Gradient-free Parameter Estimation for Hidden Markov Models with intractable likelihoods

BY ELENA EHRLICH[1], AJAY JASRA[2] & NIKOLAS KANTAS[3]

[1]Department of Mathematics, Imperial College London, London, SW7 2AZ, UK.
E-Mail: elena.ehrlich05@ic.ac.uk
[2]Department of Statistics & Applied Probability, National University of Singapore, Singapore, 117546, SG.
E-Mail: staja@nus.edu.sg
[3]Department of Statistical Science, University College London, London, WC1E 6BT, UK.
E-Mail: n.kantas@ucl.ac.uk

Abstract

In this article we focus on Maximum Likelihood estimation (MLE) for the static model parameters of hidden Markov models (HMMs). We will consider the case where one cannot or does not want to compute the conditional likelihood density of the observation given the hidden state because of increased computational complexity or analytical intractability. Instead we will assume that one may obtain samples from this conditional likelihood and hence use approximate Bayesian computation (ABC) approximations of the original HMM. Although these ABC approximations will induce a bias, this can be controlled to arbitrary precision via a positive parameter ϵ, so that the bias decreases with decreasing ϵ. We first establish that when using an ABC approximation of the HMM for a fixed batch of data, then the bias of the resulting log- marginal likelihood and its gradient is no worse than $\mathcal{O}(n\epsilon)$, where n is the total number of data-points. Therefore, when using gradient methods to perform MLE for the ABC approximation of the HMM, one may expect parameter estimates of reasonable accuracy. To compute an estimate of the unknown and fixed model parameters, we propose a gradient approach based on simultaneous perturbation stochastic approximation (SPSA) and Sequential Monte Carlo (SMC) for the ABC approximation of the HMM. The performance of this method is illustrated using two numerical examples.

Key-Words: Approximate Bayesian Computation, Hidden Markov Models, Parameter Estimation, Sequential Monte Carlo

1 Introduction

Hidden Markov models (HMMs) provide a flexible description of a wide variety of real-life phenomena when a time varying latent process is observed independently at different epochs. A HMM can be defined as a pair of discrete-time stochastic processes, $(X_t, Y_{t+1})_{t\geq 0}$, where $X_t \in \mathsf{X} \subseteq \mathbb{R}^{d_x}$ is the unobserved process and $Y_t \in \mathsf{Y} \subseteq \mathbb{R}^{d_y}$ is the observation at time t. Let $\theta \in \Theta \subset \mathbb{R}^{d_\theta}$ be a vector containing the static parameters of the model. The hidden process $(X_t)_{t\geq 0}$ is assumed to be Markov chain with initial density $\mu_\theta(x_0)$ at time 0 and transition density $f_\theta(x_t|x_{t-1})$ so that

$$\mathbb{P}_\theta(X_0 \in A) = \int_A \mu_\theta(x_0)dx_0 \quad \text{and} \quad \mathbb{P}_\theta\left(X_t \in A | (X_m, Y_{m+1})_{m\geq 0} = (x_m, y_{m+1})_{m\geq 0}\right) = \int_A f_\theta(x_t|x_{t-1})dx_t \quad t \geq 1, \tag{1}$$

where $\mathbb{P}_\theta$ denotes the probability, A belongs to the Borel σ-algebra of X, $\mathsf{B}(\mathsf{X})$, and dx_t is the Lebesgue measure. In addition, each observation Y_t is assumed to be statistically independent of every other quantity except X_t, θ :

$$\mathbb{P}_\theta\left(Y_t \in B | (X_m, Y_{m+1})_{m\geq 0} = (x_m, y_{m+1})_{m\geq 0}\right) = \int_B g_\theta(y_t|x_t)dy_t \quad t > 0 \tag{2}$$

with $B \in \mathsf{B}(\mathsf{Y})$ and $g_\theta(y_t|x_t)$ being the *conditional likelihood* density. The HMM is given by equations (1)-(2) and is often referred to in the literature also as a general state-space model. Here θ is treated as a unknown and static model parameter, which is to be estimated in using Maximum Likelihood estimation (MLE). This is an important problem with many applications ranging from financial modeling to numerical weather prediction.

Statistical inference for the class of HMMs described above is typically non-trivial. In most scenarios of practical interest one cannot calculate the *marginal likelihood* of n given observations

$$p_\theta(y_{1:n}) = \int g_\theta(y_n|x_n)p_\theta(x_n|y_{1:n-1})dx_n$$

where $y_{1:n} := (y_1, \ldots, y_n)$ are considered fixed and $p_\theta(x_n|y_{1:n-1})$ is the predictor density at time n. Hence as the likelihood is not analytically tractable, one must resort to numerical methods to both compute and to maximize

$p_\theta(y_{1:n})$ w.r.t. θ. When θ is known, a popular collection of techniques for both estimating the likelihood as well as performing *filtering* or *smoothing* are sequential Monte Carlo (SMC) methods [18, 7]. SMC techniques simulate a collection of N samples (known as *particles*) in parallel, sequentially in time and combine importance sampling and resampling to approximate a sequence of probability distributions of increasing state-space known point-wise up-to a multiplicative constant. These techniques provide a natural estimate of the likelihood $p_\theta(y_{1:n})$. The estimate is quite well understood and is known to be unbiased [13, Chapter 9]. In addition, the relative variance of this quantity is known to increase linearly with the number of data-points, n, [10, 37]. When θ is unknown, as is the case here, estimation of θ is further complicated, because of the path-degeneracy caused to the population of the samples by the resampling step of SMC. This issue has been well documented int the literature [1, 21]. However, there are still many specialized SMC techniques which can successfully be used for parameter estimation of HMMs in a wide variety of contexts; see [21] for an comprehensive overview. In particular for MLE a variety of SMC methods have been proposed in the literature [8, 16, 31]. Note that the techniques in these papers require the evaluation of $g_\theta(y|x)$ and potentially gradient vectors as well.

In this article, we consider the scenario where $g_\theta(y|x)$ is intractable. By this we mean that one cannot calculate it for given y or x either because the density does not exist or because it is computationally too expensive, e.g. due to the high-dimensionality of x. In addition, we will assume a unbiased estimator for $g_\theta(y|x)$ is also not available. Instead we will assume that one can sample from $g_\theta(\cdot|x)$ for any value of x. In this case, one cannot use the standard or the more advanced SMC methods that are mentioned above (or indeed many other simulation based approximations). Hence the problem of parameter estimation is very difficult. One approach which is designed to deal with this problem is Approximate Bayesian Computation (ABC). ABC is an approach that uses simulated samples from the likelihood to deal with the restriction of not being to evaluate its density. Although there is nothing inherently Bayesian about this, it owes its name due to its early success in Bayesian inference; see [26] and the references therein for more details. Although here we will focus only upon ABC ideas, we note that there are possible alternatives, such as [19], and refer the interested reader to [19, 20] for a discussion of the relative merits of ABC.

In the context of HMMs when the model parameters θ are known, the use of ABC approximations has appeared in [20, 28] as well as associated computational methods for filtering and smoothing in [20, 27, 9]. When the parameter is unknown, the statistical properties of ML estimators for θ based on ABC approximations has been studied in detail in [11, 12]. ABC approximations of lead to a bias, which can be controlled to arbitrary precision via a parameter $\epsilon > 0$. This bias typically goes to zero as $\epsilon \searrow 0$. In this article we aim to:

1. investigate the bias in the log-likelihood and the gradient of the log-likelihood that is induced by the ABC approximation for a fixed data set,

2. develop a gradient based approach based on SMC with computational cost $\mathcal{O}(N)$ that allows one to estimate the model parameters in either a batch or on-line fashion.

In order to implement such an approach one must obtain numerical estimates of the log- marginal likelihood as well as its gradient. Thus, it is important to understand what happens to the bias of the ABC approximation of these latter quantities, as the time parameter (or equivalently number of data-points, n) grows. We establish, under some assumptions, that this ABC bias, for both quantities is no worse than $\mathcal{O}(n\epsilon)$. This result is closely associated to the theoretical work in [11, 12]. These former results indicate that the ABC approximation is amenable to numerical implementation and parameter estimation will not necessarily be dominated by the bias. We will discuss why this is the case later in Remarks 2.1 and 2.2. For the numerical implementation of MLE we will introduce a gradient-free approach based on using finite differences with Simultaneous Perturbation Stochastic approximation (SPSA) [34, 33]. This is extending the work in [32] for the case when the likelihood is intractable and ABC approximations are used.

This paper is structured as follows. In Section 2 we discuss the estimation procedure using ABC approximations. Our bias result is also given. In Section 3 our computational strategy is outlined. In Section 4 the method is investigated from a numerical perspective. In Section 5 the article is concluded with some discussion of future work. The proofs of our results can be found in the Appendix.

2 Model and Approximation

2.1 Maximum Likelihood for Hidden Markov models

Consider first the *joint filtering* density of the HMM given by

$$\pi_\theta(x_{0:n}|y_{1:n}) = \frac{\mu_\theta(x_0)\prod_{t=1}^n g_\theta(y_t|x_t)f_\theta(x_t|x_{t-1})}{\int_{\mathsf{X}^{n+1}}\mu_\theta(x_0)\prod_{t=1}^n g_\theta(y_t|x_t)f_\theta(x_t|x_{t-1})dx_{0:n}},$$

where we recall that $\theta \in \Theta \subset \mathbb{R}^{d_\theta}$ is the vector of model parameters, $x_t \in \mathsf{X}$ are the hidden states and $y_t \in \mathsf{Y}$ the observations. The joint filtering density can be computed recursively using the well known Bayesian filtering recursions:

$$\pi_\theta(x_{0:t}|y_{1:t-1}) = \int_{\mathsf{X}} \pi_\theta(x_{0:t-1}|y_{1:t-1})f_\theta(x_t|x_{t-1})dx_t \tag{3}$$

$$\pi_\theta(x_{0:t}|y_{1:t}) = \frac{g_\theta(y_t|x_t)\pi_\theta(x_{0:t}|y_{1:t-1})}{p_\theta(y_t|y_{1:t-1})} \tag{4}$$

where the normalizing constant in (4) is referred to as *recursive likelihood* and is given as follows:

$$p_\theta(y_t|y_{1:t-1}) = \int_{\mathsf{X}} g_\theta(y_t|x_t)\pi_\theta(x_{0:t}|y_{1:t-1})dx_t \tag{5}$$

Furthermore, we write the log-(marginal) likelihood at time n:

$$l_\theta(y_{1:n}) = \log(p_\theta(y_{1:n})).$$

In the context of MLE one is usually interested computing

$$\hat{\theta} = \arg\max_{\theta\in\Theta} l_\theta(y_{1:n}).$$

Note that this is a batch or off-line procedure, which means that one needs to wait first to collect the complete data-set and then compute the ML estimate. In this paper we will focus on computing ML estimates based on gradient methods. In this case one may use iteratively for $k \geq 0$

$$\theta_{k+1} = \theta_k + a_{k+1}\, \nabla l_\theta\left(y_{1:n}\right)|_{\theta=\theta_k},$$

where $(a_k)_{k\geq1}$ is a step sequence that satisfies $\sum_k a_k = \infty$ and $\sum_k a_k^2 < \infty$, [4]. Note that this scheme is only guaranteed to converge to a local maximum and this is sensitive to initialization.

In case one expects a very long observation sequence, the computation of the gradient at each iteration of the above gradient ascent algorithm can be prohibitive. Therefore, one might prefer on-line ML methods, whereby the estimate of the parameter is updated sequentially as the data arrives. A practical alternative would be to consider maximizing instead the long run quantity

$$\lim_{n\to\infty}\frac{1}{n}l_\theta(y_{1:n}) = \lim_{n\to\infty}\sum_{t=1}^n \log\left(p_\theta(y_t|y_{1:t-1})\right).$$

Under appropriate regularity and ergodicity conditions for the augmented Markov chain $(X_t, Y_t, p_\theta(x_t|y_{1:t-1}))_{t\geq0}$ [23, 35] the average log-likelihood is an ergodic average and this leads to a gradient update scheme based on Stochastic Approximation [4]. For a similar step-size sequence $(a_t)_{t\geq1}$ one may update θ_t as follows:

$$\theta_{t+1} = \theta_t + a_{t+1}\, \nabla \log\left(p_\theta(y_t|y_{1:t-1})\right)|_{\theta=\theta_t}.$$

Upon receiving y_t, the parameter estimate is updated in the direction of ascent of the conditional density of this new observation. The algorithm in the present form is not suitable for on-line implementation due to the need to evaluate the gradient of $\log p_\theta(y_t|y_{0:t-1})$ at the current parameter estimate which would require computing the filter from time 0 to time t using the current parameter value θ_t. To bypass this problem, the *recursive ML (RML)* algorithm has been proposed originally in [2, 23, 24] for finite state spaces and in [16, 31, 17] in the context of SMC approximations. It relies on the following update scheme

$$\theta_{t+1} = \theta_t + a_{t+1}\nabla\log\left(p_{\theta_{0:t}}(y_t|y_{1:t-1})\right),$$

where the positive non-increasing step-size sequence $(a_t)_{t\geq 1}$ satisfies $\sum_t a_t = \infty$ and $\sum_t a_t^2 < \infty$, e.g. $a_t = t^{-\alpha}$ for $0.5 < \alpha \leq 1$. The quantity $\nabla \log p_{\theta_{0:t}}(y_t|y_{1:t-1})$ is defined here as

$$\nabla \log \left(p_{\theta_{0:t}}(y_t|y_{1:t-1})\right) = \nabla \log \left(p_{\theta_{0:t}}(y_{1:t})\right) - \nabla \log \left(p_{\theta_{0:t-1}}(y_{1:t-1})\right),$$

where the subscript $\theta_{0:t}$ in the notation for $\nabla \log \left(p_{\theta_{0:t}}(y_{1:t})\right)$ indicates that at each time t the quantities in (3)-(5) are computed using the current parameter estimate θ_t. The asymptotic properties of RML have been studied in [2, 22, 23, 24] for a finite state-space HMMs and [35, 36] in more general cases. It is shown that under regularity conditions this algorithm converges towards a local maximum of the average log-likelihood, whose maximum lies at the 'true' parameter value.

In this article, we would like to implement approximate versions of RML and off-line ML schemes when both the following cases hold:

- We can sample from the conditional distribution of $Y|x$, for any fixed θ and x.
- We cannot or do not want to evaluate the conditional density of $Y|x$, $g_\theta(y|x)$ and do not have access to an unbiased estimate of it.

Apart from using likelihoods which do not admit computable densities such as some stable distributions, this context might appear relevant to the context when one is interested to use SMC methods and evaluate $g_\theta(y|x)$ when d_x is large. SMC methods for filtering do not always scale well with the dimension of the hidden state d_x, often requiring a computational cost $\mathcal{O}(\kappa^{d_x})$, with $\kappa > 1$ [5, 6]. A more detailed discussion on the difficulties of using SMC methods in high dimensions is far beyond the scope of this article, but we remark the ideas in this paper can be relevant in this context.

2.2 ABC Approximations

To facilitate ML estimation when the bullet points above hold we will resort to ABC approximations of the ideal MLE procedures above. We will present a short overview here and refer the author to [11, 39] for more details.

First, we consider an ABC approximation of the joint smoothing density as in [20, 28]:

$$\pi_{\theta,\epsilon}(u_{1:n}, x_{0:n}|y_{1:n}) = \frac{\mu_\theta(x_0)\prod_{t=1}^n K_\epsilon(y_t, u_t) g_\theta(u_t|x_t) f_\theta(x_t|x_{t-1})}{p_{\theta,\epsilon}(y_{1:n})} \tag{6}$$

with the ABC marginal likelihood being

$$p_{\theta,\epsilon}(y_{1:n}) = \int_{\mathsf{X}^{n+1}\times\mathsf{Y}^n} \mu_\theta(x_0) \prod_{t=1}^n K_\epsilon(y_t, u_t) g_\theta(u_t|x_t) f_\theta(x_t|x_{t-1}) du_{1:n} x_{0:n} \tag{7}$$

and the ABC recursive likelihood

$$p_{\theta,\epsilon}(y_t|y_{1:t-1}) = \frac{p_{\theta,\epsilon}(y_{1:t})}{p_{\theta,\epsilon}(y_{1:t-1})}, \tag{8}$$

where $u_n \in \mathsf{Y}$ are pseudo observations and $K_\epsilon : \mathsf{Y}\times\mathsf{Y} \to \mathbb{R}_+ \cup \{0\}$ is some kernel function that has bandwidth that depends upon a precision parameter $\epsilon > 0$. We will also assume that the kernel is such that $K_\epsilon(y_t, u_t) = K_\epsilon(u_t, y_t)$. For example, possible choices could be:

$$K_\epsilon(y_t, u_t) = \mathbb{I}_{\{u:|y_t-u|<\epsilon\}}(u_t) \text{ or } K_\epsilon(y_t, u_t) = \exp\left(-\frac{1}{2}\left(y_t - u_t\right)^T \Sigma^{-1}\left(y_t - u_t\right)\right),$$

where $\mathbb{I}$ is the indicator function, $|\cdot|$ is the a vector norm, and Σ is a positive semi-definite $d_y \times d_y$ matrix.

Note in this context the quantity:

$$g_{\theta,\epsilon}(y_t|x_t) = \frac{1}{Z_\epsilon}\int_{\mathsf{Y}} K_\epsilon(y_t, u_t) g_\theta(u_t|x_t) du_t, \tag{9}$$

can be viewed as the likelihood of an alternative "perturbed" HMM that uses the same transition density but has $g_{\theta,\epsilon}$ as the likelihood. It can be easily shown that this HMM will admit a marginal likelihood of $\frac{1}{Z_\epsilon^n} p_{\theta,\epsilon}(y_{1:n})$ which is proportional to the one written above in (7), but the proportionality constant does not depend on θ. Note that a critical condition for this to hold is that we choose $K_\epsilon(y_t, u_t)$ such that the normalizing constant $Z_\epsilon = \int K_\epsilon(y_t, u_t) du_t$ of (9) does not depend upon x_t or θ.

The ABC-MLE approach we consider in this article will be then to use MLE for the perturbed HMM defined by $g_{\theta,\epsilon}$. For the off-line case let

$$l^{\epsilon}_{\theta}(y_{1:n}) = \log\left(p_{\theta,\epsilon}(y_{1:n})\right)$$

and denote the ABC_MLE estimate as

$$\hat{\theta^{\epsilon}} = \arg\max_{\theta\in\Theta} l^{\epsilon}_{\theta}(y_{1:n}). \tag{10}$$

Results on the consistency and efficiency of this method n grows can be found in [11, 12]. Under some regularity and other assumptions (such as the data originating from the HMM considered), the bias of the maximum likelihood estimator (MLE) is $\mathcal{O}(\epsilon)$. In addition, one may avoid encountering this bias asymptotically, if one adds appropriately noise to the observations. This procedure is referred to as *noisy ABC*, and then one can recover the true parameter. We remark that the methodology that is considered in this article can easily incorporate noisy ABC. However, there may be some reasons why one may not want to use noisy ABC:

1. the consistency results (currently) depend upon the data originating from the original HMM;

2. the current simulation-based methodology may not be able to be used efficiently for ϵ close to zero.

