Bayesian identification of state-space models via adaptive thermostats

Jack Umenberger, Thomas B. Schön, and Fredrik Lindsten

***Abstract*— Bayesian modeling has been recognized as a powerful approach to system identification, not least due to its intrinsic uncertainty quantification. However, despite many recent developments, Bayesian identification of nonlinear state space models still poses major computational challenges. We propose a new method to tackle this problem. The technique is based on simulating a so-called *thermostat*, a stochastic differential equation constructed to have the posterior parameter distribution as its limiting distribution. Simulating the thermostat requires access to *unbiased* estimates of the gradient of the log-posterior. To handle this, we make use of a recent method for debiasing particle-filter-based smoothing estimates. Numerical results show a clear benefit of this approach compared to a direct application of (biased) particle-filter-based gradient estimates within the thermostat.**

I. Introduction

Data-driven modeling of dynamic phenomena has long been a task of central importance in many areas of science and engineering. In many applications, especially those involving decision making, quantifying the accuracy of these models is essential. Bayesian inference offers a principled approach to uncertainty quantification, especially in the 'finite-data' regime, where asymptotic variance estimates from prediction error methods may not be applicable [1]. Indeed, since the early 1980s [2], Bayesian inference has been recognized as a valuable paradigm for uncertainty quantification in system identification.

For many models of interest, the requisite computations for exact Bayesian inference cannot be carried out in closed form. Among the methods for approximate inference, Markov chain Monte Carlo (MCMC), including the venerable Metropolis-Hastings algorithm, has gained immense popularity. The key to efficient application of MCMC is a proposal distribution that allows the Markov chain to rapidly reach the stationary distribution. However, designing such a proposal is a formidable task, especially in high dimensions.

Many strategies for designing effective proposals make use of the gradient of the log-posterior most notably Hamiltonian MC [3]. In applications where only stochastic approximations to the gradient are available (or desirable), a class of sampling methods known as *thermostats* have recently gained attention [4], [5]. A thermostat is a stochastic differential equation (SDE) that employs feedback control to compensate for noisy gradients; cf. §II-B. Though capable of handling noise, thermostats require *unbiased* estimates of the log-posterior gradient. For nonlinear state-space models, gradient computations based on particle approximations of the Fisher identity lead to biased estimates; cf. §III-A, and as such, are incompatible with thermostats.

In this paper, we leverage recent methods [6] for debiasing expectations with respect to particle approximations of smoothing distributions, including gradient computations, to enable the application of thermostats to the problem of Bayesian identification of nonlinear state-space models.

A. Related work

Using *dynamics* to construct efficient proposals for MCMC has a long tradition. Hamiltonian MC (HMC) drew inspiration from statistical physics [7], before being repurposed for Bayesian inference [3]. It has been recognized as one of the most powerful approaches to Monte Carlo sampling, in particular for high-dimensional problems. In the era of 'big data', there has been considerable interest in extending such methods to handle stochastic gradients, which typically arise when datasets are subsampled for computational tractability. In [8], Langevin (i.e. Brownian) dynamics driven by stochastic gradients was shown to sample from the desired distribution, provided that the sequence of stepsizes decreases to zero; cf. also [9]. Other strategies attempt to estimate the variance of the stochastic gradient and compensate by adjusting the magnitude of the artificially injected noise accordingly [10], [11]. Such methods are sensitive to errors in the variance estimate [4], which motivates the use of *thermostats* [12] to compensate for the gradient noise via feedback control, cf. [4], [5].

When it comes to Bayesian identification of nonlinear state-space models, we not only need a way to sample the model parameters, but also to marginalize out (or sample) the system states. A principled way of accomplishing this is offered by the family of methods known as particle MCMC [13], in which MCMC sampling is used for the model parameters while at the same time sequential Monte Carlo (SMC) is used to 'handle' the unknown states.

For instance, particle Metropolis-Hastings (PMH [13], [14]) uses SMC to marginalize the system states, effectively targeting the marginal parameter posterior distribution. A challenge with PMH, however, is the design of effective proposal distributions [15], i.e., those that achieve fast mixing (convergence to the stationary distribution). This is particularly challenging when the model parameter is high-

This work has been supported by the Swedish Research Council (VR) via the projects *NewLEADS – New Directions in Learning Dynamical Systems* (contract number: 621-2016-06079) and *Learning of Large-Scale Probabilistic Dynamical Models* (contract number: 2016-04278), and by the Swedish Foundation for Strategic Research (SSF) via the projects *ASSEMBLE* (contract number: RIT15-0012) and *Probabilistic Modeling and Inference for Machine Learning* (contract number: ICA16-0015).
J. Umenberger F. Lindsten and T.B. Schön are with the Dept. of Information Technology, Uppsala University, 751 05 Uppsala, Sweden. E-mails: {jack.umenberger, fredrik.lindsten, thomas.schon}@it.uu.se.

dimensional. Using, e.g., HMC in the context of PMH is far from straightforward (see, e.g., [16] for a discussion).

An alternative type of PMCMC is particle Gibbs [13], [17], [18], in which SMC is used to simulate the states alternately with the parameters using a Gibbs scheme. These methods can however suffer from slow convergence when there are strong dependencies between states and parameters.

In this work we shall circumvent these limitations of PMCMC by instead using a thermostat which is able to exploit gradient information, akin to HMC (as explained above). This is enabled by using SMC to compute *unbiased* estimates of the gradient of the log-posterior, thereby resulting in a combination of Markov chain sampling and SMC which is distinct from the family of PMCMC methods.

