*, **25**(1), 81–92. URL https://doi.org/10.1007/s11222-014-9510-0.

Del Moral P, Doucet A, Singh S (2010). "Forward smoothing using sequential Monte Carlo." *arXiv preprint arXiv:1012.5390*. URL https://doi.org/10.48550/arXiv.1012.5390.

Doucet A, Godsill S, Andrieu C (2000). "On sequential Monte Carlo sampling methods for Bayesian filtering." *Statistics and Computing*, **10**(3), 197–208. URL https://doi.org/10.1023/A:1008935410038.

Doucet A, Jacob PE, Rubenthaler S (2013). "Derivative-free Estimation of the Score Vector and Observed Information Matrix with Application to State-Space Models." *ArXiv:1304.5768*. URL https://doi.org/10.48550/arXiv.1304.5768.

Ellner SP, Bailey BA, Bobashev GV, Gallant AR, Grenfell BT, Nychka DW (1998). "Noise and Nonlinearity in Measles Epidemics: Combining Mechanistic and Statistical Approaches to Population Modeling." *American Naturalist*, **151**(5), 425–440. doi:10.1086/286130.

Ghadimi S, Lan G (2016). "Accelerated gradient methods for nonconvex nonlinear and stochastic programming." *Mathematical Programming*, **156**(1-2), 59–99. URL https://doi.org/10.1007/s10107-015-0871-8.

Gordon NJ, Salmond DJ, Smith AF (1993). "Novel approach to nonlinear/non-Gaussian Bayesian state estimation." In *IEE Proceedings F (Radar and Signal Processing)*, volume 140, pp. 107–113. IET. URL https://doi.org/10.1049/ip-f-2.1993.0015.

He D, Ionides EL, King AA (2010). "Plug-and-play inference for disease dynamics: Measles in large and small populations as a case study." *Journal of the Royal Society Interface*, **7**(43), 271–283. URL https://doi.org/10.1098/rsif.2009.0151.

Huys QJ, Paninski L (2006). "Model-based optimal interpolation and filtering for noisy, intermittent biophysical recordings." In *Fifteenth Annual Computational Neuroscience Meeting*, volume 10. URL https://doi.org/10.1371/journal.pcbi.1000379.

Ionides EL (2018). "Case study: dynamic variation in sexual contact rates." *https://kingaa.github.io/sbied/contacts/contacts.html*. URL https://doi.org/10.1093/aje/kwv044.

Ionides EL, Bhadra A, Atchadé Y, King A (2011). "Iterated Filtering." *Annals of Statistics*, **39**, 1776–1802. URL https://doi.org/10.1214/11-AOS886.

Ionides EL, Bretó C, King AA (2006). "Inference for nonlinear dynamical systems." *Proceedings of the National Academy of Sciences of the USA*, **103**, 18438–18443. URL https://doi.org/10.1073/pnas.0603181103.

Ionides EL, Nguyen D, Atchadé Y, Stoev S, King AA (2015). "Inference for dynamic and latent variable models via iterated, perturbed Bayes maps." *Proceedings of the National Academy of Sciences of the USA*, **112**(3), 719–724. URL https://doi.org/10.1073/pnas.1410597112.

Kevrekidis IG, Gear CW, Hummer G (2004). "Equation-Free: The Computer-Assisted Analysis of Complex, Multiscale Systems." *American Institute of Chemical Engineers Journal*, **50**, 1346–1354. URL https://doi.org/10.1002/aic.10106.

Kiefer J, Wolfowitz J (1952). "Stochastic Estimation of the Maximum of a Regression Function." pp. 462–466. Annals of Mathematical Statistics, 23:, URL https://doi.org/10.1214/aoms/1177729392.

King AA, Nguyen D, Ionides EL (2016). "Statistical inference for partially observed Markov processes via the R package pomp." *Journal of Statistical Software*, **69**(12). URL https://doi.org/10.18637/jss.v069.i12.