For point 1., if the data do not originate from the HMM of interest, it has not been studied what happens with regards to the asymptotics of noisy ABC for HMMs. It may be that some investigators might be uncomfortable with assuming that the data originate from the exactly the HMM being fitted. For point 2. the asymptotic bias (which is under assumptions either $\mathcal{O}(\epsilon)$ or $\mathcal{O}(\epsilon^2)$ [11, 12]) could be less than the asymptotic variance (under assumptions $\mathcal{O}(\epsilon^2)$ [11, 12]) as ϵ could be much bigger than unity when using current simulation methodology. We do not use noisy ABC in this article, but acknowledge its fundamental importance with regards to parameter estimation associated to ABC for HMMs; our approach is intended for cases where points similar to 1.-2. need to be taken into account.

For the ABC-RML we will define the time varying log- recursive likelihood as

$$r^{\epsilon}_{\theta_{0:t}}(y_{1:t}) = \log\left(p_{\theta_{0:t},\epsilon}(y_t|y_{1:t-1})\right)$$

where the subscript $\theta_{0:t}$ means again that at each time t one computes all the relevant quantities in (3)-(5) (with $g_{\theta,\epsilon}$ substituted instead of g_θ) using θ_t as the parameter value and $\theta_{0:t-1}$ has been used similarly in all the previous times. Finally we write the ABC-RML recursion for the parameter as

$$\theta_{t+1} = \theta_t + a_{t+1}\nabla r^{\epsilon}_{\theta_{0:t}}(y_{1:t}). \tag{11}$$

2.3 Bias Results

We now prove an upper-bound on the bias induced by the ABC approximation on the log- marginal likelihood and its gradient. The latter is more relevant for parameter estimation, but the mathematical arguments are considerably more involved for this quantity, in comparison to the ABC bias of the log-likelihood. Hence the log-likelihood is considered as a simple preliminary result. These results are to be taken in the context of ABC (not noisy ABC) and help to provide some guarantees associated to the numerics.

We consider for this section the scenario

$$K_\epsilon(y_t, u_t) = \mathbb{I}_{A_{\epsilon,y_t}}(u_t)$$

where the set A_{ϵ,y_t} is specified below. Here $|\cdot|$ should be understood to be an $\mathbb{L}_1-$norm. The hidden-state is assumed to lie on a *compact* set, i.e. X is compact. We use the notation $\mathcal{P}(\mathsf{X})$ to denote the class of probability measures on X and $\mathcal{M}(\mathsf{X})$ the collection of finite and signed measures on X. $\|\cdot\|$ denotes the total variation distance. The initial distribution of the hidden Markov chain is written as $\mu_\theta \in \mathcal{P}(\mathsf{X})$. In addition, we condition on the observed data and do not mention them in any mathematical statement of results (due to the assumptions below). We do not consider the instance of whether the data originate, or not, from a HMM. For the control of the bias of the gradient of the log-likelihood (Theorem 2.1), we assume that $d_\theta = 1$. This is not restrictive as one can use the arguments to prove analogous results when $d_\theta > 1$, by considering component-wise arguments for the gradient. In addition, for the gradient result, the derivative of μ_θ is written $\widetilde{\mu_\theta} \in \mathcal{M}(\mathsf{X})$ and constants $\underline{C}, \overline{C}, L$ are to be understood as arbitrary lower, higher bounds and Lipschitz constants respectively. We make the following assumptions, which are quite strong but are intended for keeping the proofs as short as possible.

(**A1**) *Lipschitz Continuity of the Likelihood.* There exist $L < +\infty$ such that for any $x \in \mathsf{X}$, $y, y' \in \mathsf{Y}$, $\theta \in \Theta$

$$|g_\theta(y|x) - g_\theta(y'|x)| \leq L|y - y'|.$$

(**A2**) *Statistic and Metric.* The set $A_{\epsilon,y}$ is:

$$A_{\epsilon,y} = \{u : |y - u| < \epsilon\}.$$

(**A3**) *Boundedness of Likelihood and Transition.* There exist $0 < \underline{C} < \overline{C} < +\infty$ such that for all $x, x' \in \mathsf{X}$, $y \in \mathsf{Y}$, $\theta \in \Theta$

$$\underline{C} \leq f_\theta(x'|x) \leq \overline{C},$$
$$\underline{C} \leq g_\theta(y|x) \leq \overline{C}.$$

(**A4**) *Lipschitz Continuity of the Gradient of the Likelihood.* $f_\theta(x'|x)$, $g_\theta(y|x')$ are differentiable in θ for each $x, x' \in \mathsf{X}$, $y \in \mathsf{Y}$. In addition, there exist $L < +\infty$ such that for any $x \in \mathsf{X}$, $y, y' \in \mathsf{Y}$, $\theta \in \Theta$

$$|\nabla g_\theta(y|x) - \nabla g_\theta(y'|x)| \leq L|y - y'|.$$

(**A5**) *Boundedness of Gradients of the Likelihood and Transition.* There exist $0 < \underline{C} < \overline{C} < +\infty$ such that for all $x, x' \in \mathsf{X}$, $y \in \mathsf{Y}$, $\theta \in \Theta$

$$\underline{C} \leq \nabla f_\theta(x'|x) \leq \overline{C},$$
$$\underline{C} \leq \nabla g_\theta(y|x) \leq \overline{C}.$$

Whilst it is fairly easy to find useful simple models where the above conditions do not hold uniformly for θ, we remark that the emphasis here is to provide intuition for the methodology and for this reason similar conditions are popular in the literature, e.g. [16, 11, 17, 35]. We first present the result on the ABC bias of the log-likelihood. The proof is in Appendix B.

Proposition 2.1. *Assume (A1-3). Then there exist a $C < +\infty$ such that for any $n \geq 1$, $\mu_\theta \in \mathcal{P}(\mathsf{X})$, $\epsilon > 0$, $\theta \in \Theta$ we have:*

$$|l_\theta(y_{1:n}) - l_\theta^\epsilon(y_{1:n})| \leq Cn\epsilon.$$

Remark 2.1. *The above proposition gives some simple guarantees on the bias of the ABC log-likelihood. When using SMC algorithms to approximate* $\log(p_\theta(y_{1:n}))$*, the overall error will be decomposed into the deterministic bias that is present from the ABC approximation (that in Proposition 2.1) and the numerical error of approximating the log-likelihood. Under some assumptions, the $\mathbb{L}_2-$error of the SMC estimate of the log-likelihood should not deteriorate any faster than linearly in time; this is due to the results cited previously. Thus, as the time parameter increases, the ABC bias of the log-likelihood will not necessarily dominate the simulation-based error that would be present even if g_θ is evaluated.*

Proposition 2.1 is reasonably straight-forward to prove, but, is of less interest in the context of parameter estimation, as one is interested in the gradient of the log-likelihood. We now have the result on the ABC bias of the gradient of the log-likelihood. The proof in Appendix C.

Theorem 2.1. *Assume (A1-5). Then there exist a $C < +\infty$ such that for any $n \geq 1$, $\mu_\theta \in \mathcal{P}(\mathsf{X})$, $\widetilde{\mu_\theta} \in \mathcal{M}(\mathsf{X})$, $\epsilon > 0$, $\theta \in \Theta$ we have:*

$$|\nabla l_\theta(y_{1:n}) - \nabla l_\theta^\epsilon(y_{1:n})| \leq Cn\epsilon(2 + \|\widetilde{\mu_\theta}\|).$$

Remark 2.2. *The above Theorem again provides some explicit guarantees when using an ABC approximation along with SMC-based numerical methods. For example, if one can consider approximating gradients in an ABC context as proposed in [38], then from the results of [17], one expects that the variance of the SMC estimates to increase only linearly in time. Again, as time increases the ABC bias does not necessarily dominate the variance that would be present even if g_θ is evaluated (i.e. one uses SMC on the true model).*

Remark 2.3. *The result in Theorem 2.1 can be found in equation (72) of [11] and direct limit (as $\epsilon \searrow 0$) in [12]. However, we adopt a new (and fundamentally different) proof technique, with a substantially more elaborate proof and an additional result of independent interest is proved. We derive the stability of the bias with time of the ABC approximation of the filter derivative; see Theorem D.1 in appendix D.*

Algorithm 1 SMC with ABC

```
• Initialization $t = 0$:
    – For $i = 1, \dots, N$ sample independently $\widetilde{x}_0^{(i)} \sim \mu_\theta$. Set $W_0^{(i)} = 1/N$.
• For $t = 1, \dots, n$
    – Step 1: For $i = 1, \dots, N$, sample next state $x_t^{(i)} \sim q_{t,\theta}(\cdot|\widetilde{x}_{t-1}^{(i)})$
        ∗ For $j = 1, \dots, M$: sample auxiliary observation samples $u_t^{(j,i)} \sim g(\cdot|x_0^i)$
    – Step 2. Compute weights
        $W_t^{(i)} \propto W_{t-1}^{(i)} \widetilde{W}_t^{(i)}, \quad \sum_{i=1}^{N} W_t^{(i)} = 1, \quad \widetilde{W}_t^{(i)} = \frac{\left(\sum_{j=1}^{M} K_\epsilon(y_t, u_t^{(j,i)})\right) f_\theta(x_t^{(i)}|x_{t-1}^{(i)})}{M\, q_{t,\theta}(x_t|x_{t-1}^{(i)})},$
    – Step. 3: If required, resample $N$ particles from
        $\widehat{\pi}_{\theta,\epsilon,t} = \sum_{i=1}^{N} W_t^{(i)} \delta_{x_{0:t}^{(i)}}, \qquad (13)$
      to get $\left\{\widetilde{x}_{0:t}^{(i)}\right\}$ and set $W_t^{(i)} = \frac{1}{N}$, else set $\widetilde{x}_{0:t}^{(i)} = x_{0:t}^{(i)}$.
```

3 Computational Strategy

We begin by considering a modified target instead of the ABC targeted filtering density in (6):

$$\pi_{\theta,\epsilon}(u_{1:n}^1, \dots, u_{1:n}^M, x_{0:n}|y_{1:n}) \propto \mu_\theta(x_0) \prod_{t=1}^{n} \left[\Big(\frac{1}{M} \sum_{j=1}^{M} K_\epsilon(y_t, u_t^j) \Big) \prod_{j=1}^{M} g_\theta(u_t^j|x_t) \right] f_\theta(x_t|x_{t-1}), \qquad (12)$$

where for every t we use this time M independent samples from the likelihood, $u_t^j \sim g_\theta(\cdot|x_t)$, $j = 1, \dots, M$. When one integrates out $u_{1:n}^1, \dots, u_{1:n}^M$ then the targeted sequence is the same as in Section 2.2, which targets a perturbed HMM with the likelihood being $g_{\theta,\epsilon}$ shown earlier in (9). Of course, in terms of estimating θ and MLE, again this yields the same bias as the original ABC approximation, but still there are substantial computational improvements. This is because as M grows we the behavior is closer to an ideal marginal SMC algorithm that targets directly the perturbed HMM without the auxiliary u variables. We proceed by presenting first SMC when the model parameters θ are known and then show how Simultaneous Perturbation Stochastic Approximation (SPSA) can be used for (off-line) gradient-free MLE and RML.

3.1 Sequential Monte Carlo

For the sake of clarity and for this sub-section only consider θ to be fixed and known. In Algorithm 1 we present the ABC-SMC algorithm of [20], which is used to perform filtering for the perturbed HMM with likelihood $g_{\theta,\epsilon}$ and transition density f_θ. The basic design elements are the important sampling proposals $q_{t,\theta}$ for the weights, the number of particles N, the number of auxiliary observation samples M and the ABC precision tolerance ϵ. The resampling step is presented here as optional, but note to get good performance it is necessary to use it when the variance of the weights or the *effective sample size* is low. For more details we refer the reader at [20].

The algorithm allows us to approximate $\pi_{\theta,\epsilon}$ in (12) using the particles. For instance, the particle approximation of the marginal of $\pi_{\theta,\epsilon}$ w.r.t. the u variables is shown in (13). In addition one obtains also particle approximations for $p_{\theta,\epsilon}(y_{1:n})$ and $p_{\theta,\epsilon}(y_t|y_{1:t-1})$ as defined in (7)-(8), which are critical quantities for parameter estimation. So we denote this SMC estimates of these quantities as $p_{\theta,\epsilon}^N(y_{1:n})$ and $p_{\theta,\epsilon}^N(y_t|y_{1:t-1})$ respectively. These are given as follows:

$$p_{\theta,\epsilon}^N(y_{1:n}) = \prod_{t=1}^{n} \frac{1}{N} \sum_{i=1}^{N} \widetilde{W}_t^{(i)}$$

with

$$p^N_{\theta,\epsilon}(y_t|y_{1:t-1}) = \frac{1}{N}\sum_{i=1}^{N}\widetilde{W}^{(i)}_t,$$

where $\widetilde{W}^{(i)}_t$ is defined in Algorithm 1. To avoid possible confusion, we remind the reader that because Z_ϵ in (9) is unknown, one $p_{\theta,\epsilon}(y_{1:n})$ coincides with the actual marginal likelihood of the perturbed HMM only up-to a proportionality constant Z^n_ϵ that is independent of θ. Of course in the context of parameter estimation this does not pose any problems.

The standard SMC approximation for the likelihood $p^N_\theta(y_{1:n})$ is an unbiased estimate in the sense

$$\mathbb{E}^N\left[p^N_\theta(y_{1:n})\right] = p_\theta(y_{1:n}),$$

where $\mathbb{E}^N[\cdot]$ denotes the expectation w.r.t the distribution of all the randomly variables in Algorithm 1. A similar result holds for $p^N_\theta(y_n|y_{1:n-1})$; see [13, Theorems 7.4.2 and 7.4.3, p. 239] for a proof and more details. Note still that $\log\left(p^N_\theta(y_{1:n})\right)$ or $\log\left(p^N_\theta(y_n|y_{1:n-1})\right)$ will be biased approximations of the ideal quantities. A usual remedy is to correct the bias up to the first order of a Taylor expansion and estimate the θ-dependendent parts of $\log\left(p_{\theta,\epsilon}(y_{1:n})\right)$ and $\log\left(p_{\theta,\epsilon}(y_n|y_{1:n-1})\right)$ instead with

$$\hat{l}^N_{\theta,\epsilon} = \log\left(p^N_{\theta,\epsilon}(y_{1:n})\right) + \frac{1}{2N}\left(p^N_{\theta,\epsilon}(y_{1:n})\right)^{-2}, \tag{14}$$

and

$$\hat{r}^N_{t,\theta,\epsilon} = \log\left(\frac{1}{N}\sum_{i=1}^{N}\widetilde{W}^{(i)}_t\right) + \frac{1}{2N}\left(\frac{1}{N}\sum_{i=1}^{N}\widetilde{W}^{(i)}_t\right)^{-2} \tag{15}$$

respectively as suggested in [30].

Remark 3.1. *The parameter ϵ determines the accuracy of the the marginal likelihoods of the perturbed HMM compared to the original one. At the same time if it is very low one may require a high value for M. This can be computed adaptively as in [15, 20]. Also it is remarked that a drawback of this algorithm is that when d_y grows with ϵ, N remaining fixed, one cannot expect the algorithm to work well for every ϵ. Typically one must increase ϵ to get reasonable results with moderate computational effort and this is at the cost of increasing the bias. To maintain ϵ at a reasonable level, one must consider more sophisticated strategies which are not investigated here.*

Remark 3.2. *We note that, after suppressing θ, if the HMM can be written in a state space model form:*

$$\begin{aligned} Y_t &= \xi(X_t, W_t) \\ X_t &= \varphi(X_{t-1}, V_t) \end{aligned}$$

where $X_0 = x_0 \in \mathsf{X}$ is known, both $(V_n)_{n\geq 1}$ and $(W_n)_{n\geq 0}$ are i.i.d. noise sequences independent of each other and ξ, φ appropriate functions. Suppose that one can evaluate:

- *the densities of W_n and V_n and sample from the associated distributions,*
- *ξ and φ point-wise.*

Similar to [29, 39], one can construct a 'collapsed' ABC approximation

$$\pi_\epsilon(w_{1:n}, v_{1:n}, u_{1:n}|y_{1:n}) \propto \prod_{t=1}^{n} K_\epsilon\left(\xi\left(\varphi^{(t)}(x_0, v_{1:t}), w_t\right), \xi\left(\varphi^{(t)}(x_0, v_{1:t}), u_t\right)\right) p(w_t)p(v_t)p(u_t).$$

Hence a version of the SMC algorithm in Figure 1 can be derived which does not need to sample from neither the dynamics of the data nor the transition density of the hidden Markov chain. This representation, however, does not always apply.

Algorithm 2 SPSA for batch ABC-MLE

```
• Initialization $k = 0$. Set $\theta_0$ and choose step size sequences $(a_k)_{k\geq 0}$, $(c_k)_{k\geq 0}$, so that $a_k > 0, a_k, c_k \to 0, \sum_{k\geq 0} a_k = \infty$, $\sum_{k\geq 0} \frac{a_k^2}{c_k^2} < \infty$.
• For $k \geq 0$
    – For $m = 1, \ldots, d_\theta$, sample independently $\Delta_k(m)$ from a Bernoulli distribution with success probability $0.5$ and support $\{-1, 1\}$.
    – Run Algorithm 1 (ABC-SMC) for $\theta_k^+ = \theta_k + c_k\Delta_k$ and $\theta_k^- = \theta_k - c_k\Delta_k$ to obtain $\hat{l}^N_{\theta_k^+,\epsilon}$ and $\hat{l}^N_{\theta_k^-,\epsilon}$ respectively.
    – For $m = 1, \ldots, d_\theta$, update $\theta(m)$
          $\theta_k(m) = \theta_k(m) + a_k \frac{\hat{l}^N_{\theta_k^+,\epsilon} - \hat{l}^N_{\theta_k^-,\epsilon}}{2c_k\Delta_k(m)}.$
```

3.2 Simultaneous Perturbation Stochastic Approximation (SPSA)

We proceed by describing SPSA as a gradient free method for off-line or batch ABC-MLE, which can be found in Algorithm 2. This algorithm does not require one to evaluate g_θ or its gradient. In this context one is interested in estimating θ' such that

$$\nabla l^\epsilon_{\theta'} = 0$$

holds, where we have dropped the dependance on $y_{1:n}$ for simplicity. Recall that here we do not have an expression for ∇l^ϵ_θ to pursue a standard Robbins-Monroe procedure [4]. One way around this would be to use a finite difference approximation to estimate the gradient w.r.t. to the m-th element of θ as $\frac{\hat{l}^\epsilon_{\theta+ce_m} - \hat{l}^\epsilon_{\theta-ce_m}}{2c}$, where e_m is a unit magnitude vector that is zero in any direction except m and $\hat{l}^\epsilon_\bullet$ an unbiased estimate of $l^\epsilon_\bullet$. To avoid having to do $2d_\theta$ evaluations of these estimates in total for each direction, SPSA has been proposed in [33] so that the gradient update requires only 2 evaluations only. Instead we perturb θ using $c_k\Delta_k$ where Δ_k is a $d_\theta-$dimensional zero mean vector, such that $\mathbb{E}\left[|\Delta_k(m)|^{-1}\right]$ or some higher inverse moment is bounded. In this case we have used the most popular choice with each entry of Δ_k being ± 1 Bernoulli distributed and the estimates for the $\hat{l}^\epsilon_\bullet$ are the bias-corrected versions as in equation (14). For more details on the conditions and the convergence details for this Stochastic Approximation method we refer the reader to [33] and for useful practical suggestions regarding the implementation to [34].