B. *Intuitive preview of main idea*

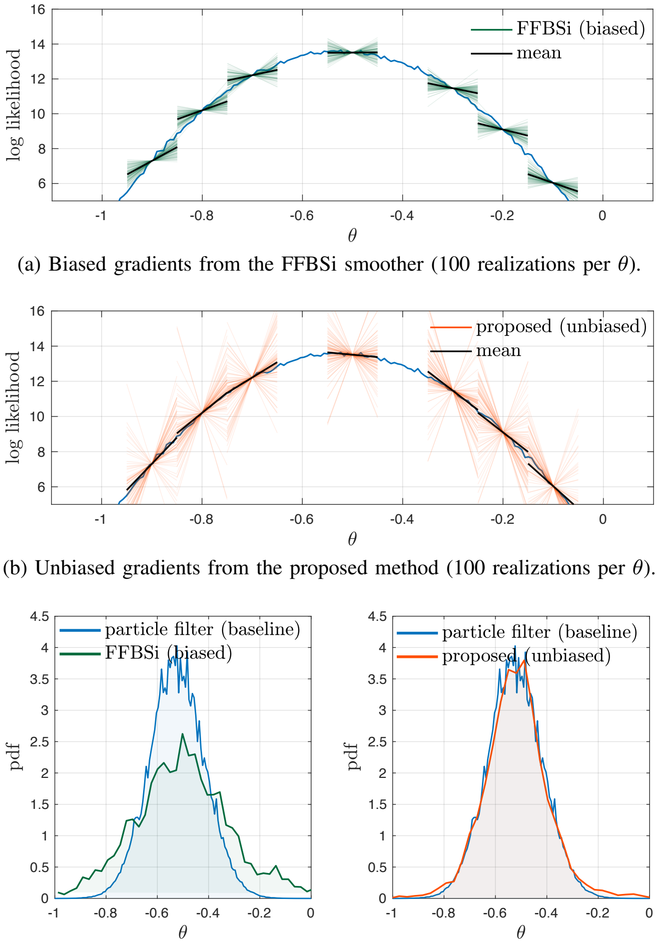
The stochastic gradient methods discussed above, including thermostats, all assume access to unbiased gradients, which is the case when stochasticity is due to subsampling i.i.d. data. For state-space models, stochasticity is inherent. Moreover, conventional methods for computing gradients, based on smoothing expectations and the Fisher identity (cf. §III-A), lead to biased estimates. The key idea in this work is to leverage recently developed debiasing techniques, cf. [19] in particular and also [6], [20], [21], [22], to generate unbiased estimates of the gradient of the log-likelihood function. These estimates can be 'plugged-in' to thermostats to create an SDE, trajectories of which correspond to samples from the parameter posterior distribution. We now briefly illustrate this key idea with a numerical example. Consider the nonlinear state-space model given by

$$x_{t+1} = \theta \frac{x_t^3 - x_t}{1 + x_t^2} + u_t + w_t, \qquad y_t = x_t^2 + v_t, \qquad (1)$$

where w_t, v_t, x_0 are normally distributed according to $\mathcal{N}\left(0, 0.1^2\right)$. Here, $\theta \in \mathbb{R}$ is the single unknown parameter of interest, with true value $\theta = 0.5$.

Given observed data y and u our goal is to draw samples from the parameter posterior distribution, using a thermostat. When *biased* estimates of the gradient (of the log-likelihood) are supplied to the thermostat, cf. Fig. 1(a), the resulting samples fail to target the correct distribution, cf. Fig 1(c). Indeed, the systematic bias in the gradient clearly overestimates the posterior variance. In contrast, when *unbiased* gradients using the proposed procedure are used, cf. Fig. 1(b), the target distribution is left invariant, cf. Fig 1(c).

The rest of the paper proceeds as follows: a precise problem statement is given in §II-A, before a brief introduction to thermostats for Bayesian inference in §II-B. The method to debias gradients is detailed in §III, and the final algorithm in presented in §IV. An additional numerical example, namely, identification of a Wiener system, is presented in §V.



(a) Biased gradients from the FFBSi smoother (100 realizations per θ).

(b) Unbiased gradients from the proposed method (100 realizations per θ).

(c) Posterior distributions from thermostat using the gradients above.

Fig. 1: Biased gradients lead to incorrect posteriors; cf. §I-B for details.

II. Preliminaries

A. *Problem statement*

This paper is concerned with identification of probabilistic state-space models of the form

$$x_t | x_{t-1}, u_t, \theta \sim f(x_t | x_{t-1}, u_t, \theta), \qquad x_0 \sim \mu(x_0, \theta). \qquad (2a)$$

$$y_t | x_t, u_t, \theta \sim g(y_t | x_t, u_t, \theta), \qquad \theta \sim p(\theta). \qquad (2b)$$

Here x_t denotes the latent (unobserved) state, u_t and y_t denote the observed inputs and outputs, respectively, and θ denotes the unknown model parameters, with prior distribution $p(\theta)$. For brevity, we shall drop explicit dependence on u_t from the notation, as is customary. The model (2) describes a very broad class of dynamical systems; in this paper we shall impose a number of further assumptions. For the purpose of applying SMC we shall assume that $g()$ can be evaluated point-wise. One way to satisfy this requirement is to assume additive measurement noise v_t, i.e.,

$$y_t = g(x_t, \theta) + v_t \implies g(y_t | x_t, \theta) = p_v(y_t - g(x_t, \theta)) \quad (3)$$

where $v \sim p_v(v)$. We also make the non-restrictive assumption that we have access to deterministic functions M and F, as well as the ability to sample random variables U_t for

$t \geq 0$, such that samples from $\mu(x_0, \theta)$ and $f(x_t|x_{t-1}, \theta)$ can be generated by evaluating $M(U_0)$ and $F(x_{t-1}, U_t, \theta)$, respectively.