Kloeden PE, Platen E (1999). *Numerical Soluion of Stochastic Differential Equations*. 3rd edition. Springer, New York. URL https://doi.org/10.1007/978-3-662-12616-5.

Kushner HJ, Clark DS (1978). *Stochastic Approximation Methods for Constrained and Unconstrained Systems*. Springer-Verlag, New York. URL https://doi.org/10.1007/978-1-4684-9352-8.

Lele SR, 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**(7), 551–563. ISSN 1461-0248. URL http://dx.doi.org/10.1111/j.1461-0248.2007.01047.x.

Lindström E, Ionides EL, Frydendall J, Madsen H (2012). "Efficient Iterated Filtering." In *16th IFAC Symposium on System Identification*. URL https://doi.org/10.3182/20120711-3-BE-2027.00300.

Nemeth C, Fearnhead P, Mihaylova L (2013). "Particle approximations of the score and observed information matrix for parameter estimation in state space models with linear computational cost." *ArXiv:1306.0735*. URL https://doi.org/10.1080/10618600.2015.1093492.

Nesterov Y (2005). "Smooth minimization of non-smooth functions." *Mathematical Programming*, **103**(1), 127–152. URL https://doi.org/10.1007/s10107-004-0552-5.

Nesterov Y (2013). *Introductory Lectures on Convex Optimization: A Basic Course*, volume 87. Springer Science & Business Media. URL https://doi.org/10.1007/s10107-004-0552-5.

Nguyen D, Ionides EL (2017). "A second-order iterated smoothing algorithm." *Statistical Computing*. URL https://doi.org/10.1007/s11222-016-9711-9.

Poyiadjis G, Doucet A, Singh SS (2011). "Particle approximations of the score and observed information matrix in state space models with application to parameter estimation." *Biometrika*, **98**(1), 65–80. URL https://doi.org/10.1093/biomet/asq062.

Ricker WE (1954). "Stock and Recruitment." *Journal of the Fisheries Research Board of Canada*, **11**, 559–623. URL https://doi.org/10.1139/f54-039.

Romero-Severson E, Volz E, Koopman J, Leitner T, Ionides E (2015). "Dynamic variation in sexual contact rates in a cohort of HIV-negative gay men." *American journal of epidemiology*, **182**(3), 255–262. URL https://doi.org/10.1093/aje/kwv044.

Roy M, Bouma MJ, Ionides EL, Dhiman RC, Pascual M (2013). "The potential elimination of plasmodium vivax malaria by relapse treatment: Insights from a transmission model and surveillance data from NW India." *PLoS Neglected Tropical Diseases*, **7**(1), e1979. URL https://doi.org/10.1371/journal.pntd.0001979.

Sisson SA, Fan Y, Tanaka MM (2007). "Sequential Monte Carlo without likelihoods." *Proceedings of the National Academy of Sciences of the USA*, **104**(6), 1760–1765. URL https://doi.org/10.1073/pnas.0607208104.

Toni T, Welch D, Strelkowa N, Ipsen A, Stumpf MPH (2009). "Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems." *Journal of the Royal Society Interface*, **6**, 187–202. URL https://doi.org/10.1098/rsif.2008.0172.

Wood SN (2010). "Statistical inference for noisy nonlinear ecological dynamic systems." *Nature*, **466**(7310), 1102–1104. URL https://doi.org/10.1038/nature09319.

Yıldırım S, Singh SS, Dean T, Jasra A (2015). "Parameter estimation in hidden Markov models with intractable likelihoods using sequential Monte Carlo." *Journal of Computational and Graphical Statistics*, **24**, 846–865. URL https://doi.org/10.1080/10618600.2014.938811.

Affiliation:

Dao Nguyen
Department of Mathematics
University of Mississippi
University, Mississippi
E-mail: dxnguyen@olemiss.edu
URL: https://math.olemiss.edu/dao-nguyen/

http://www.ajs.or.at/
http://www.osg.or.at/
Submitted: 2022-03-13
Accepted: 2022-07-26