3.2.1 Recursive ML with SPSA

Recall from (11) in Section 2.2 that the ABC-RML recursion for the parameter is given as

$$\theta_{t+1} = \theta_t + a_{t+1}\nabla r^\epsilon_{\theta_{0:t}}(y_{1:t}).$$

In Algorithm 3 we illustrate how this can be implemented using ABC-SMC. We have extended the RML procedure using both SMC and SPSA that appeared in [32] for the case where ABC approximations are used due to the intractability of $\log(g_\theta(y|x)), \nabla \log(g_\theta(y|x))$. In [39] one can find an alternative approach which implements RML and ABC that does not require to use SPSA. Although a direct comparison is beyond the scope of this paper, we expect the method in [39] to be more accurate. On the other hand, Algorithm 3 can be applied possibly to a wider class of models, but the use of SPSA means that we add an additional layer of approximation and there is a possibility of biases incurring that need to be investigated more thoroughly.

4 Numerical Simulations

We consider two numerical examples that are designed to investigate the accuracy and behavior of our numerical ABC-MLE algorithms. In order to do this, we consider scenarios where g_θ is a well behaved density, which we avoid to compute. In the first example we look at a linear Gaussian model and in the second a HMM involving the Lorenz '63 model [25].

Algorithm 3 RML with ABC-SMC

```
• Initialization $t = 0$:
    – Set $\theta_1$ and choose step size sequences $(a_t)_{t\geq 0}$, $(c_t)_{t\geq 0}$, so that $a_t > 0$, $a_t, c_t \to 0$, $\sum_{t\geq 0} a_t = \infty$, $\sum_{t\geq 0} \frac{a_t^2}{c_t^2} < \infty$.
    – For $i = 1, \ldots, N$ sample independently $\widetilde{x}_0^{(i)} \sim \mu_\theta$. Set $W_0^{(i)} = 1/N$.
• For $t = 1, \ldots, n$
    – For $m = 1, \ldots, d_\theta$, sample independently $\Delta_t(m)$ from a Bernoulli distribution with success probability $0.5$ and support $\{-1, 1\}$.
    – Set $\theta_t^+ = \theta_t + c_t\Delta_t$ and $\theta_t^- = \theta_t - c_t\Delta_t$. For each value use $\left\{\left(\widetilde{x}_{0:t-1}^{(i)}, W_{t-1}^{(i)}\right)\right\}$ to compute Steps 1 and 2 of Algorithm 1 (ABC-SMC) returning $\left\{\widetilde{W}_t^{(i)}(\theta_t^+), W_t^{(i)}(\theta_t^+)\right\}$ and $\left\{\widetilde{W}_t^{(i)}(\theta_t^-), W_t^{(i)}(\theta_t^-)\right\}$ respectively.
    – Compute $\hat{r}^N_{t,\theta_t^+,\epsilon}$ and $\hat{r}^N_{t,\theta_t^-,\epsilon}$ respectively using (15).
    – Update $\theta_t$. For $m = 1, \ldots, d_\theta$
          $\theta_{t+1}(m) = \theta_t(m) + a_t \frac{\hat{r}^N_{t,\theta_t^+,\epsilon} - \hat{r}^N_{t,\theta_t^-,\epsilon}}{2c_t\Delta_t(m)}.$
    – Compute Steps 1 to 3 of Algorithm 1 (ABC-SMC) using $\theta_{t+1}$ to get $\left\{\widetilde{x}_{0:t}^{(i)}, W_{t-1}^{(i)}\right\}$.
```

4.1 Linear Gaussian Model

We consider the following linear Gaussian HMM, with $\mathsf{Y} = \mathsf{X} = \mathbb{R}$, $t \geq 1$:

$$\begin{aligned} Y_t &= X_t + \sigma_w W_t \\ X_t &= \phi X_{t-1} + \sigma_v V_t, \end{aligned}$$

with W_t, V_t independent and $W_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$, $V_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$. In the subsequent examples, we will use a simulated dataset obtained with $\theta = (\sigma_v, \phi, \sigma_w) = (0.2, 0.9, 0.3)$, which is the same example as in [32].

4.1.1 Batch MLE

We begin by considering a short data set, of $n = 1000$ data points. The off-line scenario is the one for which we can expect the best possible performance of the ABC-MLE. If one could not obtain reasonable parameter estimates in this example one would not expect ABC to be very useful in practice. In Algorithm 3.2 recall $\Delta_k(m)$ is the m^{th}-entry of the ±1, zero mean Bernoulli variable and for the step-sizes we chose $c_k = k^{-0.1}$, $a_k = 1$ for $k < 10^4$, and $a_k = (k - 10^4)^{-0.8}$ for $k \geq 10^4$. In Figure 1, we compare offline ML estimates of the following cases:

1. Kalman Filtering (KF) for the original HMM is used to compute $\hat{l}_\theta$ for SPSA,

2. Standard SMC (without ABC) with $N = 1000$ for the original HMM is used to compute $\hat{l}_\theta$ for SPSA,

3. ABC-SMC with $N = 200$, $M = 10$, $\epsilon = 0.1$is used to compute $\hat{l}_\theta^\epsilon$ for SPSA.

The horizontal lines in Figure 1 show also Maximum Likelihood estimates (MLE) obtained from an offline grid search optimization that uses KF. All procedures seem to be very accurate at estimating the MLE obtained from the grid search. This allows us to investigate RML, which is a more challenging problem.

4.1.2 RML

We now consider a larger data set with $n = 50,000$ data points, simulated with the previously indicated parameter values. We use Algorithm 3 described in Section 3.2. Again we compare the same three procedures outlines above using fifty independent runs in each case. The standard SMC and ABC-SMC algorithms were employed with the

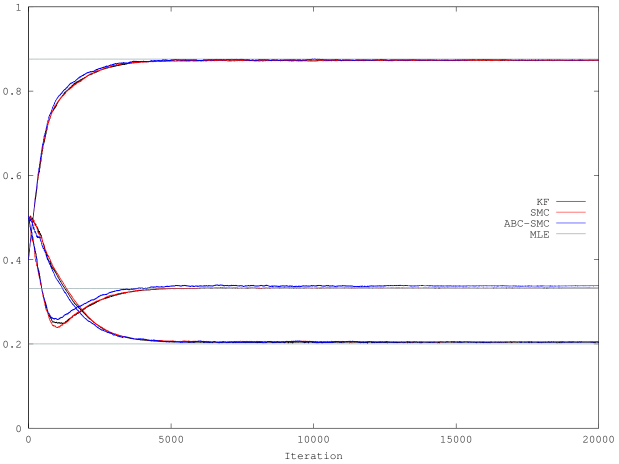
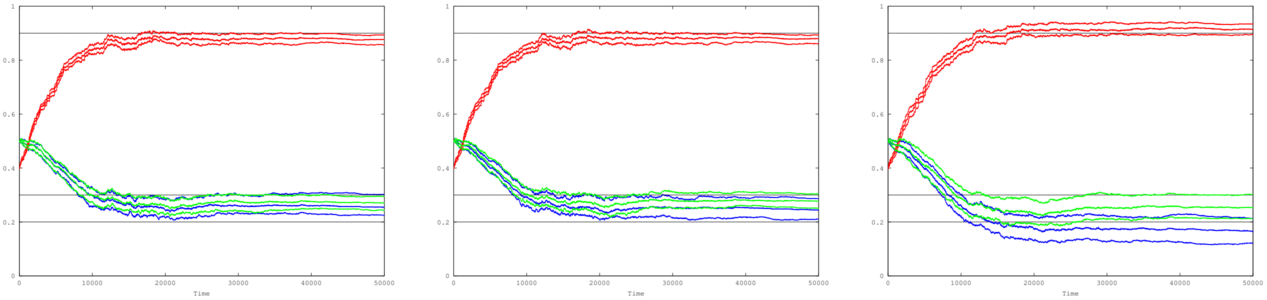


Figure 1: A typical run of the offline parameter estimates obtained by the KF, SMC, and ABC-SMC for the linear Gaussian HMM, along with the ML estimators for θ.

same for N and M, ϵ as in the off-line case. Also for each case we used the same the step-size sequences for SPSA, which were similar to their off-line counterparts in Section 4.1.1. In Figure 2, we plot the medians and credible intervals for the 5-95% percentiles of the parameter estimates (across the independent runs). The $\widehat{\theta}_t$ converge after $t = 20000$ time steps, with the KF and SMC yielding similarly valued estimates. Note there seems to be an apparent bias in both cases relative to the true parameters (the MLE for the data-set used has been checked that it converges to the true parameters by $n = 5 \times 10^4$). A similar bias has appeared in [32] for this particular model. The theoretical justification in [33] applies directly when SPSA is used for off-line MLE (as in Section 4.1.1) with a finite and fixed data-set. For RML the argument to be maximized is an ergodic average [23, 24, 35, 36], so we believe the bias accumulated here is due to the step-sizes of SPSA decreasing much faster than the gradient to be estimated reaches stationarity. Ideally, one would like to run this algorithm for a much longer n, slower decreasing step-sizes and also delay updating θ until stationarity is reached, but this would make using multiple runs prohibitive. In [32] it seemed that this bias was not considerable for other models, such as the popular stochastic volatility model. In any case, it would be useful to examine precisely under what conditions SPSA can be used within RML, but this is beyond the scope of this paper that puts more emphasis on the relative accuracy of ABC. In Figure 2 we also observe increased variance from left to right in Figure 2, which we attribute to the progressively added randomness of SMC and ABC-SMC respectively. In particular, the expected reduced accuracy of ABC-SMC against SMC is apparent, but, the bias does not appear to be substantial (for ABC-SMC) in this particular example.



(a) Kalman

(b) Sequential Monte Carlo

(c) SMC-ABC

Figure 2: Credible intervals for the 5-95% percentiles and the medians after multiple runs of parameter estimates using RML with KF, SMC, and ABC-SMC for the linear Gaussian HMM.

4.2 Lorenz '63 Model

4.2.1 Model and Data

We now consider the following non-linear state-space model with $\mathsf{X} = \mathsf{Y} = \mathbb{R}^3$. The original model is such that hidden process evolves deterministically according to the Lorenz '63 system of ordinary differential equations,

$$\begin{aligned}\dot{X}(1) &= \sigma_{63}\big(X(2) - X(1)\big)\\ \dot{X}(2) &= \rho X(1) - X(2) - X(1)X(3)\\ \dot{X}(3) &= X(1)X(2) - \beta X(3).\end{aligned}$$

where $X(m), \dot{X}(m)$ are the $m^{th}-$components of the state and velocity at any time respectively. We discretize the model to a discrete-time Markov chain with dynamics:

$$X_t = f_t(X_{t-1}) + V_t, \quad t \geq 0$$

where f_t is the 4^{th}-order approximation Runge-Kutta approximation of the Lorenz '63 system, $V_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \tau I_{d_x})$ and X_0 is taken as known. Here τ is used to represent the time-discretization.

For the observations we use:

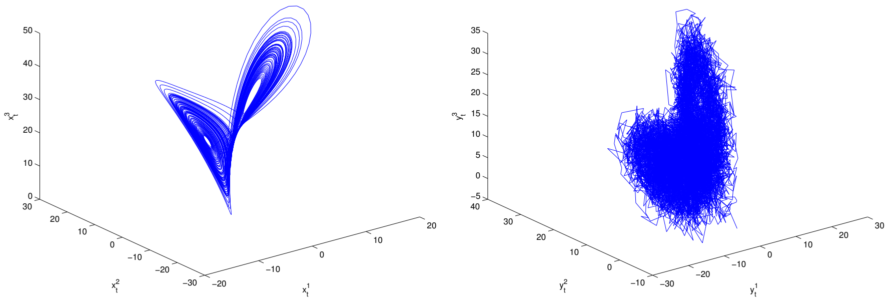
$$Y_t = HX_t + QW_t, \quad t \geq 1$$

where $W_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I_{d_y})$, W_t is independent of V_t and Q is the Cholesky root of a Toeplitz matrix defined by the parameters κ and σ as follows:

$$\begin{aligned}Q_{ij} &= \sigma S\left(\kappa^{-1}\min(|i-j|, d_y - |i-j|)\right), \quad i,j \in \{1,\ldots,d_y\}\\ S(z) &= \left\{\begin{array}{ll} 1 - \frac{3}{2}z + \frac{1}{2}z^3, & 0 \leq z \leq 1\\ 0, & z > 1\end{array}\right. ,\end{aligned}$$

and

$$H_{ij} = \left\{\begin{array}{ll} \frac{1}{2}, & i = j\\ \frac{1}{2}, & i = j-1\\ 0, & i \neq j\end{array}\right. .$$



(a) Hidden Markov chain

(b) Observations

Figure 3: Evolution of the 3-dimensional Lorenz '63 HMM with $n = 5000$.

When $\theta = (\kappa, \sigma, \sigma_{63}, \rho, \beta) = (2.5, 2, 10, 28, \frac{8}{3})$, $n = 5000$ and $\tau = 0.05$, a visualization of the Lorenz '63 (hidden) dynamics is shown in Figure 3(a) and the associated simulated dataset in 3(b).

For the simulated data-set in Figure 3(b) and its extension for longer n, in the remainder we will use ABC-SMC to obtain parameter estimates from RML. In the subsequent sub-section we will study the performance of these estimates under different settings. We will use $\widehat{\theta}_{\epsilon,n}^{N,M}$ to denote the estimate of θ at time n, that was estimated using N particles, M pseudo-observations and a Gaussian kernel with covariance ϵI_{d_y}. We will compare the behavior of the algorithm as each of N, M, n, ϵ varies.

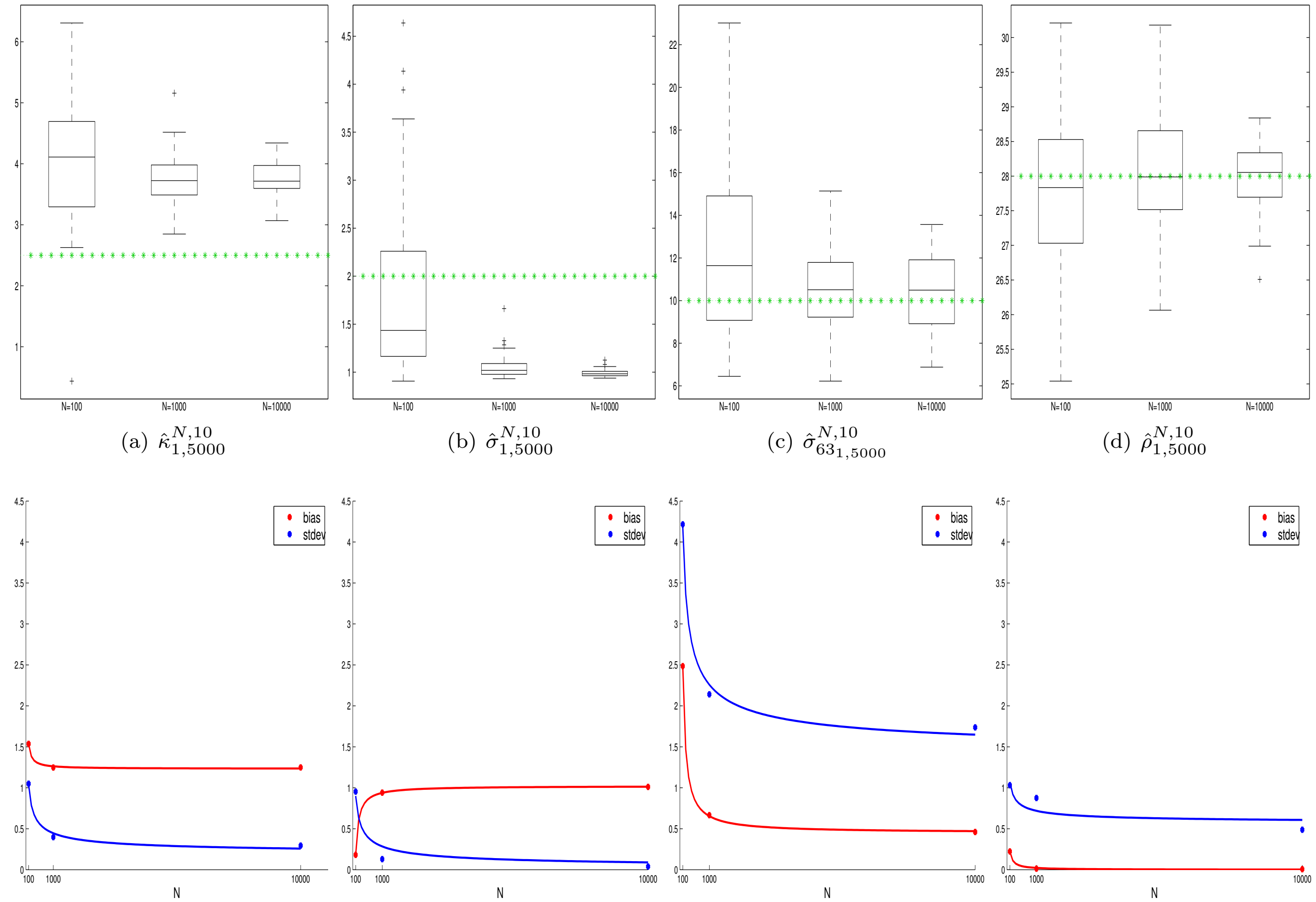
4.2.2 Numerical Results

We now examine the performance of the algorithm for $N \in \{100, 1000, 10000\}$. For each value of N, we ran fifty independent runs of ABC-SMC, using $M = 10$ and $\epsilon = 1$. In Figures 4(a)-4(d) we plot box-plots of the terminal parameter estimates, $\widehat{\theta}_{1,5000}^{N,10}$, against their true values marked by dotted green lines. In Figures 4(e)-4(h) we plot the absolute value of the Monte Carlo (MC) bias (that is, the absolute difference between the estimate and true value), in red, and the MC standard deviation, in blue. The MC bias and standard deviation points are fitted with least-squares curves proportional to $\frac{1}{\sqrt{N}}$, the standard MC rates with which the accuracy of the estimates is expected to improve. With regards to the variability of the estimates one sees the expected reduction in variability as N increases. The bias is harder to quantify; it will not necessarily be the case that as N grows the bias falls. This is because there is a Monte Carlo bias (from the SMC), an optimization bias (from the SPSA), an approximation bias (from the ABC). Increasing N can only deal with the SMC bias (which for estimates with parameters fixed is $\mathcal{O}(N^{-1})$), but the addition of parameter estimation complicates things here. The main point is that as expected one obtains significantly more reproducible/consistent results as N grows.