Given observed data $y = \{y_t\}_{t=1}^T$ (and $u = \{u_t\}_{t=1}^T$) our goal is to approximate the parameter posterior distribution

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{\int p(y|\theta)p(\theta)d\theta}, \text{ where } \ p(y|\theta) = \int p(x, y|\theta)dx.$$

These integrals are, in general, intractable. As an alternative to exact computation, we instead wish to draw samples $\theta \sim p(\theta|y)$ from the posterior. Many strategies for parameter posterior sampling have been proposed, cf. §I-A. In this work, we adopt a MCMC approach based on adaptive thermostats, which we review next.

B. Adaptive thermostats for Bayesian inference

In this subsection, we provide a brief introduction to adaptive thermostats for Bayesian inference (cf. §I-A for references). The methods were developed in the field of statistical physics for simulating molecular dynamics, and we shall explain the key ideas in this context. Consider a physical system, in thermal equilibrium with a heat bath at constant temperature $\mathcal{T}$. The position(s) $q \in \mathbb{R}^d$ and momentum(s) $\rho \in \mathbb{R}^d$ of the molecules constitute the state of the system. In thermal equilibrium, these states are distributed according to the *canonical* distribution $p_\beta(q, \rho) \propto \exp(\beta^{-1} H(q, \rho))$, where $H(q, \rho) = U(q) + K(\rho)$ denotes the Hamiltonian with potential energy U and kinetic energy $K = \frac{1}{2}\rho' \mathcal{M}^{-1}\rho$, and β is proportional to the reciprocal temperature. Here $\mathcal{M} = mI$, where m denotes the mass of each molecule. It has been shown [23] that the following (second order) Langevin dynamics

$$dq = \mathcal{M}^{-1}\rho \ dt, \tag{4a}$$

$$d\rho = -\nabla U(q) \ dt - \gamma\rho dt + \sqrt{2\beta^{-1}\gamma}\mathcal{M}^{1/2}dW, \tag{4b}$$

ergodically samples the canonical distribution $p_\beta(q, \rho)$. Here, γ is a 'friction constant' for the damping due to interaction with the heat bath, and dW is a vector of Wiener increments.

The key point is: the stochastic dynamical system (4) generates trajectories q that represent samples from the distribution $\bar{p}_\beta(q) \propto \exp(-\beta^{-1}U(q))$. From the perspective of statistical physics, a given potential energy U leads to a particular canonical distribution $\bar{p}_\beta$. For the purpose of statistical *inference*, given a probability distribution of interest (a 'target distribution') $\phi(\theta)$, one can let q denote θ and choose $U(q) \propto -\log\phi(q)$ such that the canonical distribution $\bar{p}_\beta$ conincides with ϕ.

If the potential gradient ∇U is replaced with a noisy approximation $\nabla\tilde{U}$, (4) will no longer sample from the canonical distribution. The canonical ensemble satisfies the thermal equilibrium condition $k_B\mathcal{T} = \frac{2}{d}\mathbb{E}[K(\rho)]$, where k_B is the Boltzmann constant. Roughly speaking, the noise associated with $\nabla\tilde{U}$ constitutes energy perturbations that disrupt the requisite thermal equilibrium. To continue to sample from p_β, it is necessary to correct for these perturbations by regulating the temperature. One possibility is to introduce feedback control in the form of a *thermostat*, which leads to the augmented system

$$dq = \mathcal{M}^{-1}\rho \ dt, \tag{5a}$$

$$d\rho = -\nabla\tilde{U}(q) \ dt - \zeta\rho dt + \sigma_a\mathcal{M}^{1/2}dW, \tag{5b}$$

$$d\zeta = k_i\left(\rho'\mathcal{M}^{-1}\rho - dk_B\mathcal{T}\right) \ dt. \tag{5c}$$

One can interpret ζ as a variable friction coefficient, adapted via (integral) feedback control (with gain k_i) to regulate the temperature (kinetic energy) of the system (5c), e.g., when K is too large, ζ will grow, increasing the viscous damping so as to reduce ρ. Injection of the artificial noise $\sigma_a\mathcal{M}^{1/2}dW$ enables (5) to be related to a standard Itō stochastic differential equation [5].

III. Unbiased gradient approximation

The adaptive-thermostat method described in §II-B assumes access to *unbiased* stochastic estimates of the gradient ∇U. For the general class of state-space methods considered in this paper, methods to approximate the gradient of the log-likelihood yield biased estimates, cf. §III-A. In this section, we present a method for generating unbiased gradient estimates, based on recently developed Rhee-Glynn debiasing techniques.

A. Biased gradients from the Fisher identity

To apply the adaptive-thermostat (5) to the problem of posterior sampling for state-space models, one must choose $U(\theta) \propto -\log p(\theta|y)$. From Bayes' rule, a natural choice is $U(\theta) = -\log p(y|\theta)p(\theta)$. Let $\ell(\theta) := \log p(y|\theta)$ denote the log-likelihood. The gradient of $\ell(\theta)$ is given by the Fisher identity [24],

$$\nabla_\theta\ell(\theta) = \int \nabla_\theta \log p(x, y|\theta)p(x|y, \theta)dx, \tag{6}$$

where $p(x|y, \theta)$ denotes the so-called *smoothing* distribution. With the exception of the linear, Gaussian case, for models of the form (2) the smoothing distribution is not available in closed-form, and thus (6) must be approximated, e.g., by Monte Carlo methods:

$$\nabla_\theta\ell(\theta) \approx \frac{1}{M}\sum\nolimits_{i=1}^{M} \nabla_\theta \log p(\tilde{x}^i, y|\theta), \tag{7}$$

where $\tilde{x}^i \overset{\text{approx.}}{\sim} p(x|y, \theta)$ denote *approximate* samples from the smoothing distribution, generated by, e.g., a particle smoother [25], [26]. The stochastic gradient in (7) gives a *biased* approximation of $\nabla_\theta\ell(\theta)$, and is therefore unsuitable for use in the adaptive-thermostat; cf. Fig 1 for an illustration.

B. Debiased gradients

To obtain unbiased estimates of $\nabla_\theta\ell(\theta)$ we employ the so-called Rhee-Glynn debiasing technique [6]. Given a functional $h(x)$ and smoothing distribution $\pi(x)$ the purpose of the procedure is to produce unbiased estimates of the expectation $\int h(x)\pi(x)dx =: \pi(h)$. To approximate $\nabla_\theta\ell(\theta)$ via the Fisher identity (6), we choose $h(x) = \nabla_\theta \log p(x, y|\theta)$ and $\pi = p(x|y, \theta)$. The debiasing method proceeds as follows: Suppose we have two Markov chains, $(X^{(n)})_{n\geq 0}$

and $(\tilde{X}^{(n)})_{n\geq 0}$ with the following properties: i) each has invariant distribution π, ii) for all n, $X^{(n)}$ and $\tilde{X}^{(n)}$ have the same marginal distribution, i.e., $\mathbb{E}[X^{(n)}] = \mathbb{E}[\tilde{X}^{(n)}]$, iii) the two chains are coupled in such a way that there exists an almost surely finite meeting time τ, i.e., $X^{(n)} = \tilde{X}^{(n-1)}$ for all $n \geq \tau$. These properties imply:

$$\begin{aligned}\pi(h) &= \lim_{n\to\infty} \mathbb{E}[h(X^{(n)})] \qquad (8)\\ &= \mathbb{E}[h(X^{(0)})] + \sum\nolimits_{n=1}^{\infty} \mathbb{E}[h(X^{(n)})] - \mathbb{E}[h(\tilde{X}^{(n-1)})]\\ &= \mathbb{E}\left[h(X^{(0)})] + \sum\nolimits_{n=1}^{\tau} h(X^{(n)}) - h(\tilde{X}^{(n-1)})\right].\end{aligned}$$

The first equality follows from property (i), the second from property (ii), and the third by property (iii), after interchanging expectation and summation, as $X^{(n)} = \tilde{X}^{(n-1)}$ for all $n \geq \tau$ allows us to truncate the infinite sum. We define the *Rhee-Glynn estimator* of $\pi(h)$ as

$$H_0 := h(X^{(0)}) + \sum\nolimits_{n=1}^{\tau} h(X^{(n)}) - h(\tilde{X}^{(n-1)}), \qquad (9)$$

which, from (8), is an unbiased estimate. In the following subsections we describe how Markov chains $(X^{(n)})_{n\geq 0}$ and $(\tilde{X}^{(n)})_{n\geq 0}$ with the requisite properties (i)-(iii) can be constructed for state-space models of the form (2).

C. Conditional particle filter for smoothing

The first requirement is that the Markov chains $(X^{(n)})_{n\geq 0}$ and $(\tilde{X}^{(n)})_{n\geq 0}$ are invariant to the smoothing distribution π. We make use of a variant of the particle filter (PF) [27], [28], [29] called a *conditional* particle filter (CPF) [13], cf. Algorithm 1. The procedure amounts to producing a weighted particle system $\{w_t^j, x_t^j\}_{t=1:T}^{j=1:N}$, as in the regular PF, except that one of the trajectories (i.e., sequence of particles) is fixed to a given 'reference trajectory'. The CPF returns a trajectory by randomly selecting one of the particles with probability given by the weights, $w_T^{1:N}$, at the final time. For a reference trajectory $X_{\text{ref}} = \bar{x}$, sampling a new trajectory $X = x$ from the CPF is denoted $X \sim \text{CPF}(X_{\text{ref}}, \cdot)$. The CPF defines a Markov kernel on the space of trajectories that leaves π invariant. A Markov chain $(X^{(n)})_{n\geq 0}$ constructed by sampling $X^{(n)} \sim \text{CPF}(X^{(n-1)}, \cdot)$ thereby satisfies satisfies property (i).

Line 4 of Algorithm 1 is the resampling step. Here, $r(w_{t-1}^{1:N})$ denotes a distribution over $\{1, \dots, N\}^{N-1}$. Ancestor a_t^j is drawn from a distribution over $\{1, \dots, N\}$ with probabilities given by the weights $w_t^{1:N}$, for all $j = 1, \dots, N-1$ and the Nth ancestor is set deterministically as $a_t^N = N$ according to the conditioning.

D. Coupled conditional particle filters

To satisfy properties (ii) and (iii), i.e. the requirement that the Markov chains $(X^{(n)})_{n\geq 0}$ and $(\tilde{X}^{(n)})_{n\geq 0}$ have the same marginal distribution for each n and 'meet' in finite time τ, respectively, it is necessary to couple the CPF chains in some way. One possibility is to employ the so-called *coupled conditional particle filter* (CCPF) [19], cf. Algorithm 2. Coupling between the CPF chains occurs via two mechanisms in the CCPF. First, the same realization of the random variable U is used to propagate the state of both chains, cf. line 5. Second, the resampling (line 4) and trajectory selection (line 8) is performed via a joint distribution over the indices for both chains. Specifically, the distribution $\bar{r}(w_{t-1}^{1:N}, \hat{w}_{t-1}^{1:N})$ is such that ancestors $(a_t^j, \hat{a}_t^j)$ are sampled in order to maximize the probability of $a_t^j = \hat{a}_t^j$ under the marginal constraints $\mathbb{P}(a_t^j = k) = w_t^k$ and $\mathbb{P}(\hat{a}_t^j = k) = \hat{w}_t^k$, for all $k = 1, \dots, N$ (cf. [19] for details). This sampling strategy ensures that each CPF kernel leaves π invariant. The two coupling mechanisms combined are sufficient to ensure finite meeting time τ, cf. [19, §3]. However, the joint distribution over $\{1, \dots, N\}^2$ can be designed to maximize the probability that $a_t^j = \hat{a}_t^j$. This in turn minimizes τ which reduces the number of iterations (and hence computational complexity) of the unbiased estimator, cf. (9). For specific coupled sampling strategies, cf. [19, §2.2].