Next we look at the influence of the number of auxiliary observations samples. For $M \in \{1, 3, 5, 10, 25, 50\}$, we show in Figures 5(a)-5(d) the box-plots of the terminal estimates $\widehat{\theta}_{1,5000}^{5000,M}$ from fifty independent runs of ABC-SMC, using $N = 5000$ and $\epsilon = 1$. The dotted green lines marks the true θ values which generate the data. In Figures 5(e)-5(h), the MC biases and the MC standard deviations of the $\widehat{\theta}_{1,5000}^{5000,M}$ are plotted as discrete points, in red and blue, with lines of least squared-error fitted around them. As M increases, we see reductions in the MC variance. This reduction in variance can be attributed to the fact that the ABC-SMC algorithm approximates the ideal SMC algorithm that targets the perturbed HMM. Hence by a Rao-Blackwellization type argument, one expects a reduction in variance. These results are consistent with [15]. For this example, after $M \geq 5$, there seems to be little impact on the accuracy of the parameter estimates, but this is example specific.

We now vary n. For $n \in \{5000, 10,000, 15,000\}$ we ran fifty independent runs of ABC-SMC using $N = 200$, $M = 10$, and $\epsilon = 1$, and plotted box-plots of the terminal estimates $\widehat{\theta}_{1,n}^{200,10}$, in Figures 6(a)-6(d), against the true values of θ marked in dotted green lines. Recall that RML estimation tries to maximize $\frac{1}{n}\log(p_{\theta,\epsilon}(y_{1:n}))$, so we expect n not to have a great effect on the bias nor the variance when it is above some value. This can also be explained by the bias results in Section 2.3 and the theoretical results in [11, 12]. In Figures 6(e)-6(h) the absolute value of the MC biases and the MC standard deviations have been plotted in red and blue, and fitted with linear lines of least squared-error.

Finally, we investigate the influence of $\epsilon \in \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 50\}$. For each ϵ, we again ran fifty independent runs of ABC-SMC with $N = 200$ and $M = 10$, for the dataset $n = 5000$. The box-plot of the parameter estimates are plotted, in Figures 7(a)-7(d), against dotted green lines which indicate the true θ. Figures 7(e)-7(h) show the



(a) $\hat{\kappa}^{N,10}_{1,5000}$

(b) $\hat{\sigma}^{N,10}_{1,5000}$

(c) $\hat{\sigma}63^{N,10}_{1,5000}$

(d) $\hat{\rho}^{N,10}_{1,5000}$

(e) $\hat{\kappa}^{N,10}_{1,5000}$

(f) $\hat{\sigma}^{N,10}_{1,5000}$

(g) $\hat{\sigma}63^{N,10}_{1,5000}$

(h) $\hat{\rho}^{N,10}_{1,5000}$

Figure 4: $\widehat{\theta}^{N,10}_{1,5000}$ when estimating $\theta = (\kappa, \sigma, \sigma_{63}, \rho)$ of the Lorenz '63 HMM, using ABC-SMC with values of $N \in \{100, 1000, 10000\}$. Figures 4(a)-4(d) show the $\widehat{\theta}^{N,10}_{1,5000}$ in box-plots and their true values in dotted green lines. Figures 4(e)-4(h) show the MC bias and MC standard deviation of the $\widehat{\theta}^{N,10}_{1,5000}$, in red and blue, with curves of least squared-error $\propto \frac{1}{\sqrt{N}}$.

absolute value of MC biases in red, and the MC standard deviations in blue. Fitted to the MC biases is a non-linear least squares curve proportional to $\epsilon + \frac{1}{\epsilon}$. The result we presented in Section 2.3 states that as ϵ increases, the bias will increase on $\mathcal{O}(\epsilon)$, hence the term proportional to ϵ of the fitted curve. However, the ABC-SMC algorithm becomes less stable for ϵ too small (in the sense that, for example, the variance of the weights will become larger as ϵ grows), incurring more varied estimates. We conjecture this will affect biases according to a term proportional to $\frac{1}{\epsilon}$. Similarly, we fitted to the MC standard deviations non-linear least squares curves proportional to $\frac{1}{\epsilon}$ and note that the MC standard deviation decreases at this rate as ϵ increases.

5 Conclusions

In this article we have presented how to perform ML parameter estimation using ABC-SMC and SPSA for HMMs. For batch MLE the method appears to be very accurate when a well-selected step-sized is used. In the on-line case and RML the method again appears to be sensitive to the tuning of the step-sizes and for moderately long runs one should expect a bias, which in the examples here and in[32] seems small but not negligible. We believe this bias is due to using SPSA within another Stochastic Approximation algorithm, i.e. the RML. A theoretical investigation of identifying the source of this bias should be an interesting extension our work. Furthermore, except the obvious case when the likelihood in the HMM is intractable, these ideas could be also useful for models where the parameter and observations are of moderate dimension and the state-dimension is high. Such models have wide application in data assimilation and numerical weather prediction. In addition, the work related here is closely related to [39], where following a representation similar to Remark 3.2, the authors provide a RML algorithm without using SPSA and also show how on-line Expectation-Maximization techniques like [16, 38] are relevant for ABC-MLE for HMMs. We conclude by mentioning that current ongoing work is trying to address the limitations in efficiency of the presented ABC-SMC algorithm when small ϵ is used. Two potential ways to proceed can be 1) to introduce approximations by the expectation-propagation algorithm in [3] and potentially removing SMC and 2) to consider combining ABC with more advanced SMC approaches such as [14] to allow use of much lower ϵ.

Acknowledgements

We thank the referee for comments that have vastly improved the paper. We also acknowledge useful discussions on this material with Sumeetpal Singh. The second author was funded by an MOE grant and acknowledges useful conversations with David Nott. The third author acknowledges support from EPSRC under grant EP/J01365X/1 since July 2012 and under the programme grant on Control For Energy and Sustainability EP/G066477/1 during earlier stages of this work when he was employed at Imperial College.

A Notations

We will introduce a round of notations. Firstly, we alert the reader that throughout appendix k is used as a time index instead of t used earlier. As our analysis will rely upon that in [35] our notations will follow that article. It is remarked that under our assumptions, one can establish the same assumptions as in [35]. Moreover, the time-inhomogenous upper-bounds in that paper can be made time-homogenous (albeit less tight) under our assumptions. In addition, our proof strategy follows ideas in the expanded technical report of [1].

$\mathcal{B}_b(\mathsf{X})$ is the class of bounded and real-valued measurable functions on X. Throughout, for $\varphi \in \mathcal{B}_b(\mathsf{X})$, $\|\varphi\|_\infty := \sup_{x\in\mathsf{X}} |\varphi(x)|$. For $\varphi \in \mathcal{B}_b(\mathsf{X})$ and any operator $Q : \mathsf{X} \to \mathcal{M}(\mathsf{X})$, $Q(\varphi)(x) := \int_{\mathsf{X}} \varphi(y)Q(x, dy)$. In addition for $\mu_\theta \in \mathcal{M}(\mathsf{X})$, $\mu_\theta Q(\varphi) := \int_{\mathsf{X}} \mu_\theta(dx)Q(\varphi)(x)$.

We introduce the non-negative operator:

$$R_{\theta,n}(x, dx') := g_\theta(y_n|x')f_\theta(x'|x)dx'$$

with the ABC equivalent $R_{\theta,\epsilon,n}(x, dx') := g_{\theta,\epsilon}(y_n|x')f_\theta(x'|x)dx'$, $g_{\theta,\epsilon}(y|x) = \int_{A_{\epsilon,y}} g(u|x)dy/\int_{A_{\epsilon,y}} dy$. To keep consistency with [35] and to allow the reader to follow the proofs, we note that the filter at time $n \geq 0$, $F_\theta^n(\mu_\theta)$ (respectively ABC filter, at time n, $F_{\theta,\epsilon}^n(\mu_\theta)$) is exactly, with initial distribution $\mu_\theta \in \mathcal{P}(\mathsf{X})$ and test function $\varphi \in \mathcal{B}_b(\mathsf{X})$

$$F_\theta^n(\mu_\theta)(\varphi) = \frac{\mu_\theta R_{1,n,\theta}(\varphi)}{\mu_\theta R_{1,n,\theta}(1)}$$

respectively

$$F_{\theta,\epsilon}^n(\mu_\theta)(\varphi) = \frac{\mu_\theta R_{1,n,\theta,\epsilon}(\varphi)}{\mu_\theta R_{1,n,\theta,\epsilon}(1)}$$

where $F_\theta^0(\mu_\theta) = F_{\theta,\epsilon}^0(\mu_\theta) = \mu_\theta$, $R_{1,n,\theta}(\varphi)(x_0) = \int \prod_{k=1}^n R_{k,\theta}(x_{k-1}, dx_k)\varphi(x_n)$. In addition, we write the filter derivatives as $\widetilde{F}_\theta^n(\mu_\theta, \widetilde{\mu_\theta})(\varphi)$, $\widetilde{F}_{\theta,\epsilon}^n(\mu_\theta, \widetilde{\mu_\theta})(\varphi)$ where the second argument is the gradient of the initial measure.

The following operators will be used below, for $n \geq 1$:

$$\begin{aligned} \widetilde{G}^n(\mu_\theta, \widetilde{\mu_\theta})(\varphi) &:= (\mu_\theta R_{1,n,\theta}(1))^{-1}[\widetilde{\mu_\theta} R_{1,n,\theta}(\varphi) - \widetilde{\mu_\theta} R_{1,n,\theta}(1)F_\theta^n(\mu_\theta)(\varphi)] &(16)\\ \widetilde{H}^n(\mu_\theta)(\varphi) &:= F_\theta^{n-1}(\mu_\theta)R_{n,\theta}(1)^{-1}[F_\theta^{n-1}(\mu_\theta)\widetilde{R}_{n,\theta}(\varphi) - F_\theta^{n-1}(\mu_\theta)\widetilde{R}_{n,\theta}(1)F_\theta^n(\mu_\theta)(\varphi)] &(17)\end{aligned}$$

with the convention $\widetilde{G}^0(\mu_\theta, \widetilde{\mu_\theta})(\varphi) = \widetilde{\mu_\theta}$. In addition, we set

$$\widetilde{G}^{(n)}(\mu_\theta, \widetilde{\mu_\theta})(\varphi) := (\mu_\theta R_{n,\theta}(1))^{-1}[\widetilde{\mu_\theta} R_{n,\theta}(\varphi) - \widetilde{\mu_\theta} R_{n,\theta}(1)F_\theta^{(n)}(\mu_\theta)(\varphi)].$$

where $F_\theta^{(n)}(\mu_\theta) = \mu_\theta R_{n,\theta}/\mu_\theta R_{n,\theta}(1)$. Finally, an important notational convention is as follows. Throughout we use C to denote a constant whose value may change from line-to-line in the calculations. This constant will typically not depend upon important parameters such as ϵ and n and any important dependencies will be highlighted.

B Bias of the Log-Likelihood

Proof. [Proof of Proposition 2.1] We begin with the equality

$$\log(p_\theta(y_{1:n})) - \log(p_{\theta,\epsilon}(y_{1:n})) = \sum_{k=1}^{n}\Big(\log(p_\theta(y_k|y_{1:k-1})) - \log(p_{\theta,\epsilon}(y_k|y_{1:k-1}))\Big) \tag{18}$$

with, for $1 \leq k \leq n$

$$\begin{aligned} p_\theta(y_k|y_{1:k-1}) &= \int_{\mathsf{X}^2} g_\theta(y_k|x_k) f_\theta(x_k|x_{k-1}) F_\theta^{k-1}(\mu_\theta)(dx_{k-1})dx_k \\ p_{\theta,\epsilon}(y_k|y_{1:k-1}) &= \int_{\mathsf{X}^2} g_\theta^\epsilon(y_k|x_k) f_\theta(x_k|x_{k-1}) F_{\theta,\epsilon}^{k-1}(\mu_\theta)(dx_{k-1})dx_k. \end{aligned}$$

We will consider each summand in (18). The case $k \geq 2$ is only considered; the scenario $k = 1$ will follow a similar and simpler argument.

Using the inequality $|\log(x) - \log(y)| \leq |x-y|/(x \wedge y)$ for every $x, y > 0$ we have

$$|\log(p_\theta(y_k|y_{1:k-1})) - \log(p_{\theta,\epsilon}(y_k|y_{1:k-1}))| \leq \frac{|p_\theta(y_k|y_{1:k-1}) - p_{\theta,\epsilon}(y_k|y_{1:k-1})|}{p_\theta(y_k|y_{1:k-1}) \wedge p_{\theta,\epsilon}(y_k|y_{1:k-1})}.$$

Note that

$$p_\theta(y_k|y_{1:k-1}) \wedge p_\theta(y_k|y_{1:k-1}) =$$

$$\int_{\mathsf{X}^2} g_\theta(y_k|x_k) f_\theta(x_k|x_{k-1}) F_\theta^{k-1}(\mu_\theta)(dx_{k-1})dx_k \wedge \int_{\mathsf{X}^2} g_\theta^\epsilon(y_k|x_k) f_\theta(x_k|x_{k-1}) F_{\theta,\epsilon}^{k-1}(\mu_\theta)(dx_{k-1})dx_k \geq C > 0 \tag{19}$$

where we have applied (A3) and C does not depend upon ϵ. Thus we consider

$$|p_{\theta,\epsilon}(y_k|y_{1:k-1}) - p_\theta(y_k|y_{1:k-1})| =$$

$$|\int_{\mathsf{X}^2} g_\theta(y_k|x_k) f_\theta(x_k|x_{k-1}) F_\theta^{k-1}(\mu_\theta)(dx_{k-1})dx_k - \int_{\mathsf{X}^2} g_{\theta,\epsilon}(y_k|x_k) f_\theta(x_k|x_{k-1}) F_{\theta,\epsilon}^{k-1}(\mu_\theta)(dx_{k-1})dx_k|.$$

The R.H.S. can be upper-bounded by the sum of

$$|\int_{\mathsf{X}^2} [g_\theta(y_k|x_k) - g_{\theta,\epsilon}(y_k|x_k)] f_\theta(x_k|x_{k-1}) F_\theta^{k-1}(\mu_\theta)(dx_{k-1})dx_k|$$

and

$$|\int_{\mathsf{X}^2} g_{\theta,\epsilon}(y_k|x_k) f_\theta(x_k|x_{k-1}) [F_{\theta,\epsilon}^{k-1}(\mu_\theta)(dx_{k-1}) - F_{\theta,\epsilon}^{k-1}(\mu_\theta)(dx_{k-1}])dx_k|.$$

The first expression can be dealt with by using (A1), which implies

$$\sup_{x\in\mathsf{X}} |g_{\theta,\epsilon}(y_k|x) - g_{\theta,\epsilon}(y_k|x)| \leq C\epsilon. \tag{20}$$

The second expression can be controlled by [20, Theorem 2]:

$$\sup_{k\geq 1} \|F_\theta^{k-1}(\mu_\theta) - F_{\theta,\epsilon}^{k-1}(\mu_\theta)\| \leq C\epsilon \tag{21}$$

to yield that

$$|p_{\theta,\epsilon}(y_k|y_{1:k-1}) - p_\theta(y_k|y_{1:k-1})| \leq C\epsilon. \tag{22}$$

One can thus conclude. □

C Bias of the Gradient of the Log-Likelihood

Proof. [Proof of Theorem 2.1] We have that

$$\nabla\Big(\log p_\theta(y_{1:n}) - \log p_{\theta,\epsilon}(y_{1:n})\Big) = \nabla\Big\{\sum_{k=1}^n\Big(\log[p_\theta(y_k|y_{1:k-1}) - \log[p_{\theta,\epsilon}(y_k|y_{1:k-1})\Big)\Big\}.$$

It then follows that

$$\nabla\Big(\log p_\theta(y_{1:n}) - \log p_{\theta,\epsilon}(y_{1:n})\Big) =$$

$$\sum_{k=1}^n\Big(\frac{[\nabla p_\theta(y_k|y_{1:k-1}) - \nabla p_{\theta,\epsilon}(y_k|y_{1:k-1})]}{p_\theta(y_k|y_{1:k-1})} + \frac{\nabla p_{\theta,\epsilon}(y_k|y_{1:k-1})}{p_\theta(y_k|y_{1:k-1})p_{\theta,\epsilon}(y_k|y_{1:k-1})}[p_{\theta,\epsilon}(y_k|y_{1:k-1}) - p_\theta(y_k|y_{1:k-1})]\Big). \quad (23)$$

We will deal with the two terms on the R.H.S. of (23) in turn. The scenario $k \geq 2$ is only considered; the case $k = 1$ follows a similar and simpler argument.