Algorithm 1: Conditional particle filter (CPF)

```
   Data: outputs $y$, reference trajectory $\bar{x}$, parameters $\theta$,
         number of particles $N$
1  Draw $U_0^j$ and set $x_0^j = M(U_0^j, \theta)$ for $j = 1 : N-1$,
     $x_0^N = \bar{x}_0$ ;
2  Set $w_0^j = N^{-1}$ for $j = 1 : N$ ;
3  for  $t = 1 : T$ do
4      Draw ancestors $a_{t-1}^{1:N-1} \sim r(w_{t-1}^{1:N})$, and set
         $a_{t-1}^N = N$ ;
5      Draw $U_t^j$ and set $x_t^j = F(x_{t-1}^{a_{t-1}^j}, U_t^j, \theta)$ for
         $j = 1 : N-1$, $x_t^N = \bar{x}_t$ ;
6      Compute $w_t^j = g(y_t | x_t^j, \theta)$ for $j = 1 : N$, and
         normalize: $\bar{w}_t = \sum_j w_t^j$, $w_t^j \leftarrow w_t^j / \bar{w}_t$ ;
7  end
8  Draw $b_T$ from discrete distribution over $1 : N$ with
     probabilities $w_T^{1:N}$ ;
9  Set $b_t = a_t^{b_{t+1}}$ for $t = T-1, \dots, 0$ ;
10 return $x = (x_0^{b_0}, \dots, x_T^{b_T})$ ;
```

To summarize, the CCPF takes two reference trajectories $X_{\text{ref}} = \bar{x}$ and $\tilde{X}_{\text{ref}} = \tilde{x}$, and returns two new trajectories X and $\tilde{X}$ from CPFs, coupled as described in Algorithm 2. We denote this operation by $(X, \tilde{X}) \sim \text{CCPF}(X_{\text{ref}}, \tilde{X}_{\text{ref}}, \cdot)$. The CCPF then defines a Markov kernel on the space of *pairs* of trajectories. The Markov chain formed by evolving the kernel $(X^{(n+1)}, \tilde{X}^{(n)}) \sim \text{CCPF}(X^{(n)}, \tilde{X}^{(n-1)}, \cdot)$ is such that $X^{(n)} \sim \text{CPF}(X^{(n-1)}, \cdot)$, $\tilde{X}^{(n)} \sim \text{CPF}(\tilde{X}^{(n-1)}, \cdot)$, with $\tau := \inf\{n : X^{(n)} = \tilde{X}^{(n-1)}\}$ finite in expectation.

E. Final Rhee-Glynn debiased gradient approximation

With the specification of the CCPF in §III-D we now have the Markov chains required to implement the debiasing method described in §III-B. The procedure is detailed in Algorithm 3. Line 1 requires sampling the initial trajectories from a 'particle filter'. This amounts to running a PF and then returning a randomly selected trajectory from the weighted particle system, with probability given by the weights at the final time, as in line 9 of Algorithm 1. In fact, the entire

Algorithm 2: Coupled conditional PF (CCPF)

```
   Data: outputs $y$, reference trajectories $\bar{x}$ and $\tilde{x}$,
         parameters $\theta$, number of particles $N$
1  Draw $U_0^j$, set $x_0^j = M(U_0^j, \theta)$ and $\hat{x}_0^j = M(U_0^j, \theta)$,
     for $j = 1 : N-1$, $x_0^N = \bar{x}_0$ and $\hat{x}_0^N = \tilde{x}_0$ ;
2  Set $w_0^j = N^{-1}$ and $\hat{w}_0^j = N^{-1}$ for $j = 1 : N$ ;
3  for  $t = 1 : T$ do
4      Draw ancestors
         $(a_{t-1}^{1:N-1}, \hat{a}_{t-1}^{1:N-1}) \sim \bar{r}(w_{t-1}^{1:N}, \hat{w}_{t-1}^{1:N})$, and set
         $a_{t-1}^N = N$ and $\hat{a}_{t-1}^N = N$ ;
5      Draw $U_t^j$ and set $x_t^j = F(x_{t-1}^{a_{t-1}^j}, U_t^j, \theta)$ and
         $\hat{x}_t^j = F(\hat{x}_{t-1}^{\hat{a}_{t-1}^j}, U_t^j, \theta)$ for $j = 1 : N-1$,
         $x_t^N = \bar{x}_t$ and $\hat{x}_t^N = \tilde{x}_t$ ;
6      Compute $w_t^j = g(y_t | x_t^j, \theta)$ and $\hat{w}_t^j = g(y_t | \hat{x}_t^j, \theta)$
         for $j = 1 : N$, and normalize: $\bar{w}_t = \sum_j w_t^j$,
         $w_t^j \leftarrow w_t^j / \bar{w}_t$, $\bar{\hat{w}}_t = \sum_j \hat{w}_t^j$, $\hat{w}_t^j \leftarrow \hat{w}_t^j / \bar{\hat{w}}_t$ ;
7  end
8  Draw $(b_T, \hat{b}_T) \sim \bar{r}(w_T^{1:N}, \hat{w}_T^{1:N})$ ;
9  Set $b_t = a_t^{b_{t+1}}$ and $\hat{b}_t = \hat{a}_t^{\hat{b}_{t+1}}$ for $t = T-1, \ldots, 0$ ;
10 return $x = (x_0^{b_0}, \ldots, x_T^{b_T})$ and $\hat{x} = (\hat{x}_0^{\hat{b}_0}, \ldots, \hat{x}_T^{\hat{b}_T})$ ;
```

process of sampling the trajectories from the PF, as in line 1, is identical to the CPF described in Algorithm 1, if one simply ignores the Nth particle (i.e. the reference trajectory).

Algorithm 3: Rhee-Glynn gradient estimation

```
Data: gradient of joint log-likelihood
      $h(x) = \nabla_\theta \log p(x, y | \theta)$, parameters $\theta$
Draw $X^{(0)}$ and $\tilde{X}^{(0)}$ from the particle filter ;
Draw $X^{(1)} \sim$ CPF$(X^{(0)}, \cdot)$ ;
Set $n = 1$ ;
while  $n \le \tau$, $\tau := \inf\{n : X^{(n)} = \tilde{X}^{(n-1)}\}$ do
    Draw $(X^{(n+1)}, \tilde{X}^{(n)}) \sim$ CCPF$(X^{(n)}, \tilde{X}^{(n-1)}, \cdot)$ ;
    $n \leftarrow n + 1$
end
return
  $H_0 := h(X^{(0)}) + \sum_{n=1}^{\tau} h(X^{(n)}) - h(\tilde{X}^{(n-1)})$ ;
```

IV. Adaptive-thermostat for state-space models

We are now in a position to present the main contribution of this paper: an adaptive-thermostat for parameter posterior sampling in state-space models. We combine the adaptive-thermostat outlined in §II-B with the method for debiasing log-likelihood gradient estimates presented in §III. To implement the scheme, it is necessary to employ a discretization of the continuous time Nosé-Hoover dynamics in (5). A simple forward Euler discretization, with stepsize ϵ, yields:

$$\begin{aligned} q_{t+1} &= q_t + \mathcal{M}^{-1} \rho_t \epsilon, \qquad (10) \\ \rho_{t+1} &= \rho_t - \nabla \tilde{U}(q_{t+1}) \epsilon - \zeta_t \rho_t \epsilon + \sigma_a \mathcal{M}^{1/2} e, \quad e \sim \mathcal{N}(0, I), \\ \zeta_{t+1} &= \zeta_t + k_i \left(\rho_{t+1}' \mathcal{M}^{-1} \rho_{t+1} - d k_B \mathcal{T} \right) \epsilon. \end{aligned}$$

We note that it is possible to employ more advanced (e.g. higher order) discretization strategies, which permit the use of larger stepsizes [5]. The stochastic gradient $\nabla \tilde{U}(q_{t+1})$ in (10) is given by

$$\nabla \tilde{U}(q_{t+1}) = H_0 + \nabla_\theta \log p(\theta)|_{\theta = q_{t+1}} \qquad (11)$$

where H_0 is computed by running Algorithm 3 with input $\theta = q_{t+1}$. Samples q_t generated by simulating (10) are then *approximately* distributed according to $p(\theta | y)$. We say *approximately*, as (10) is only guaranteed to leave the target canonical distribution invariant when the stochastic gradient $\nabla \tilde{U}$ is given by the exact gradient corrupted with additive Gaussian noise of unknown but constant covariance, i.e. $\nabla \tilde{U}(q) = \nabla U(q) + e$ with $e \sim \mathcal{N}(0, \Sigma)$.

A. Discussion

We conclude with a brief discussion of some extensions, limitations and practical considerations associated with the proposed approach, beginning with the assumption that stochasticity in the gradient is due to additive Gaussian noise.

1) State-dependent noise covariance: In some applications, it is more appropriate to model the stochasticity in the gradient $\nabla \tilde{U}$ by additive Gaussian noise with covariance that is *state-dependent*, i.e., $\nabla \tilde{U}(q) = \nabla U(q) + e$ with $e \sim \mathcal{N}\left(0, \bar{\Sigma}(q)\right)$. In this setting, it is desirable to introduce additional damping in (5), that is a function of the empirical covariance of the stochastic gradient. The resulting sampler is known as the Covariance-Controlled Adaptive Langevin thermostat and, under these assumptions, has been shown to leave the canonical distribution invariant; cf. [30] for details.

2) Algorithm parameters: We can distinguish between two types of parameters the must be specified to operate the thermostat. First, there are those that arise due to the statistical mechanics interpretation of the thermostat as a method for simulating from a canonical ensemble, e.g., m, $\mathcal{T}$, k_i. The parameters affect the molecular dynamics, including ρ and ζ, but not the distribution $\bar{p}_\beta$ over q which is of primary interest for inference. It is common to set m, $dk_B\mathcal{T}$, and k_i to unity [4]. Second, there are the parameters that have a greater influence on the efficacy of the thermostat to sample from the desired distribution, namely the stepsize ϵ and artificial noise variance σ_a^2. Too small a stepsize can result in slow exploration of the parameter space, whereas too large a stepsize can result in excessive relative error or even divergence. For a discussion on balancing this tradeoff via selection of ϵ and σ_a^2, cf. [5, §4].

3) Computational complexity: Roughly speaking, the computational complexity of the proposed method is comparable to other particle MCMC methods, in that it uses the same underlying machinery, namely, the particle filter; cf., e.g., [15, §3.1] for a discussion on computational complexity of such methods. One of the key factors that determines the computational cost of the debiasing method presented in §III is the meeting time, τ. Specifically, the sooner the chains meet, the fewer the required realizations from the CCPF. For coupled sampling strategies to minimize τ, cf., [19, §2.2].