First starting with summand

$$\frac{[\nabla p_\theta(y_k|y_{1:k-1}) - \nabla p_{\theta,\epsilon}(y_k|y_{1:k-1})]}{p_\theta(y_k|y_{1:k-1})}.$$

Noting (19), we need only upper-bound the $\mathbb{L}_1$ norm of the following expression

$$\int_{\mathsf{X}^2}\nabla\{g_\theta(y_k|x_k)\}f_\theta(x_k|x_{k-1})F_\theta^{k-1}(\mu_\theta)(dx_{k-1})dx_k - \int_{\mathsf{X}^2}\nabla\{g_{\theta,\epsilon}(y_k|x_k)\}f_\theta(x_k|x_{k-1})F_{\theta,\epsilon}^{k-1}(\mu_\theta)(dx_{k-1})dx_k \quad (24)$$

$$+\int_{\mathsf{X}^2}g_\theta(y_k|x_k)\nabla\{f_\theta(x_k|x_{k-1})\}F_\theta^{k-1}(\mu_\theta)(dx_{k-1})dx_k - \int_{\mathsf{X}^2}g_{\theta,\epsilon}(y_k|x_k)\nabla\{f_\theta(x_k|x_{k-1})\}F_{\theta,\epsilon}^{k-1}(\mu_\theta)(dx_{k-1})dx_k \quad (25)$$

$$+\int_{\mathsf{X}^2}g_\theta(y_k|x_k)f_\theta(x_k|x_{k-1})\widetilde{F}_\theta^{k-1}(\mu_\theta,\widetilde{\mu_\theta})(dx_{k-1})dx_k - \int_{\mathsf{X}^2}g_{\theta,\epsilon}(y_k|x_k)f_\theta(x_k|x_{k-1})\widetilde{F}_{\theta,\epsilon}^{k-1}(\mu_\theta,\widetilde{\mu_\theta})(dx_{k-1})dx_k. \quad (26)$$

We start with (24). Using (A4) we can establish that for each $k \geq 1$

$$\sup_{x\in\mathsf{X}}|\nabla\{g_\theta(y_k|x_k)\} - \nabla\{g_{\theta,\epsilon}(y_k|x_k)\}| \leq C\epsilon \quad (27)$$

where C does not depend upon k, ϵ. Hence

$$|\int_{\mathsf{X}^2}[\nabla\{g_\theta(y_k|x_k)\} - \nabla\{g_{\theta,\epsilon}(y_k|x_k)\}]f_\theta(x_k|x_{k-1})F_\theta^{k-1}(\mu_\theta)(dx_{k-1})dx_k| \leq C\epsilon.$$

Then we note that by [20, Theorem 2] (see (21)) and (A5)

$$|\int_{\mathsf{X}^2}\nabla\{g_{\theta,\epsilon}(y_k|x_k)\}f_\theta(x_k|x_{k-1})[F_\theta^{k-1}(\mu_\theta)(dx_{k-1}) - F_{\theta,\epsilon}^{k-1}(\mu_\theta)(dx_{k-1})]dx_k| \leq C\epsilon$$

Thus we have shown that

$$|\int_{\mathsf{X}^2}\nabla\{g_\theta(y_k|x_k)\}f_\theta(x_k|x_{k-1})F_\theta^{k-1}(\mu_\theta)(dx_{k-1})dx_k - \int_{\mathsf{X}^2}\nabla\{g_{\theta,\epsilon}(y_k|x_k)\}f_\theta(x_k|x_{k-1})F_{\theta,\epsilon}^{k-1}(\mu_\theta)(dx_{k-1})dx_k| \leq C\epsilon.$$

Now, moving onto (25), by (20) we have

$$|\int_{\mathsf{X}^2}[g_\theta(y_k|x_k) - g_{\theta,\epsilon}(y_k|x_k)]\nabla\{f_\theta(x_k|x_{k-1})\}F_\theta^{k-1}(\mu_\theta)(dx_{k-1})dx_k| \leq C\epsilon.$$

and can again use [20, Theorem 2] (i.e. (21)) to deduce that

$$|\int_{\mathsf{X}^2}g_{\theta,\epsilon}(y_k|x_k)\nabla\{f_\theta(x_k|x_{k-1})\}[F_\theta^{k-1}(\mu_\theta)(dx_{k-1}) - F_{\theta,\epsilon}^{k-1}(\mu_\theta)(dx_{k-1})]dx_k| \leq C\epsilon$$

and thus that

$$|\int_{\mathsf{X}^2} g_\theta(y_k|x_k)\nabla\{f_\theta(x_k|x_{k-1})\}F_\theta^{k-1}(\mu_\theta)(dx_{k-1})dx_k - \int_{\mathsf{X}^2} g_{\theta,\epsilon}(y_k|x_k)\nabla\{f_\theta(x_k|x_{k-1})\}F_{\theta,\epsilon}^{k-1}(\mu_\theta)(dx_{k-1})dx_k| \leq C\epsilon$$

which upper-bounds the expression in (25). We now move onto (26), which upper-bounded by

$$|\int_{\mathsf{X}^2}[g_\theta(y_k|x_k) - g_{\theta,\epsilon}(y_k|x_k)]f_\theta(x_k|x_{k-1})\widetilde{F}_\theta^{k-1}(\mu_\theta,\widetilde{\mu_\theta})(dx_{k-1})dx_k| +$$

$$|\int_{\mathsf{X}^2} g_{\theta,\epsilon}(y_k|x_k)f_\theta(x_k|x_{k-1})[\widetilde{F}_\theta^{k-1}(\mu_\theta,\widetilde{\mu_\theta})(dx_{k-1}) - \widetilde{F}_{\theta,\epsilon}^{k-1}(\mu_\theta,\widetilde{\mu_\theta})(dx_{k-1})]dx_k|.$$

For the first expression, we can write:

$$(\sup_{x\in\mathsf{X}}|g_\theta(y_k|x) - g_{\theta,\epsilon}(y_k|x)|)|\int_{\mathsf{X}}\left(\int_{\mathsf{X}}\frac{[g_\theta(y_k|x_k) - g_{\theta,\epsilon}(y_k|x_k)]}{(\sup_{x\in\mathsf{X}}|g_\theta(y_k|x) - g_{\theta,\epsilon}(y_k|x)|)}f_\theta(x_k|x_{k-1})dx_k\right)\widetilde{F}_\theta^{k-1}(\mu_\theta,\widetilde{\mu_\theta})(dx_{k-1})|.$$

Then we can apply (20) and, noting that

$$\left(\int_{\mathsf{X}}\frac{[g_\theta(y_k|x_k) - g_{\theta,\epsilon}(y_k|x_k)]}{(\sup_{x\in\mathsf{X}}|g_\theta(y_k|x) - g_{\theta,\epsilon}(y_k|x)|)}f_\theta(x_k|x_{k-1})dx_k\right) \leq 1$$

one can also use Lemma D.3 to deduce that

$$|\int_{\mathsf{X}^2}[g_\theta(y_k|x_k) - g_{\theta,\epsilon}(y_k|x_k)]f_\theta(x_k|x_{k-1})\widetilde{F}_\theta^{k-1}(\mu_\theta,\widetilde{\mu_\theta})(dx_{k-1})dx_k| \leq C(1 + \|\widetilde{\mu_\theta}\|)\epsilon.$$

Then, one can easily apply Theorem D.1 to show that

$$|\int_{\mathsf{X}^2} g_{\theta,\epsilon}(y_k|x_k)f_\theta(x_k|x_{k-1})[\widetilde{F}_\theta^{k-1}(\mu_\theta,\widetilde{\mu_\theta})(dx_{k-1}) - \widetilde{F}_{\theta,\epsilon}^{k-1}(\mu_\theta,\widetilde{\mu_\theta})(dx_{k-1})]dx_k| \leq C(2 + \|\widetilde{\mu_\theta}\|)\epsilon.$$

Thus we have upper-bounded the $\mathbb{L}_1-$norm of the sum of the expressions (24)-(26) and we have established that

$$\frac{[\nabla p_\theta(y_k|y_{1:k-1}) - \nabla p_{\theta,\epsilon}(y_k|y_{1:k-1})]}{p_\theta(y_k|y_{1:k-1})} \leq C(2 + \|\widetilde{\mu_\theta}\|)\epsilon. \tag{28}$$

Moving onto the second summand on the R.H.S. of (23),

$$\frac{\nabla p_{\theta,\epsilon}(y_k|y_{1:k-1})}{p_\theta(y_k|y_{1:k-1})p_{\theta,\epsilon}(y_k|y_{1:k-1})}[p_{\theta,\epsilon}(y_k|y_{1:k-1}) - p_\theta(y_k|y_{1:k-1}).$$

By (22), we need only consider upper-bounding, in $\mathbb{L}_1$, $\nabla p_{\theta,\epsilon}(y_k|y_{1:k-1})$. This can be decomposed into the sum of three expressions:

$$\int_{\mathsf{X}^2}\nabla\{g_{\theta,\epsilon}(y_k|x_k)\}f_\theta(x_k|x_{k-1})F_{\theta,\epsilon}^{k-1}(\mu_\theta)(dx_{k-1})dx_k$$

$$\int_{\mathsf{X}^2} g_{\theta,\epsilon}(y_k|x_k)\nabla\{f_\theta(x_k|x_{k-1})\}F_{\theta,\epsilon}^{k-1}(\mu_\theta)(dx_{k-1})dx_k$$

and

$$\int_{\mathsf{X}^2} g_{\theta,\epsilon}(y_k|x_k)f_\theta(x_k|x_{k-1})\widetilde{F}_{\theta,\epsilon}^{k-1}(\mu_\theta,\widetilde{\mu_\theta})(dx_{k-1})dx_k.$$

As $\nabla\{g_{\theta,\epsilon}(y_k|x_k)\}$ and $g_{\theta,\epsilon}(y_k|x_k)\nabla\{f_\theta(x_k|x_{k-1})\}$ are upper-bounded as well as X being compact the first two expressions are upper-bounded in $\mathbb{L}_1$. In addition as $\int_{\mathsf{X}} g_{\theta,\epsilon}(y_k|x_k)f_\theta(x_k|x_{k-1})dx_k$ is upper-bounded, we can apply Lemma D.3 to see that the third expression is upper-bounded in $\mathbb{L}_1$. Hence, we have shown that

$$\left|\frac{\nabla p_{\theta,\epsilon}(y_k|y_{1:k-1})}{p_\theta(y_k|y_{1:k-1})p_{\theta,\epsilon}(y_k|y_{1:k-1})}[p_{\theta,\epsilon}(y_k|y_{1:k-1}) - p_\theta(y_k|y_{1:k-1})]\right| \leq C(1 + \|\widetilde{\mu_\theta}\|)\epsilon. \tag{29}$$

Combining the results (28)-(29) and noting (23) we can conclude. □

D Bias of the Gradient of the Filter

Theorem D.1. *Assume (A1-5). Then there exist a $C < +\infty$ such that for any $n \geq 1$, $\mu_\theta \in \mathcal{P}(\mathsf{X})$, $\widetilde{\mu_\theta} \in \mathcal{M}(\mathsf{X})$, $\epsilon > 0$, $\theta \in \Theta$:*

$$\|\widetilde{F}_\theta^n(\mu_\theta, \widetilde{\mu_\theta}) - \widetilde{F}_{\theta,\epsilon}^n(\mu_\theta, \widetilde{\mu_\theta})\| \leq C\epsilon(2 + \|\widetilde{\mu_\theta}\|).$$

Proof. We have the following telescoping sum decomposition (e.g. [13]) for the differences in the filters, with $\varphi \in \mathcal{B}_b(\mathsf{X})$:

$$F_\theta^n(\mu_\theta)(\varphi) - F_{\theta,\epsilon}^n(\mu_\theta)(\varphi) = \sum_{p=1}^n \Big[F_\theta^{n-p+1,n}(F_{\theta,\epsilon}^{n-p}(\mu_\theta))(\varphi) - F_\theta^{n-p+2,n}(F_{\theta,\epsilon}^{n-p+1}(\mu_\theta))(\varphi)\Big]$$

where we are using the notation $F_\theta^{q,n}(\mu_\theta)(\varphi) = \frac{\mu_\theta R_{q,n,\theta}(\varphi)}{\mu_\theta R_{q,n,\theta}(1)}$, for $1 \leq q \leq n$. Hence, taking gradients and swapping the order of summation and differentiation we have and omitting the second arguments of $\widetilde{F}$ on the R.H.S. (to reduce the notational burden)

$$\begin{aligned}\widetilde{F}_\theta^n(\mu_\theta, \widetilde{\mu_\theta})(\varphi) - \widetilde{F}_{\theta,\epsilon}^n(\mu_\theta, \widetilde{\mu_\theta})(\varphi) \quad = \quad & \sum_{p=1}^n \Big[\widetilde{F}_\theta^{n-p+2,n}(F_\theta^{(n-p+1)}[F_{\theta,\epsilon}^{n-p}(\mu_\theta)], \widetilde{F}_\theta^{(n-p+1)}[F_{\theta,\epsilon}^{n-p}(\mu_\theta)])(\varphi) - \\ & \widetilde{F}_\theta^{n-p+2,n}(F_{\theta,\epsilon}^{(n-p+1)}[F_{\theta,\epsilon}^{(n-p)}(\mu_\theta)], \widetilde{F}_{\theta,\epsilon}^{(n-p+1)}[F_{\theta,\epsilon}^{(n-p)}(\mu_\theta)])(\varphi)\Big]. \end{aligned} \tag{30}$$

To continue with the proof we will adopt [35, Lemma 6.4]:

$$\widetilde{F}_\theta^n(\mu_\theta, \widetilde{\mu_\theta})(\varphi) = \widetilde{G}_\theta^n(\mu_\theta, \widetilde{\mu_\theta}) + \sum_{q=1}^n \widetilde{G}_\theta^{q+1,n}(F_\theta^q(\mu_\theta), \widetilde{H}^q(\mu_\theta))(\varphi)$$

with $\widetilde{G}_\theta^n$ and $\widetilde{H}^q(\mu_\theta)$ defined in (16)-(17) and $\widetilde{G}_\theta^{q+1,n}$ similar extension to the notation as for the filter $F_\theta^{q+1,n}$ and the convention $\widetilde{G}_\theta^{n+1,n}(\mu_\theta, \widetilde{\mu_\theta}) = \widetilde{\mu_\theta}$. Returning to (30) and again omitting the second arguments of $\widetilde{F}$ on the R.H.S.:

$$\widetilde{F}_\theta^n(\mu_\theta, \widetilde{\mu_\theta})(\varphi) - \widetilde{F}_{\theta,\epsilon}^n(\mu_\theta, \widetilde{\mu_\theta})(\varphi) =$$

$$\begin{aligned}\sum_{p=1}^n \Big[&\widetilde{G}_\theta^{n-p+2,n}\{F_\theta^{(n-p+1)}(F_{\theta,\epsilon}^{n-p}(\mu_\theta)), \widetilde{F}_\theta^{(n-p+1)}(F_{\theta,\epsilon}^{n-p}(\mu_\theta))\}(\varphi) - \widetilde{G}_\theta^{n-p+2,n}\{F_{\theta,\epsilon}^{(n-p+1)}[F_{\theta,\epsilon}^{n-p}(\mu_\theta)], \widetilde{F}_{\theta,\epsilon}^{(n-p+1)}[F_{\theta,\epsilon}^{n-p}(\mu_\theta)]\}(\varphi) + \\ &\sum_{q=n-p+2}^n \Big\{\widetilde{G}_\theta^{q+1,n}\{F_\theta^{n-p+2,q}[F_\theta^{(n-p+1)}(F_{\theta,\epsilon}^{n-p}(\mu_\theta))], \widetilde{H}_\theta^{n-p+2,q}[F_\theta^{(n-p+1)}(F_{\theta,\epsilon}^{n-p}(\mu_\theta))]\}(\varphi) - \\ &\widetilde{G}_\theta^{q+1,n}\{F_\theta^{n-p+2,q}[F_{\theta,\epsilon}^{(n-p+1)}(F_{\theta,\epsilon}^{n-p}(\mu_\theta))], \widetilde{H}_\theta^{n-p+2,q}[F_{\theta,\epsilon}^{(n-p+1)}(F_{\theta,\epsilon}^{n-p}(\mu_\theta))]\}(\varphi)\Big\}\Big]. \end{aligned} \tag{31}$$

We start first with the summand on the R.H.S. of the second line of (31), which we compactly denote as:

$$\widetilde{G}_\theta^{p-1}\{F_\theta[F_{\theta,\epsilon}^{n-p}(\mu_\theta)], \widetilde{F}_\theta[F_{\theta,\epsilon}^{n-p}(\mu_\theta)]\}(\varphi) - \widetilde{G}_\theta^{p-1}\{F_{\theta,\epsilon}[F_{\theta,\epsilon}^{n-p}(\mu_\theta)], \widetilde{F}_{\theta,\epsilon}[F_{\theta,\epsilon}^{n-p}(\mu_\theta)]\}(\varphi).$$

This can be decomposed further into the sum of

$$\widetilde{G}_\theta^{p-1}\{F_\theta[F_{\theta,\epsilon}^{n-p}(\mu_\theta)], \widetilde{F}_\theta[F_{\theta,\epsilon}^{n-p}(\mu_\theta)]\}(\varphi) - \widetilde{G}_\theta^{p-1}\{F_{\theta,\epsilon}[F_{\theta,\epsilon}^{n-p}(\mu_\theta)], \widetilde{F}_\theta[F_{\theta,\epsilon}^{n-p}(\mu_\theta)]\}(\varphi) \tag{32}$$

and

$$\widetilde{G}_\theta^{p-1}\{F_{\theta,\epsilon}[F_{\theta,\epsilon}^{n-p}(\mu_\theta)], \widetilde{F}_\theta[F_{\theta,\epsilon}^{n-p}(\mu_\theta)]\}(\varphi) - \widetilde{G}_\theta^{p-1}\{F_{\theta,\epsilon}[F_{\theta,\epsilon}^{n-p}(\mu_\theta)], \widetilde{F}_{\theta,\epsilon}[F_{\theta,\epsilon}^{n-p}(\mu_\theta)]\}(\varphi). \tag{33}$$

Beginning with (32), by [35, Lemma 6.7], equation (43) we have

$$|\widetilde{G}_\theta^{p-1}\{F_\theta[F_{\theta,\epsilon}^{n-p}(\mu_\theta)], \widetilde{F}_\theta[F_{\theta,\epsilon}^{n-p}(\mu_\theta)]\}(\varphi) - \widetilde{G}_\theta^{p-1}\{F_{\theta,\epsilon}[F_{\theta,\epsilon}^{n-p}(\mu_\theta)], \widetilde{F}_\theta[F_{\theta,\epsilon}^{n-p}(\mu_\theta)]\}(\varphi)|$$

$$\leq C\|\varphi\|_\infty \rho^{p-1}\|F_\theta[F_{\theta,\epsilon}^{n-p}(\mu_\theta)] - F_{\theta,\epsilon}[F_{\theta,\epsilon}^{n-p}(\mu_\theta)]\|\|\widetilde{F}_\theta[F_{\theta,\epsilon}^{n-p}(\mu_\theta)]\|$$

where $\rho \in (0,1)$ and C do not depend upon μ_θ, ϵ or n, p. Applying Lemma D.2 we have

$$|\widetilde{G}_\theta^{p-1}\{F_\theta[F_{\theta,\epsilon}^{n-p}(\mu_\theta)], \widetilde{F}_\theta[F_{\theta,\epsilon}^{n-p}(\mu_\theta)]\}(\varphi) - \widetilde{G}_\theta^{p-1}\{F_{\theta,\epsilon}[F_{\theta,\epsilon}^{n-p}(\mu_\theta)], \widetilde{F}_\theta[F_{\theta,\epsilon}^{n-p}(\mu_\theta)]\}(\varphi)|$$

$$\leq C\|\varphi\|_\infty \rho^{p-1}\epsilon\|\widetilde{F}_\theta[F_{\theta,\epsilon}^{n-p}(\mu_\theta)]\|$$

where C does not depend upon μ_θ, ϵ or n, p. Then by Remark D.1 and Lemma D.3 $\|\widetilde{F}_\theta[F_{\theta,\epsilon}^{n-p}(\mu_\theta)]\| \leq C(2 + \|\widetilde{\mu_\theta}\|)$ and thus the upper-bound on the $\mathbb{L}_1-$norm of (32):

$$|\widetilde{G}_\theta^{p-1}\{F_\theta[F_{\theta,\epsilon}^{n-p}(\mu_\theta)], \widetilde{F}_\theta[F_{\theta,\epsilon}^{n-p}(\mu_\theta)]\}(\varphi) - \widetilde{G}_\theta^{p-1}\{F_{\theta,\epsilon}[F_{\theta,\epsilon}^{n-p}(\mu_\theta)], \widetilde{F}_\theta[F_{\theta,\epsilon}^{n-p}(\mu_\theta)]\}(\varphi)| \leq C\|\varphi\|_\infty \epsilon\rho^{p-1}(2 + \|\widetilde{\mu_\theta}\|). \quad (34)$$