V. Numerical examples

A. Illustration on scalar nonlinear system

We now return briefly to the illustration presented in Fig 1, and fill-in the remaining details. We assume a uniform prior over $\theta \in \mathbb{R}$, i.e. $p(\theta) \propto 1$. This implies that $p(\theta|y) \propto p(y|\theta)$, which means that we can approximately evaluate the posterior (pointwise, after normalization) using a PF. This PF approximation provides our baseline for comparison. Problem data is generated by sampling $u_t \sim \mathcal{N}(0,1)$ for $t = 1 : 50$ and simulating (1) with the true parameter, to obtain a realization of y. The biased gradient estimates in Fig 1(a) are generated using the MC approximation of the Fisher identity in (7), with samples $\tilde{x}^i$ from a forward filtering backward simulation (FFBSi) particle smoother. The posteriors in Fig 1(c) represent histograms (50 bins) of the position trajectory q_t, after running the thermostat (10) for 10^4 timesteps (discarding the first 10^3 samples for burn-in).

Finally, we note that the debiasing method introduces additional variance to the stochastic gradient, compared to the FFBSi. Nevertheless, the thermostat (with the same tuning parameters for each gradient method, $\epsilon = 0.05$ and $\sigma_a = 1$) is able to effectively dissipate this additional noise.

B. Wiener system

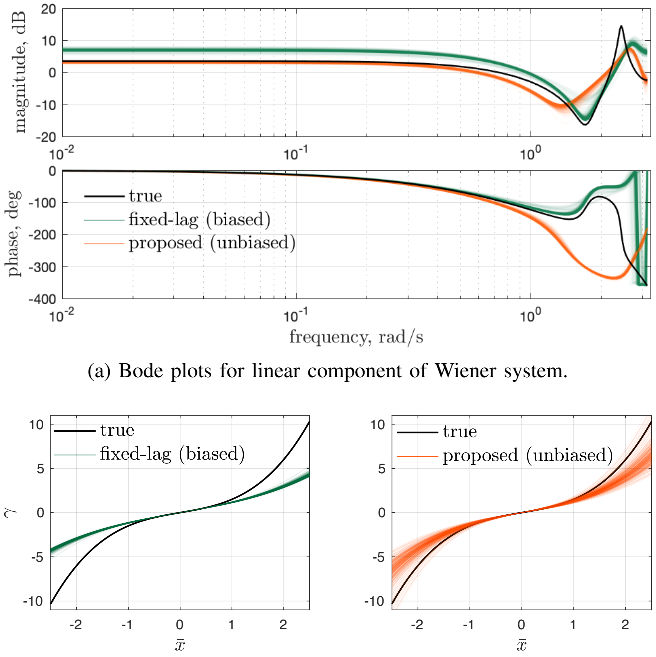
In this experiment we consider identification of a Wiener system, comprised of the transfer function

$$\bar{x}_t = G(z) = \frac{\theta_5 z^{-1} + \theta_6 z^{-2} + \theta_7 z^{-3} + \theta_8 z^{-4}}{1 + \theta_1 z^{-1} + \theta_2 z^{-2} + \theta_3 z^{-3} + \theta_4 z^{-4}} u_t + w_t,$$

in series with the static output nonlinearity $y = \gamma(\bar{x}) + v_t = \theta_9 \bar{x}_t^2 + \theta_{10} \bar{x}_t^3 + v_t$. Here, w_t and v_t denote process and measurement noise, respectively, and are each distributed according to $\mathcal{N}\left(0, 0.1^2\right)$. The true parameters are given by $\theta = (0.8, 0.1, -0.4, 0.1, 0.4, 0.9, 0.5, 0.6, 1, 0.5)$. We assume a uniform prior over $\theta \in \mathbb{R}$, i.e. $p(\theta) \propto 1$. Problem data y was obtained by simulating the system, with the true parameters, excited by $u_t \sim \mathcal{N}(0, 0.4)$ for $t = 1 : 50$.

To draw samples from the posterior, we apply two thermostats (10), with $\epsilon = 0.05$, $\sigma_a = 1$ and all other parameters set to unity. The first uses stochastic gradients computed using the approximate Fisher identity (7) with states $\tilde{x}$ from a fixed-lag smoother [31] ($N = 2$ particles, lag = 5). The second uses stochastic gradients computed with the proposed debiasing method ($N = 2$ particles). Each thermostat was simulated for 10×10^3 timesteps; after discarding the first 10^3 samples for burn-in, the resulting position trajectories $q_t = \theta \in \mathbb{R}^{10}$, converted to Bode representations of G and (static) $\gamma(\cdot)$, are depicted in Fig 2.

Unlike the illustration in §I-B, (an approximation) of the 'ground truth' posterior is not available for comparison. Nonetheless, at lower frequencies in the Bode plots, especially for the magnitude, as well as the static nonlinearity γ, we observe less bias (relative to the true system) in samples generated using the proposed debiased gradients.



(a) Bode plots for linear component of Wiener system.

(b) Static output map for Wiener system.

Fig. 2: Samples from two thermostats targeting the parameter posterior distribution of the Wiener system; cf. §V-B for details. 'Fixed-lag' makes use of biased stochastic gradients from a fixed-lag smoother; 'proposed' denotes stochastic gradients from the debiasing method outlined in this paper.

References

[1] L. Ljung, *System Identification: Theory for the User*, 2nd ed. Prentice Hall, 1999.

[2] V. Peterka, "Bayesian system identification," *Automatica*, vol. 17, no. 1, pp. 41–53, 1981.

[3] R. M. Neal, "MCMC using Hamiltonian dynamics," *Handbook of Markov Chain Monte Carlo*, pp. 113–162, 2011.