Now, moving onto (33), by [35, Lemma 6.7], equation (42):

$$|\widetilde{G}_\theta^{p-1}\{F_{\theta,\epsilon}[F_{\theta,\epsilon}^{n-p}(\mu_\theta)], \widetilde{F}_\theta[F_{\theta,\epsilon}^{n-p}(\mu_\theta)]\}(\varphi) - \widetilde{G}_\theta^{p-1}\{F_{\theta,\epsilon}[F_{\theta,\epsilon}^{n-p}(\mu_\theta)], \widetilde{F}_{\theta,\epsilon}[F_{\theta,\epsilon}^{n-p}(\mu_\theta)]\}(\varphi)|$$

$$\leq C\rho^{p-1}\|\varphi\|_\infty \|\widetilde{F}_\theta[F_{\theta,\epsilon}^{n-p}(\mu_\theta)] - \widetilde{F}_{\theta,\epsilon}[F_{\theta,\epsilon}^{n-p}(\mu_\theta)]\|.$$

Applying Lemma D.1

$$|\widetilde{G}_\theta^{p-1}\{F_{\theta,\epsilon}[F_{\theta,\epsilon}^{n-p}(\mu_\theta)], \widetilde{F}_\theta[F_{\theta,\epsilon}^{n-p}(\mu_\theta)]\}(\varphi) - \widetilde{G}_\theta^{p-1}\{F_{\theta,\epsilon}[F_{\theta,\epsilon}^{n-p}(\mu_\theta)], \widetilde{F}_{\theta,\epsilon}[F_{\theta,\epsilon}^{n-p}(\mu_\theta)]\}(\varphi)|$$

$$\leq C\|\varphi\|_\infty \epsilon\rho^{p-1}(1 + \|\widetilde{F}_{\theta,\epsilon}^{n-p}(\mu_\theta)\|).$$

Then by Lemma D.3, we deduce that

$$|\widetilde{G}_\theta^{p-1}\{F_{\theta,\epsilon}[F_{\theta,\epsilon}^{n-p}(\mu_\theta)], \widetilde{F}_\theta[F_{\theta,\epsilon}^{n-p}(\mu_\theta)]\}(\varphi) - \widetilde{G}_\theta^{p-1}\{F_{\theta,\epsilon}[F_{\theta,\epsilon}^{n-p}(\mu_\theta)], \widetilde{F}_{\theta,\epsilon}[F_{\theta,\epsilon}^{n-p}(\mu_\theta)]\}(\varphi)| \leq C\|\varphi\|_\infty \epsilon\rho^{p-1}(2 + \|\widetilde{\mu_\theta}\|). \quad (35)$$

Combining (34) and (35)

$$|\widetilde{G}_\theta^{p-1}\{F_\theta[F_{\theta,\epsilon}^{n-p}(\mu_\theta)], \widetilde{F}_\theta[F_{\theta,\epsilon}^{n-p}(\mu_\theta)]\}(\varphi) - \widetilde{G}_\theta^{p-1}\{F_{\theta,\epsilon}[F_{\theta,\epsilon}^{n-p}(\mu_\theta)], \widetilde{F}_{\theta,\epsilon}[F_{\theta,\epsilon}^{n-p}(\mu_\theta)]\}(\varphi)| \leq C\|\varphi\|_\infty \epsilon\rho^{p-1}(2 + \|\widetilde{\mu_\theta}\|). \quad (36)$$

We now consider the summands over q in the second and third lines of (31). Again, adopting the compact notation above we can decompose the summands over q into the sum of

$$\widetilde{G}_\theta^{n-q}\{F_\theta^s[F_\theta(F_{\theta,\epsilon}^{n-p}(\mu_\theta))], \widetilde{H}_\theta^s[F_\theta(F_{\theta,\epsilon}^{n-p}(\mu_\theta))]\}(\varphi) - \widetilde{G}_\theta^{n-q}\{F_\theta^s[F_{\theta,\epsilon}(F_{\theta,\epsilon}^{n-p}(\mu_\theta))], \widetilde{H}_\theta^s[F_\theta(F_{\theta,\epsilon}^{n-p}(\mu_\theta))]\}(\varphi) \quad (37)$$

and

$$\widetilde{G}_\theta^{n-q}\{F_\theta^s[F_{\theta,\epsilon}(F_{\theta,\epsilon}^{n-p}(\mu_\theta))], \widetilde{H}_\theta^s[F_\theta(F_{\theta,\epsilon}^{n-p}(\mu_\theta))]\}(\varphi) - \widetilde{G}_\theta^{n-q}\{F_\theta^s[F_{\theta,\epsilon}(F_{\theta,\epsilon}^{n-p}(\mu_\theta))], \widetilde{H}_\theta^s[F_{\theta,\epsilon}(F_{\theta,\epsilon}^{n-p}(\mu_\theta))]\}(\varphi) \quad (38)$$

where $s = q - n + p - 1$. We start with (37); by [35, Lemma 6.7] equation (43), we have

$$|\widetilde{G}_\theta^{n-q}\{F_\theta^s[F_\theta(F_{\theta,\epsilon}^{n-p}(\mu_\theta))], \widetilde{H}_\theta^s[F_\theta(F_{\theta,\epsilon}^{n-p}(\mu_\theta))]\}(\varphi) - \widetilde{G}_\theta^{n-q}\{F_\theta^s[F_{\theta,\epsilon}(F_{\theta,\epsilon}^{n-p}(\mu_\theta))], \widetilde{H}_\theta^s[F_\theta(F_{\theta,\epsilon}^{n-p}(\mu_\theta))]\}(\varphi)|$$

$$\leq C\|\varphi\|_\infty \rho^{n-q}\|F_\theta^s[F_\theta(F_{\theta,\epsilon}^{n-p}(\mu_\theta))] - F_\theta^s[F_{\theta,\epsilon}(F_{\theta,\epsilon}^{n-p}(\mu_\theta))]\|\|\|\widetilde{H}_\theta^s[F_\theta(F_{\theta,\epsilon}^{n-p}(\mu_\theta))\|.$$

Then we will use the stability of the filter (e.g. [35, Theorem 3.1])

$$\|F_\theta^s[F_\theta(F_{\theta,\epsilon}^{n-p}(\mu_\theta))] - F_\theta^s[F_{\theta,\epsilon}(F_{\theta,\epsilon}^{n-p}(\mu_\theta))]\| \leq C\rho^s\|F_\theta(F_{\theta,\epsilon}^{n-p}(\mu_\theta)) - F_{\theta,\epsilon}(F_{\theta,\epsilon}^{n-p}(\mu_\theta))\|.$$

By Lemma D.2 $\|F_\theta(F_{\theta,\epsilon}^{n-p}(\mu_\theta)) - F_{\theta,\epsilon}(F_{\theta,\epsilon}^{n-p}(\mu_\theta))\| \leq C\epsilon$ and thus

$$|\widetilde{G}_\theta^{n-q}\{F_\theta^s[F_\theta(F_{\theta,\epsilon}^{n-p}(\mu_\theta))], \widetilde{H}_\theta^s[F_\theta(F_{\theta,\epsilon}^{n-p}(\mu_\theta))]\}(\varphi) - \widetilde{G}_\theta^{n-q}\{F_\theta^s[F_{\theta,\epsilon}(F_{\theta,\epsilon}^{n-p}(\mu_\theta))], \widetilde{H}_\theta^s[F_\theta(F_{\theta,\epsilon}^{n-p}(\mu_\theta))]\}(\varphi)|$$

$$\leq C\|\varphi\|_\infty \epsilon\rho^{p-1}\|\widetilde{H}_\theta^s[F_\theta(F_{\theta,\epsilon}^{n-p}(\mu_\theta))]\|.$$

By [35, Lemma 6.8] we have $\|\widetilde{H}_\theta^s[F_\theta(F_{\theta,\epsilon}^{n-p}(\mu_\theta))]\| \leq C$, where C does not depend upon $F_\theta(F_{\theta,\epsilon}^{n-p}(\mu_\theta))$ or ϵ and hence

$$|\widetilde{G}_\theta^{n-q}\{F_\theta^s[F_\theta(F_{\theta,\epsilon}^{n-p}(\mu_\theta))], \widetilde{H}_\theta^s[F_\theta(F_{\theta,\epsilon}^{n-p}(\mu_\theta))]\}(\varphi) - \widetilde{G}_\theta^{n-q}\{F_\theta^s[F_{\theta,\epsilon}(F_{\theta,\epsilon}^{n-p}(\mu_\theta))], \widetilde{H}_\theta^s[F_\theta(F_{\theta,\epsilon}^{n-p}(\mu_\theta))]\}(\varphi)| \leq C\|\varphi\|_\infty \epsilon \rho^{p-1}.$$

Now, turning to (38) and applying [35, Lemma 6.7] (42) we have

$$|\widetilde{G}_\theta^{n-q}\{F_\theta^s[F_{\theta,\epsilon}(F_{\theta,\epsilon}^{n-p}(\mu_\theta))], \widetilde{H}_\theta^s[F_\theta(F_{\theta,\epsilon}^{n-p}(\mu_\theta))]\}(\varphi) - \widetilde{G}_\theta^{n-q}\{F_\theta^s[F_{\theta,\epsilon}(F_{\theta,\epsilon}^{n-p}(\mu_\theta))], \widetilde{H}_\theta^s[F_{\theta,\epsilon}(F_{\theta,\epsilon}^{n-p}(\mu_\theta))]\}(\varphi)|$$

$$\leq C\|\varphi\|_\infty \rho^{n-q} \|\widetilde{H}_\theta^s[F_\theta(F_{\theta,\epsilon}^{n-p}(\mu_\theta))] - \widetilde{H}_\theta^s[F_{\theta,\epsilon}(F_{\theta,\epsilon}^{n-p}(\mu_\theta))]\|. \tag{39}$$

Then by [35, Lemma 6.8] we have

$$\|\widetilde{H}_\theta^s[F_\theta(F_{\theta,\epsilon}^{n-p}(\mu_\theta))] - \widetilde{H}_\theta^s[F_{\theta,\epsilon}(F_{\theta,\epsilon}^{n-p}(\mu_\theta))]\| \leq C\rho^s \|F_\theta(F_{\theta,\epsilon}^{n-p})(\mu_\theta) - F_{\theta,\epsilon}(F_{\theta,\epsilon}^{n-p}(\mu_\theta))\|$$

and then on applying Lemma D.2 we thus have that

$$\|\widetilde{H}_\theta^q(F_\theta(F_{\theta,\epsilon}^{n-p})(\mu_\theta)) - \widetilde{H}_\theta^q(F_{\theta,\epsilon}^{n-p+1})(\mu_\theta)\| \leq C\epsilon\rho^s.$$

Returning to (39), it follows by the above calculations that:

$$|\widetilde{G}_\theta^{n-q}\{F_\theta^s[F_\theta(F_{\theta,\epsilon}^{n-p}(\mu_\theta))], \widetilde{H}_\theta^s[F_\theta(F_{\theta,\epsilon}^{n-p}(\mu_\theta))]\}(\varphi) - \widetilde{G}_\theta^{n-q}\{F_\theta^s[F_{\theta,\epsilon}(F_{\theta,\epsilon}^{n-p}(\mu_\theta))], \widetilde{H}_\theta^s[F_\theta(F_{\theta,\epsilon}^{n-p}(\mu_\theta))]\}(\varphi)| \leq C\|\varphi\|_\infty \epsilon \rho^{p-1}.$$

Thus we have proved that

$$|\widetilde{G}_\theta^{n-q}\{F_\theta^s[F_\theta(F_{\theta,\epsilon}^{n-p}(\mu_\theta))], \widetilde{H}_\theta^s[F_\theta(F_{\theta,\epsilon}^{n-p}(\mu_\theta))]\}(\varphi) - \widetilde{G}_\theta^{n-q}\{F_\theta^s[F_{\theta,\epsilon}(F_{\theta,\epsilon}^{n-p}(\mu_\theta))], \widetilde{H}_\theta^s[F_{\theta,\epsilon}(F_{\theta,\epsilon}^{n-p}(\mu_\theta))]\}(\varphi)\| \leq C\|\varphi\|_\infty \epsilon \rho^{p-1}. \tag{40}$$

Then, returning to (31) and noting (36), (40) we have the upper-bound

$$\|\widetilde{F}_\theta^n(\mu_\theta, \widetilde{\mu_\theta}) - \widetilde{F}_{\theta,\epsilon}^n(\mu_\theta, \widetilde{\mu_\theta})\| \leq C\epsilon(2 + \|\widetilde{\mu_\theta}\|)\sum_{p=1}^{n}[\rho^{p-1} + \sum_{q=n-p}^{n} \rho^{p-1}] \leq C\epsilon(2 + \|\widetilde{\mu_\theta}\|).$$

□

D.1 Technical Results for ABC Bias of the Filter-Derivative

Lemma D.1. *Assume (A1-5). Then there exist a* $C < +\infty$ *such that for any* $n \geq 1$, $\mu_\theta \in \mathcal{P}(\mathsf{X})$, $\widetilde{\mu_\theta} \in \mathcal{M}(\mathsf{X})$, $\epsilon > 0$ $\theta \in \Theta$*:*

$$\|\widetilde{F}_\theta^{(n)}(\mu_\theta, \widetilde{\mu_\theta}) - \widetilde{F}_{\theta,\epsilon}^{(n)}(\mu_\theta, \widetilde{\mu_\theta})\| \leq C\epsilon(1 + \|\widetilde{\mu_\theta}\|).$$

Proof. By [35, Lemma 6.7] we have the decomposition, for $\varphi \in \mathcal{B}_b(\mathsf{X})$:

$$\widetilde{F}_\theta^{(n)}(\mu_\theta, \widetilde{\mu_\theta})(\varphi) = \widetilde{G}_\theta^{(n)}(\mu_\theta, \widetilde{\mu_\theta})(\varphi) - \widetilde{H}_\theta^{(n)}(\mu_\theta)(\varphi)$$

where

$$\widetilde{H}^{(n)}(\mu_\theta)(\varphi) := \mu_\theta R_{n,\theta}(1)^{-1}[\mu_\theta \widetilde{R}_{n,\theta}(\varphi) - \mu_\theta \widetilde{R}_{n,\theta}(1)\mu_\theta(\varphi).$$

Thus to control the difference, we can consider the two differences $\widetilde{G}_\theta^{(n)}(\mu_\theta, \widetilde{\mu_\theta})(\varphi) - \widetilde{G}_{\theta,\epsilon}^{(n)}(\mu_\theta, \widetilde{\mu_\theta})(\varphi)$ and $\widetilde{H}_\theta^{(n)}(\mu_\theta)(\varphi) - \widetilde{H}_{\theta,\epsilon}^{(n)}(\mu_\theta)(\varphi)$.

Control of $\widetilde{G}_\theta^{(n)}(\mu_\theta, \widetilde{\mu_\theta})(\varphi) - \widetilde{G}_{\theta,\epsilon}^{(n)}(\mu_\theta, \widetilde{\mu_\theta})(\varphi)$. We will use the Hahn-Jordan decomposition: $\widetilde{\mu_\theta} = \widetilde{\mu_\theta}^{+} - \widetilde{\mu_\theta}^{-}$. It is assumed that both $\widetilde{\mu_\theta}^{+}(1), \widetilde{\mu_\theta}^{-}(1) > 0$. The scenario with either $\widetilde{\mu_\theta}^{+}(1) = 0$ or $\widetilde{\mu_\theta}^{+}(1) = 0$ is straightforward and omitted for brevity. We can write:

$$\widetilde{G}_\theta^{(n)}(\mu_\theta, \widetilde{\mu_\theta})(\varphi) = \frac{\widetilde{\mu_\theta}^{+} R_{n,\theta}(1)}{\mu_\theta R_{n,\theta}(1)}[F_\theta^{(n)}(\widetilde{\bar{\mu}_\theta}^{+})(\varphi) - F_\theta^{(n)}(\mu_\theta)(\varphi)] + \frac{\widetilde{\mu_\theta}^{-} R_{n,\theta}(1)}{\mu_\theta R_{n,\theta}(1)}[F_\theta^{(n)}(\widetilde{\bar{\mu}_\theta}^{-})(\varphi) - F_\theta^{(n)}(\mu_\theta)(\varphi)]$$

where $\widetilde{\bar{\mu}_\theta}^+(\cdot) = \widetilde{\mu_\theta}^+(\cdot)/\widetilde{\mu_\theta}^+(1)$ and $\widetilde{\bar{\mu}_\theta}^-(\cdot) = \widetilde{\mu_\theta}^-(\cdot)/\widetilde{\mu_\theta}^-(1)$. Thus we have

$$\begin{aligned}
\widetilde{G}_\theta^{(n)}(\mu_\theta, \widetilde{\mu_\theta})(\varphi) - \widetilde{G}_{\theta,\epsilon}^{(n)}(\mu_\theta, \widetilde{\mu_\theta})(\varphi) \quad = \quad & \left[\frac{\widetilde{\mu_\theta}^+ R_{n,\theta}(1)}{\mu_\theta R_{n,\theta}(1)} - \frac{\widetilde{\mu_\theta}^+ R_{n,\theta,\epsilon}(1)}{\mu_\theta R_{n,\theta,\epsilon}(1)}\right][F_\theta^{(n)}(\widetilde{\bar{\mu}_\theta}^+)(\varphi) - F_\theta^{(n)}(\mu_\theta)(\varphi)] \\
& + \frac{\widetilde{\mu_\theta}^+ R_{n,\theta,\epsilon}(1)}{\mu_\theta R_{n,\theta,\epsilon}(1)}[F_\theta^{(n)}(\widetilde{\bar{\mu}_\theta}^+)(\varphi) - F_\theta^{(n)}(\mu_\theta)(\varphi) - F_{\theta,\epsilon}^{(n)}(\widetilde{\bar{\mu}_\theta}^+)(\varphi) + F_{\theta,\epsilon}^{(n)}(\mu_\theta)(\varphi)] \\
& + \left[\frac{\widetilde{\mu_\theta}^- R_{n,\theta}(1)}{\mu_\theta R_{n,\theta}(1)} - \frac{\widetilde{\mu_\theta}^- R_{n,\theta,\epsilon}(1)}{\mu_\theta R_{n,\theta,\epsilon}(1)}\right][F_\theta^{(n)}(\widetilde{\bar{\mu}_\theta}^-)(\varphi) - F_\theta^{(n)}(\mu_\theta)(\varphi)] \\
& + \frac{\widetilde{\mu_\theta}^- R_{n,\theta,\epsilon}(1)}{\mu_\theta R_{n,\theta,\epsilon}(1)}[F_\theta^{(n)}(\widetilde{\bar{\mu}_\theta}^-)(\varphi) - F_\theta^{(n)}(\mu_\theta)(\varphi) - F_{\theta,\epsilon}^{(n)}(\widetilde{\bar{\mu}_\theta}^-)(\varphi) + F_{\theta,\epsilon}^{(n)}(\mu_\theta)(\varphi)]
\end{aligned} \tag{41}$$