[4] N. Ding, Y. Fang, R. Babbush, C. Chen, R. D. Skeel, and H. Neven, "Bayesian sampling using stochastic gradient thermostats," in *Advances in neural information processing systems (NIPS)*, 2014, pp. 3203–3211.

[5] B. Leimkuhler and X. Shang, "Adaptive thermostats for noisy gradient systems," *SIAM Journal on Scientific Computing*, vol. 38, no. 2, pp. A712–A736, 2016.

[6] P. W. Glynn and C.-h. Rhee, "Exact estimation for Markov chain equilibrium expectations," *Journal of Applied Probability*, vol. 51, no. A, pp. 377–389, 2014.

[7] S. Duane, A. D. Kennedy, B. J. Pendleton, and D. Roweth, "Hybrid Monte Carlo," *Physics letters B*, vol. 195, no. 2, pp. 216–222, 1987.

[8] M. Welling and Y. W. Teh, "Bayesian learning via stochastic gradient Langevin dynamics," in *Proceedings of the 28th International Conference on Machine Learning (ICML)*, 2011, pp. 681–688.

[9] S. Patterson and Y. W. Teh, "Stochastic gradient Riemannian Langevin dynamics on the probability simplex," in *Advances in neural information processing systems (NIPS)*, 2013, pp. 3102–3110.

[10] T. Chen, E. Fox, and C. Guestrin, "Stochastic gradient Hamiltonian Monte Carlo," in *International Conference on Machine Learning (ICML)*, 2014, pp. 1683–1691.

[11] S. J. Vollmer, K. C. Zygalakis *et al.*, "(non-) asymptotic properties of stochastic gradient Langevin dynamics," *arXiv preprint arXiv:1501.00438*, 2015.

[12] A. A. Samoletov, C. P. Dettmann, and M. A. Chaplain, "Thermostats for "slow" configurational modes," *Journal of Statistical Physics*, vol. 128, no. 6, pp. 1321–1336, 2007.

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

[14] J. Fernández-Villaverde and J. F. Rubio-Ramírez, “Estimating macroeconomic models: A likelihood approach,” *The Review of Economic Studies*, vol. 74, no. 4, pp. 1059–1087, 2007.

[15] N. Kantas, A. Doucet, S. S. Singh, J. Maciejowski, N. Chopin *et al.*, “On particle methods for parameter estimation in state-space models,” *Statistical science*, vol. 30, no. 3, pp. 328–351, 2015.

[16] F. Lindsten and A. Doucet, “Pseudo-marginal Hamiltonian Monte Carlo,” arXiv.org, arXiv:1607.02516, 2016.

[17] N. Whiteley, C. Andrieu, and A. Doucet, “Efficient Bayesian inference for switching state-space models using discrete particle Markov chain Monte Carlo methods,” *arXiv preprint arXiv:1011.2437*, 2010.

[18] F. Lindsten, M. I. Jordan, and T. B. Schön, “Particle Gibbs with ancestor sampling,” *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 2145–2184, 2014.

[19] P. E. Jacob, F. Lindsten, and T. B. Schön, “Smoothing with couplings of conditional particle filters,” *arXiv preprint arXiv:1701.02002*, 2017.

[20] D. McLeish, “A general method for debiasing a Monte Carlo estimator,” *arXiv preprint arXiv:1005.2228*, 2010.

[21] C.-h. Rhee and P. W. Glynn, “A new approach to unbiased estimation for SDE’s,” in *Proceedings of the Winter Simulation Conference*. Winter Simulation Conference, 2012, p. 17.

[22] M. Vihola, “Unbiased estimators and multilevel Monte Carlo,” *Operations Research*, vol. 66, no. 2, pp. 448–462, 2017.

[23] J. C. Mattingly, A. M. Stuart, and D. J. Higham, “Ergodicity for sdes and approximations: locally Lipschitz vector fields and degenerate noise,” *Stochastic processes and their applications*, vol. 101, no. 2, pp. 185–232, 2002.

[24] O. Cappé, E. Moulines, and T. Rydén, *Inference in hidden Markov models*. Berlin, Germany: Springer, 2005.

[25] M. Briers, A. Doucet, and S. Maskell, “Smoothing algorithms for state–space models,” *Annals of the Institute of Statistical Mathematics*, vol. 62, no. 1, p. 61, 2010.

[26] F. Lindsten and T. B. Schön, “Backward simulation methods for Monte Carlo statistical inference,” *Foundations and Trends in Machine Learning*, vol. 6, no. 1, pp. 1–143, 2013.

[27] L. Stewart and P. McCarty, “The use of Bayesian belief networks to fuse continuous and discrete information for target recognition and discrete information for target recognition, tracking, and situation assessment,” in *Proceedings of SPIE Signal Processing, Sensor Fusion and Target Recognition*, vol. 1699, 1992, pp. 177–185.

[28] N. J. Gordon, D. J. Salmond, and A. F. M. Smith, “Novel approach to nonlinear/non-Gaussian Bayesian state estimation,” in *IEE Proceedings on Radar and Signal Processing*, vol. 140, 1993, pp. 107–113.

[29] G. Kitagawa, “A Monte Carlo filtering and smoothing method for non-Gaussian nonlinear state space models,” in *Proceedings of the 2nd US-Japan joint Seminar on Statistical Time Series Analysis*, 1993, pp. 110–131.

[30] X. Shang, Z. Zhu, B. Leimkuhler, and A. J. Storkey, “Covariance-controlled adaptive Langevin thermostat for large-scale Bayesian sampling,” in *Advances in Neural Information Processing Systems (NIPS)*, 2015, pp. 37–45.

[31] G. Kitagawa and S. Sato, “Monte Carlo smoothing and self-organising state-space model,” in *Sequential Monte Carlo methods in practice*. Springer, 2001, pp. 177–195.