By symmetry, we need only consider the terms including $\widetilde{\mu_\theta}^+$; one can treat those with $\widetilde{\mu_\theta}^-$ by using similar arguments. First dealing with term on the first line of the R.H.S. of (41). We have that

$$\left[\frac{\widetilde{\mu_\theta}^+ R_{n,\theta}(1)}{\mu_\theta R_{n,\theta}(1)} - \frac{\widetilde{\mu_\theta}^+ R_{n,\theta,\epsilon}(1)}{\mu_\theta R_{n,\theta,\epsilon}(1)}\right][F_\theta^{(n)}(\widetilde{\bar{\mu}_\theta}^+)(\varphi) - F_\theta^{(n)}(\mu_\theta)(\varphi)] =$$

$$\left[\frac{\widetilde{\mu_\theta}^+ R_{n,\theta}(1) - \widetilde{\mu_\theta}^+ R_{n,\theta,\epsilon}(1)}{\mu_\theta R_{n,\theta}(1)} + \widetilde{\mu_\theta}^+ R_{n,\theta,\epsilon}(1)\frac{\mu_\theta R_{n,\theta,\epsilon}(1) - \mu_\theta R_{n,\theta}(1)}{\mu_\theta R_{n,\theta,\epsilon}(1)\mu_\theta R_{n,\theta}(1)}\right][F_\theta^{(n)}(\widetilde{\bar{\mu}_\theta}^+)(\varphi) - F_\theta^{(n)}(\mu_\theta)(\varphi)]$$

Now by (A1), for any n

$$\sup_{x\in\mathsf{X}} |R_{n,\theta}(1)(x) - R_{n,\theta,\epsilon}(1)(x)| \leq C\epsilon \tag{42}$$

thus

$$\left[\frac{\widetilde{\mu_\theta}^+ R_{n,\theta}(1) - \widetilde{\mu_\theta}^+ R_{n,\theta,\epsilon}(1)}{\mu_\theta R_{n,\theta}(1)} + \widetilde{\mu_\theta}^+ R_{n,\theta,\epsilon}(1)\frac{\mu_\theta R_{n,\theta,\epsilon}(1) - \mu_\theta R_{n,\theta}(1)}{\mu_\theta R_{n,\theta,\epsilon}(1)\mu_\theta R_{n,\theta}(1)}\right] \leq \frac{C\epsilon\widetilde{\mu_\theta}^+(1)}{\mu_\theta R_{n,\theta}(1)} + C\epsilon\frac{\widetilde{\mu_\theta}^+ R_{n,\theta,\epsilon}(1)}{\mu_\theta R_{n,\theta,\epsilon}(1)\mu_\theta R_{n,\theta}(1)}.$$

Now one can show that there exist a $C < +\infty$ such that for any $x, y \in \mathsf{X}$

$$R_{n,\theta}(1)(x) \geq C R_{n,\theta}(1)(y) \qquad R_{n,\theta,\epsilon}(1)(x) \geq C R_{n,\theta,\epsilon}(1)(y). \tag{43}$$

Then it follows that

$$\frac{C\epsilon\widetilde{\mu_\theta}^+(1)}{\mu_\theta R_{n,\theta}(1)} + C\epsilon\frac{\widetilde{\mu_\theta}^+ R_{n,\theta,\epsilon}(1)}{\mu_\theta R_{n,\theta,\epsilon}(1)\mu_\theta R_{n,\theta}(1)} \leq C\epsilon\widetilde{\mu_\theta}^+(1).$$

Hence we have shown that

$$\left[\frac{\widetilde{\mu_\theta}^+ R_{n,\theta}(1)}{\mu_\theta R_{n,\theta}(1)} - \frac{\widetilde{\mu_\theta}^+ R_{n,\theta,\epsilon}(1)}{\mu_\theta R_{n,\theta,\epsilon}(1)}\right][F_\theta^{(n)}(\widetilde{\bar{\mu}_\theta}^+)(\varphi) - F_\theta^{(n)}(\mu_\theta)(\varphi)] \leq C\|\varphi\|_\infty\epsilon\widetilde{\mu_\theta}^+(1).$$

Second, the second line of the R.H.S. of (41). By Lemma D.2, for any $\mu_\theta \in \mathcal{P}(\mathsf{X})$, $\|F_\theta^{(n)}(\mu_\theta) - F_{\theta,\epsilon}^{(n)}(\mu_\theta)\| \leq C\epsilon$, with C independent of μ_θ, and in addition using (43) we have

$$\frac{\widetilde{\mu_\theta}^+ R_{n,\theta,\epsilon}(1)}{\mu_\theta R_{n,\theta,\epsilon}(1)}[F_\theta^{(n)}(\widetilde{\bar{\mu}_\theta}^+)(\varphi) - F_\theta^{(n)}(\mu_\theta)(\varphi) - F_{\theta,\epsilon}^{(n)}(\widetilde{\bar{\mu}_\theta}^+)(\varphi) + F_{\theta,\epsilon}^{(n)}(\mu_\theta)(\varphi)] \leq C\|\varphi\|_\infty\epsilon\widetilde{\mu_\theta}^+(1).$$

Thus we have shown:

$$\|\widetilde{G}_\theta^{(n)}(\mu_\theta, \widetilde{\mu_\theta})(\varphi) - \widetilde{G}_{\theta,\epsilon}^{(n)}(\mu_\theta, \widetilde{\mu_\theta})(\varphi)\| \leq C\epsilon[\widetilde{\mu_\theta}^+(1) + \widetilde{\mu_\theta}^-(1)] = C\epsilon\|\widetilde{\mu_\theta}\|. \tag{44}$$

Control of $\widetilde{H}_\theta^{(n)}(\mu_\theta)(\varphi) - \widetilde{H}_{\theta,\epsilon}^{(n)}(\mu_\theta)(\varphi)$. We have

$$\widetilde{H}_\theta^{(n)}(\mu_\theta)(\varphi) - \widetilde{H}_{\theta,\epsilon}^{(n)}(\mu_\theta)(\varphi) = \left[\frac{\mu_\theta\widetilde{R}_{n,\theta}(\varphi)}{\mu_\theta R_{n,\theta}(1)} - \frac{\mu_\theta\widetilde{R}_{n,\theta,\epsilon}(\varphi)}{\mu_\theta R_{n,\theta,\epsilon}(1)}\right] + \left[\frac{\mu_\theta\widetilde{R}_{n,\theta,\epsilon}(1)F_{\theta,\epsilon}^{(n)}(\mu_\theta)(\varphi)}{\mu_\theta R_{n,\theta,\epsilon}(1)} - \frac{\mu_\theta\widetilde{R}_{n,\theta}(1)F_\theta^{(n)}(\mu_\theta)(\varphi)}{\mu_\theta R_{n,\theta}(1)}\right]. \tag{45}$$

We start with the first bracket on the R.H.S. of (45). We first note that

$$\widetilde{R}_{n,\theta}(\varphi)(x) - \widetilde{R}_{n,\theta,\epsilon}(\varphi)(x) = \int f_\theta(x'|x)\varphi(x')[\nabla g_\theta(y_n|x') - \nabla g_{\theta,\epsilon}(y_n|x')]dx' \leq C\|\varphi\|_\infty \epsilon \tag{46}$$

where we have applied (27). Then we have

$$\frac{\mu_\theta \widetilde{R}_{n,\theta}(\varphi)}{\mu_\theta R_{n,\theta}(1)} - \frac{\mu_\theta \widetilde{R}_{n,\theta,\epsilon}(\varphi)}{\mu_\theta R_{n,\theta,\epsilon}(1)} = \frac{\mu_\theta \widetilde{R}_{n,\theta}(\varphi) - \mu_\theta \widetilde{R}_{n,\theta,\epsilon}(\varphi)}{\mu_\theta R_{n,\theta}(1)} + \mu_\theta \widetilde{R}_{n,\theta,\epsilon}(\varphi)\frac{\mu_\theta R_{n,\theta,\epsilon}(1) - \mu_\theta R_{n,\theta}(1)}{\mu_\theta R_{n,\theta,\epsilon}(1)\mu_\theta R_{n,\theta}(1)}.$$

By using (46) on the first term on the R.H.S. of the above equation and by using (42) in the numerator for the second, along with (43) in the denominator, we have

$$\left|\frac{\mu_\theta \widetilde{R}_{n,\theta}(\varphi)}{\mu_\theta R_{n,\theta}(1)} - \frac{\mu_\theta \widetilde{R}_{n,\theta,\epsilon}(\varphi)}{\mu_\theta R_{n,\theta,\epsilon}(1)}\right| \leq C\epsilon[\|\varphi\|_\infty + |\mu_\theta \widetilde{R}_{n,\theta,\epsilon}(\varphi)|].$$

Then as

$$\widetilde{R}_{n,\theta,\epsilon}(\varphi)(x) = \int \varphi(x')[\nabla g_{\theta,\epsilon}(y_n|x')f_\theta(x'|x) - g_{\theta,\epsilon}(y_n|x)\nabla f_\theta(x'|x)]dx' \leq C\|\varphi\|_\infty \int_{\mathsf{X}} dx' \leq C\|\varphi\|_\infty \tag{47}$$

where the compactness of X and (A5) have been used, we have the upper-bound

$$\left|\frac{\mu_\theta \widetilde{R}_{n,\theta}(\varphi)}{\mu_\theta R_{n,\theta}(1)} - \frac{\mu_\theta \widetilde{R}_{n,\theta,\epsilon}(\varphi)}{\mu_\theta R_{n,\theta,\epsilon}(1)}\right| \leq C\|\varphi\|_\infty \epsilon. \tag{48}$$

Moving onto the second bracket on the R.H.S. of (45), this is equal to

$$\left[\frac{\mu_\theta \widetilde{R}_{n,\theta,\epsilon}(1)}{\mu_\theta R_{n,\theta,\epsilon}(1)} - \frac{\mu_\theta \widetilde{R}_{n,\theta}(1)}{\mu_\theta R_{n,\theta}(1)}\right] F_{\theta,\epsilon}^{(n)}(\mu_\theta)(\varphi) + \frac{\mu_\theta \widetilde{R}_{n,\theta}(1)}{\mu_\theta R_{n,\theta}(1)}[F_{\theta,\epsilon}^{(n)}(\mu_\theta)(\varphi) - F_\theta^{(n)}(\mu_\theta)(\varphi)]$$

By using the inequality (48), we have

$$\left[\frac{\mu_\theta \widetilde{R}_{n,\theta,\epsilon}(1)}{\mu_\theta R_{n,\theta,\epsilon}(1)} - \frac{\mu_\theta \widetilde{R}_{n,\theta}(1)}{\mu_\theta R_{n,\theta}(1)}\right] F_{\theta,\epsilon}^{(n)}(\mu_\theta)(\varphi) \leq C\epsilon|F_{\theta,\epsilon}^{(n)}(\mu_\theta)(\varphi)| \leq C\|\varphi\|_\infty \epsilon.$$

Using Lemma D.2 and in addition using (43) in the denominator and (47) in the numerator we have

$$\frac{\mu_\theta \widetilde{R}_{n,\theta}(1)}{\mu_\theta R_{n,\theta}(1)}[F_{\theta,\epsilon}^{(n)}(\mu_\theta)(\varphi) - F_\theta^{(n)}(\mu_\theta)(\varphi)] \leq C\|\varphi\|_\infty \epsilon$$

where C does not depend upon μ_θ and ϵ. Thus we have established that

$$\frac{\mu_\theta \widetilde{R}_{n,\theta,\epsilon}(1)F_{\theta,\epsilon}^{(n)}(\mu_\theta)(\varphi)}{\mu_\theta R_{n,\theta,\epsilon}(1)} - \frac{\mu_\theta \widetilde{R}_{n,\theta}(1)F_\theta^{(n)}(\mu_\theta)(\varphi)}{\mu_\theta R_{n,\theta}(1)} \leq C\|\varphi\|_\infty \epsilon. \tag{49}$$

One can put together the results of (48) and (49) and establish that

$$|\widetilde{H}_\theta^{(n)}(\mu_\theta)(\varphi) - \widetilde{H}_{\theta,\epsilon}^{(n)}(\mu_\theta)(\varphi)| \leq C\|\varphi\|_\infty \epsilon. \tag{50}$$

On combining the results (44) and (50) and noting (45) we conclude the proof. □

Lemma D.2. *Assume (A1-3). Then there exist a $C < +\infty$ such that for any $n \geq 1$, $\mu_\theta \in \mathcal{P}(\mathsf{X})$, $\epsilon > 0$, $\theta \in \Theta$:*

$$\|F_\theta^{(n)}(\mu_\theta) - F_{\theta,\epsilon}^{(n)}(\mu_\theta)\| \leq C\epsilon.$$

Proof. For $\varphi \in \mathcal{B}_b(\varphi)$

$$F_\theta^{(n)}(\mu_\theta)(\varphi) - F_{\theta,\epsilon}^{(n)}(\mu_\theta)(\varphi) = \frac{\mu_\theta R_{n,\theta}(\varphi) - \mu_\theta R_{n,\theta,\epsilon}(\varphi)}{\mu_\theta R_{n,\theta}(1)} + \mu_\theta R_{n,\theta,\epsilon}(\varphi)\left[\frac{\mu_\theta R_{n,\theta,\epsilon}(1) - \mu_\theta R_{n,\theta}(1)}{\mu_\theta R_{n,\theta,\epsilon}(1)\mu_\theta R_{n,\theta}(1)}\right].$$

Then by applying (42) on both terms on the R.H.S. we have the upper-bound

$$\frac{C\|\varphi\|_\infty \epsilon}{\mu_\theta R_{n,\theta}(1)}.$$

One can conclude by using the inequality (43) for $R_{n,\theta}(1)(\cdot)$. □

Lemma D.3. *Assume (A1-5). Then there exist a $C < +\infty$ such that for any $n \geq 1$, $\mu_\theta \in \mathcal{P}(\mathsf{X})$, $\widetilde{\mu_\theta} \in \mathcal{M}(\mathsf{X})$, $\epsilon > 0$, $\theta \in \Theta$:*

$$\|\widetilde{F}_\theta^n(\mu_\theta, \widetilde{\mu_\theta})\| \vee \|\widetilde{F}_{\theta,\epsilon}^n(\mu_\theta, \widetilde{\mu_\theta})\| \leq C(1 + \|\widetilde{\mu_\theta}\|).$$

Proof. We will consider only $F_\theta^n(\mu_\theta, \widetilde{\mu_\theta})$ as the ABC filter derivative will follow similar calculations, for any $\epsilon > 0$ (with upper-bounds that are independent of ϵ). By [35, Lemma 6.4] we have for $\varphi \in \mathcal{B}_b(\mathsf{X})$

$$\widetilde{F}_\theta^n(\mu_\theta, \widetilde{\mu_\theta})(\varphi) = \widetilde{G}_\theta^n(\mu_\theta, \widetilde{\mu_\theta})(\varphi) + \sum_{p=1}^n \widetilde{G}_\theta^{n-p}(F_\theta^p(\mu_\theta), \widetilde{H}_\theta^p(\mu_\theta))(\varphi).$$

By [35, Lemma 6.6] we have the upper-bound

$$\|\widetilde{F}_\theta^n(\mu_\theta, \widetilde{\mu_\theta})\| \leq C\Big(\rho^n\|\widetilde{\mu_\theta}\| + \sum_{p=1}^n \rho^{n-p}\|\widetilde{H}_\theta^p(\mu_\theta)\|\Big)$$

with $\rho \in (0,1)$. Then by [35, Lemma 6.8], it follows that

$$\|\widetilde{F}_\theta^n(\mu_\theta, \widetilde{\mu_\theta})\| \leq C\Big(\rho^n\|\widetilde{\mu_\theta}\| + \sum_{p=1}^n \rho^{n-p}\Big)$$

from which one concludes. □

Remark D.1. *Using the proof above, one can also show that there exist a $C < +\infty$ such that for any $n \geq 1$, $\mu_\theta \in \mathcal{P}(\mathsf{X})$, $\widetilde{\mu_\theta} \in \mathcal{M}(\mathsf{X})$, $\epsilon > 0$, $\theta \in \Theta$*

$$\|\widetilde{F}_\theta^{(n)}(\mu_\theta, \widetilde{\mu_\theta})\| \vee \|\widetilde{F}_{\theta,\epsilon}^{(n)}(\mu_\theta, \widetilde{\mu_\theta})\| \leq C(1 + \|\widetilde{\mu_\theta}\|).$$

References

[1] Andrieu, C., Doucet, A. & Tadic, V. B. (2005). On-line simulation-based algorithms for parameter estimation in general state-space models, In *Proc. of the 44th IEEE Conference on Decision and Control and European Control Conference (CDC-ECC '05),* 332 – 337, and expanded Technical Report, available at URL http://www.maths.bris.ac.uk/˜maxca/preprints/andrieu_doucet_tadic_2007.pdf.

[2] Arapostathis, A. & Marcus, S.I. (1990) Analysis of an identification algorithm arising in the adaptive estimation of Markov chains, *Mathematics of Control, Signals and Systems,* **3**, 1–29.

[3] Barthelmé, S. & chopin, N. (2011). Expectation-Propagation for summary-less, likelihood-free inference. arXiv:1107.5959 [stat.CO].

[4] Benveniste, A., Métivier, M. and Priouret, P. (1990). *Adaptive Algorithms and Stochastic Approximation.* New York: Springer-Verlag..

[5] Beskos, A., Crisan, D., Jasra, A. & Whiteley, N. (2011). Error bounds and normalizing constants for sequential Monte carlo in high-dimensions. arXiv:1112.1544 [stat.CO].

[6] Bickel, P., Li, B. & Bengtsson, T. (2008). Sharp failure rates for the bootstrap particle filter in high dimensions. In *Pushing the Limits of Contemporary Statistics*, B. Clarke & S. Ghosal, Eds, 318–329, IMS.

[7] Cappé, O., Ryden, T, & Moulines, É. (2005). *Inference in Hidden Markov Models.* Springer: New York.

[8] Cappé, O. (2009). Online sequential Monte Carlo EM algorithm. In *Proc. of IEEE Workshop Statist. Signal Process. (SSP), Cardiff, Wales, UK.*

[9] Calvet, C. & Czellar, V. (2012). Accurate methods for approximate Bayesian computation filtering. Technical Report, HEC Paris.

[10] Cérou, F., Del Moral, P. & Guyader, A. (2011). A non-asymptotic variance theorem for un-normalized Feynman-Kac particle models. *Ann. Inst. Henri Poincare*, **47**, 629–649.

[11] Dean, T. A., Singh, S. S., Jasra, A. & Peters G. W. (2010). Parameter estimation for Hidden Markov models with intractable likelihoods. arXiv:1103.5399 [math.ST].

[12] Dean, T.A. & Singh, S.S. (2011) Asymptotic behavior of approximate Bayesian estimators. arXiv:1105.3655 [math.ST].

[13] Del Moral, P. (2004). *Feynman-Kac Formulae: Genealogical and Interacting Particle Systems with Applications.* Springer: New York.

[14] Del Moral, P., Doucet, A. & Jasra, A. (2006). Sequential Monte Carlo samplers. *J. R. Statist. Soc. B*, **68**, 411–436.

[15] Del Moral, P., Doucet, A., & Jasra, A. (2012). An adaptive sequential Monte Carlo method for approximate Bayesian computation. *Statist. Comp.*, **22**, 1009-1020.

[16] Del Moral, P., Doucet, A. & Singh, S. S. (2009). Forward only smoothing using Sequential Monte Carlo. arXiv:1012.5390 [stat.ME].

[17] Del Moral, P., Doucet, A. & Singh, S. S. (2011). Uniform stability of a particle approximation of the optimal filter derivative. arXiv:1106.2525 [math.ST].

[18] Doucet, A., Godsill, S. & Andrieu, C (2000). On sequential Monte Carlo sampling methods for Bayesian filtering. *Statist. Comp.*, **10**, 197–208.

[19] Gauchi, J. P. & Vila, J. P. (2012). Nonparametric filtering approaches for identification and inference in nonlinear dynamic systems. *Statist. Comp.* (to appear).

[20] Jasra, A., Singh, S. S., Martin, J. S. & McCoy, E. (2012). Filtering via approximate Bayesian computation. *Statist. Comp.*, (to appear).

[21] Kantas, N., Doucet, A., Singh, S.S., Maciejowski, J.M. & Chopin, N. (2011) On Particle Methods for Parameter Estimation in General State-Space Models, submitted.

[22] Le Gland, F. & Mevel, M. (2000) Exponential Forgetting and Geometric Ergodicity in Hidden Markov Models, *Mathematics of Control, Signals and Systems*, 13 , pp 63 – 93.

[23] Le Gland, F. & Mevel, M. (1997). Recursive identification in hidden Markov models. *Proc. 36th IEEE Conf. Decision and Control*, 3468-3473.

[24] Le Gland, F. & Mevel, M. (1995) Recursive Identification of HMM's with Observations in a Finite Set, *Proc. of the 34th Conference on Decision and Control*, pp. 216 – 221.

[25] Lorenz, E.N. (1963) Deterministic Nonperiodic Flow, *Journal of the Atmospherical Sciences,* **20**, 130–141.

[26] Marin, J.-M., Pudlo, P., Robert, C.P., & Ryder, R (2012). Approximate Bayesian Computational methods.*Statist. Comp.*, (to appear).

[27] Martin, J. S., Jasra, A., Singh, S. S., Whiteley, N. & McCoy, E. (2012). Approximate Bayesian computation for smoothing, arXiv:1206.5208 [stat.CO].

[28] McKinley, J., Cook, A. & Deardon, R. (2009). Inference for epidemic models without likelihooods. *Intl. J. Biostat.*, **5**, a24.

[29] Murray, L. M., Jones, E. & Parslow, J. (2011). On collapsed state-space models and the particle marginal Metropolis-Hastings sampler. arXiv:1202.6159 [stat.CO].

[30] Pitt, M. K. (2002). Smooth particle filters for likelihood evaluation and maximization, Technical Report, University of Warwick.

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

[32] Poyiadjis, G., Singh, S. S. & Doucet, A. (2006). Gradient-free maximum likelihood parameter estimation with particle filters.*In proc Amer. Control Conf.*, 6-9.

[33] Spall, J.C. (1992) Multivariate stochastic approximation using a simultaneous perturbation gradient approximation, Automatic Control, IEEE Transactions on, 37-3, 332–341.

[34] Spall, J. (2003). *Introduction to Stochastic Search and Optimization* (1st ed), Wiley: New York.

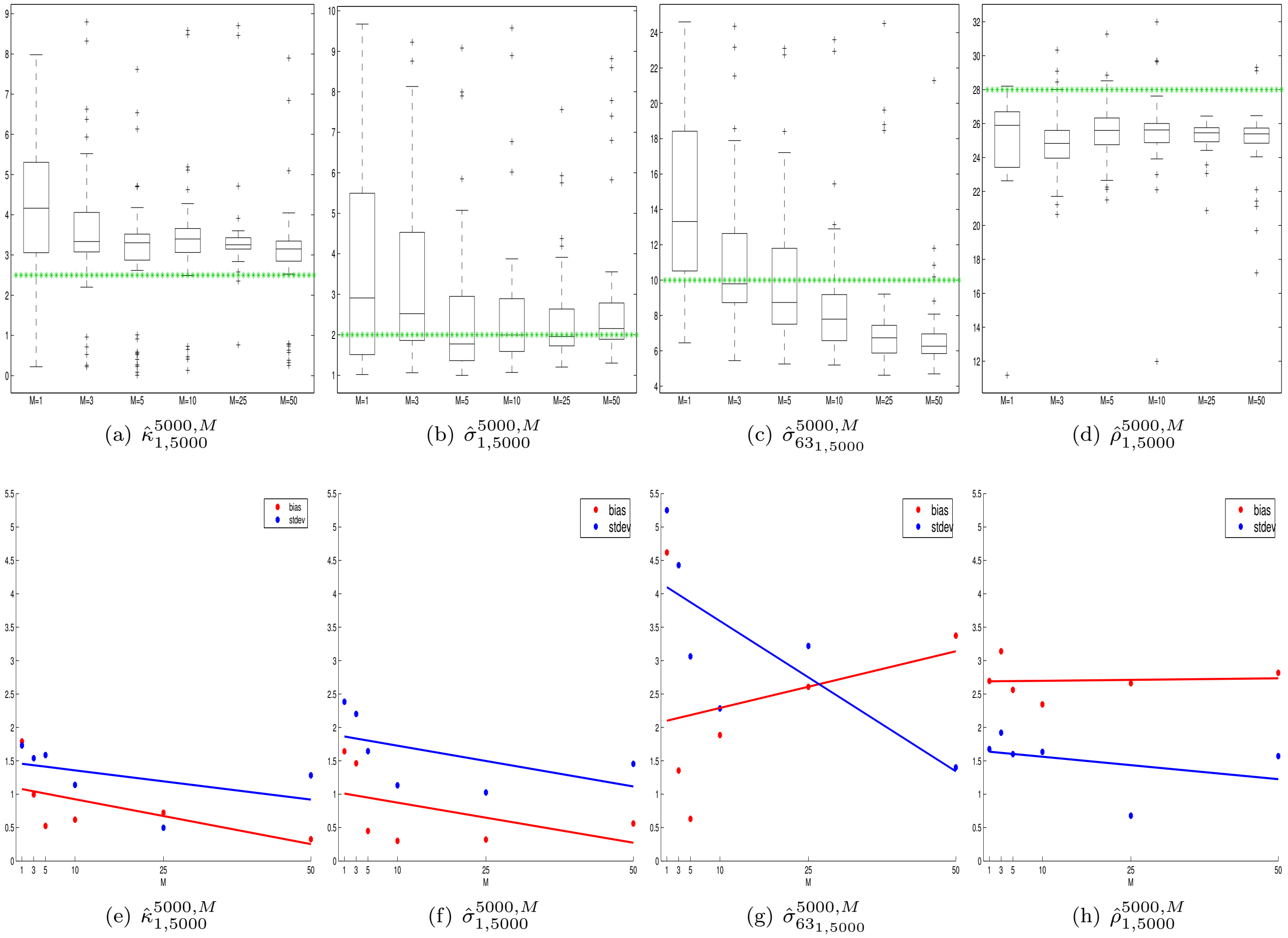
[35] Tadic, V. B. & Doucet, A. (2005). Exponential forgetting and geometric ergodicity for optimal filtering in general state-space models. *Stoch. Proc. Appl.*, **115**, 1408–1436.

[36] Tadic, V. B. (2009) Analyticity, Convergence and Convergence Rate of Recursive Maximum Likelihood Estimation in Hidden Markov Models, arXiv:0904.4264.

[37] Whiteley, N., Kantas, N. & Jasra, A. (2012). Linear variance bounds for particle approximations of time-homogeneous Feynman-Kac formulae. *Stoch. Proc. Appl.*, **122**, 1840–1865.

[38] Yildirim, S., Singh, S.S. & Doucet, A. (2013). An Online Expectation-Maximisation Algorithm for Changepoint Models, *Journal of Computational and Graphical Statistics,* to appear.

[39] Yildirim, S., Dean, T.A., Singh, S.S. & Jasra, A. (2013). Approximate Bayesian Computation for Recursive Maximum Likelihood Estimation in Hidden Markov Models, Technical Report, University of Cambridge.



(a) $\hat{\kappa}^{5000,M}_{1,5000}$

(b) $\hat{\sigma}^{5000,M}_{1,5000}$

(c) $\hat{\sigma}^{5000,M}_{63_{1,5000}}$

(d) $\hat{\rho}^{5000,M}_{1,5000}$

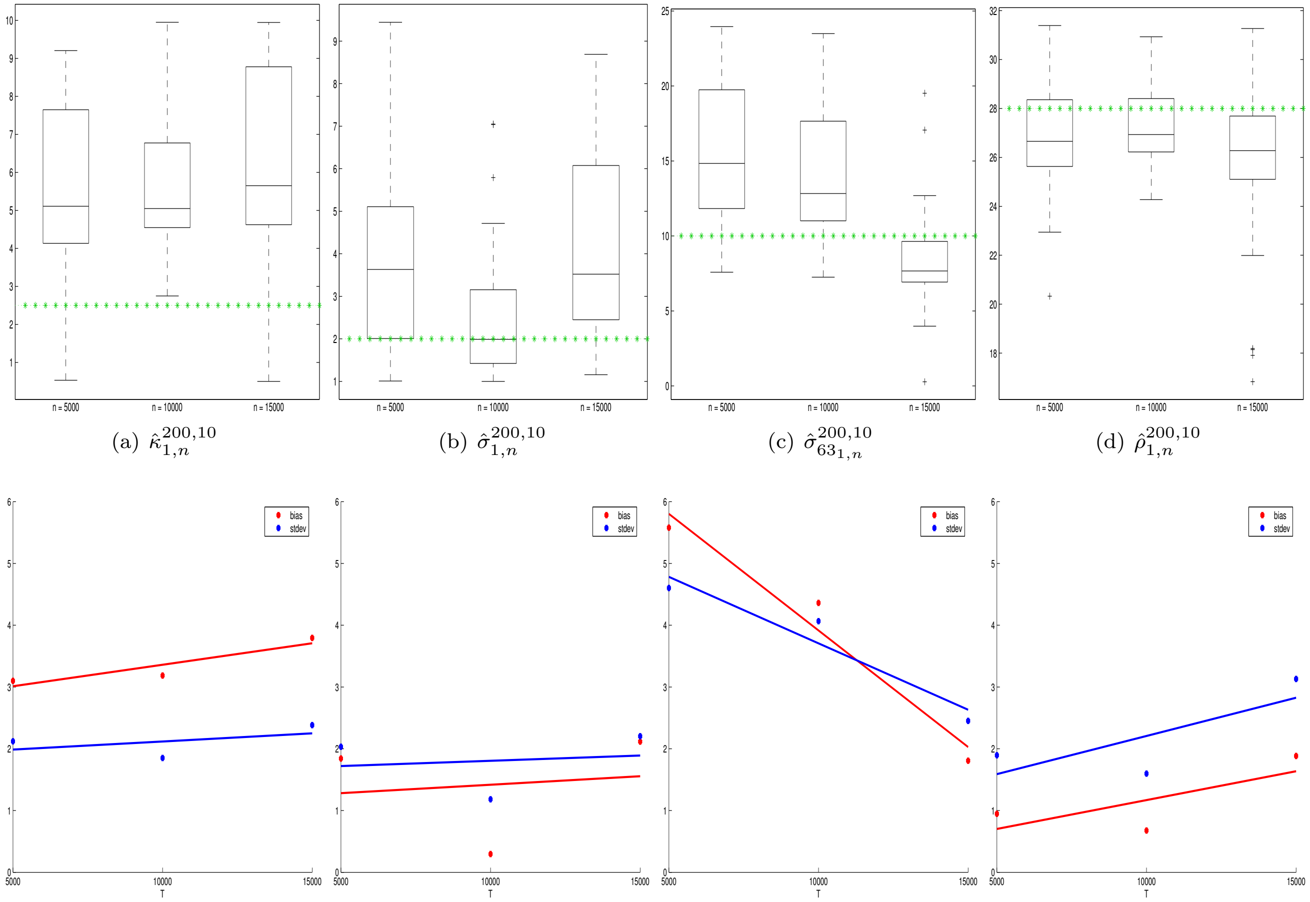
(e) $\hat{\kappa}^{5000,M}_{1,5000}$

(f) $\hat{\sigma}^{5000,M}_{1,5000}$

(g) $\hat{\sigma}^{5000,M}_{63_{1,5000}}$

(h) $\hat{\rho}^{5000,M}_{1,5000}$

Figure 5: $\widehat{\theta}^{5000,M}_{1,5000}$ when estimating $\theta = (\kappa, \sigma, \sigma_{63}, \rho)$ of the Lorenz '63 HMM, using ABC-SMC with values of $M \in \{1, 3, 5, 10, 25, 50\}$. Figures 5(a)-5(d) show the $\widehat{\theta}^{5000,M}_{1,5000}$ in box-plots and their true values in dotted green lines. Figures 5(e)-5(h) show the MC bias and MC standard deviation of the $\widehat{\theta}^{5000,M}_{1,5000}$, in red and blue, with lines of least squared-error.



(a) $\hat{\kappa}_{1,n}^{200,10}$

(b) $\hat{\sigma}_{1,n}^{200,10}$

(c) $\hat{\sigma}_{63_{1,n}}^{200,10}$

(d) $\hat{\rho}_{1,n}^{200,10}$

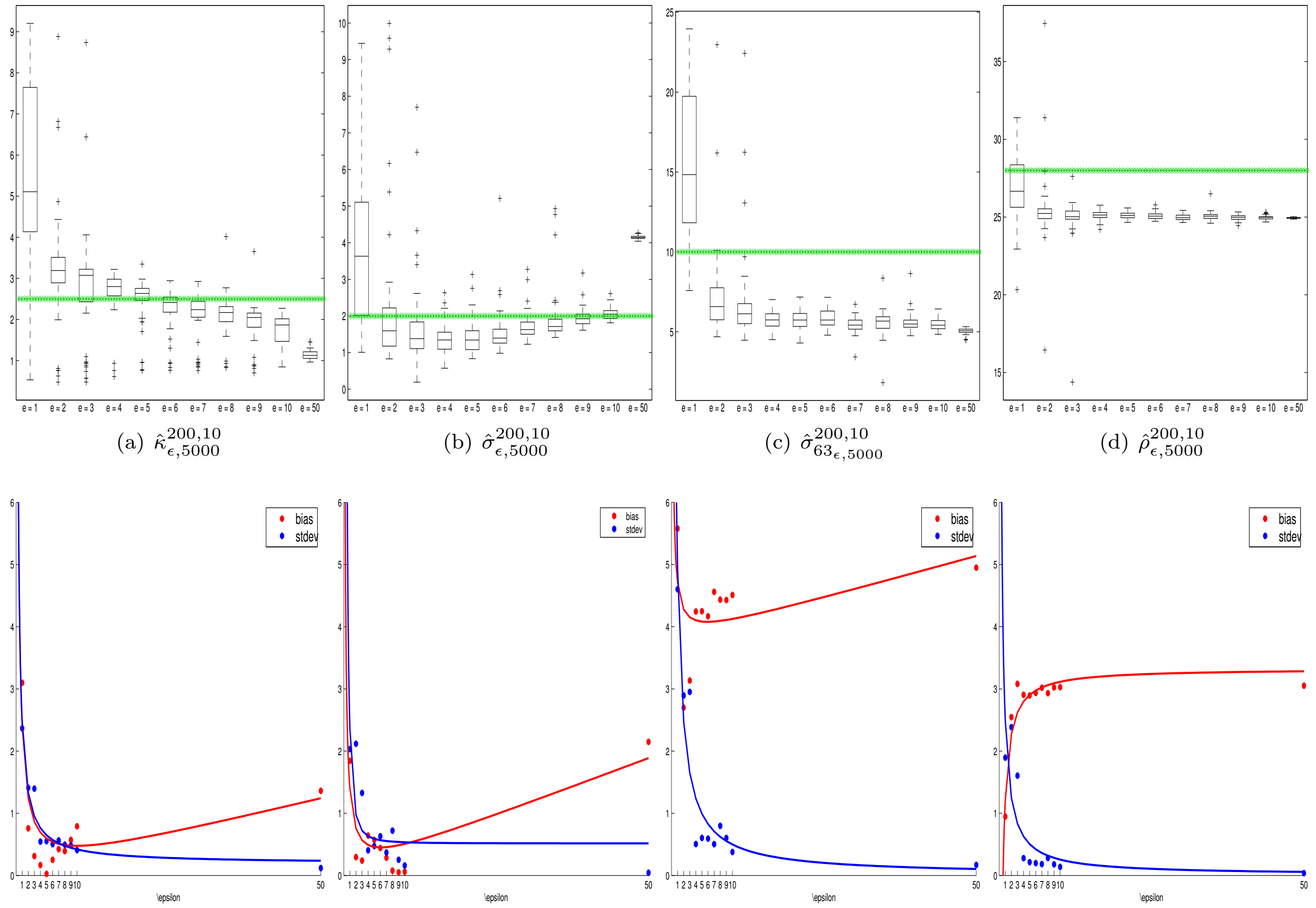
(e) $\hat{\kappa}_{1,n}^{200,10}$

(f) $\hat{\sigma}_{1,n}^{200,10}$

(g) $\hat{\sigma}_{63_{1,n}}^{200,10}$

(h) $\hat{\rho}_{1,n}^{200,10}$

Figure 6: $\widehat{\theta}_{1,n}^{200,10}$ when using ABC-SMC to estimate $\theta = (\kappa, \sigma, \sigma_{63}, \rho)$ of the Lorenz '63 HMM, for datasets of length $n \in \{5000, 10000, 15000\}$. Figures 6(a)-6(d) show the $\widehat{\theta}_{1,n}^{200,10}$ in box-plots and their true values in dotted green lines. Figures 6(e)-6(h) show the MC bias and MC standard deviation of the $\widehat{\theta}_{1,n}^{200,10}$, in red and blue, with lines of least squared-error.



(a) $\hat{\kappa}^{200,10}_{\epsilon,5000}$

(b) $\hat{\sigma}^{200,10}_{\epsilon,5000}$

(c) $\hat{\sigma}^{200,10}_{63_{\epsilon,5000}}$

(d) $\hat{\rho}^{200,10}_{\epsilon,5000}$

(e) $\hat{\kappa}^{200,10}_{\epsilon,5000}$

(f) $\hat{\sigma}^{200,10}_{\epsilon,5000}$

(g) $\hat{\sigma}^{200,10}_{63_{\epsilon,5000}}$

(h) $\hat{\rho}^{200,10}_{\epsilon,5000}$

Figure 7: $\widehat{\theta}^{200,10}_{\epsilon,5000}$ when estimating $\theta = (\kappa, \sigma, \sigma_{63}, \rho)$ of the Lorenz '63 HMM, using ABC-SMC with values of $\epsilon \in \{1, 2, 3, \ldots, 10, 50\}$. Figures 7(a)-7(d) show the MC biases and their curves of non-linear least squared-error proportional to $\epsilon + \frac{1}{\epsilon}$ in red, and the MC standard deviations with their curves of non-linear least squared-error proportional to $\frac{1}{\epsilon}$ in blue.