Numerical methods for sensitivity analysis of Feynman-Kac models

Pierre-Arnaud Coquelin, Romain Deguest, Rémi Munos

Numerical methods for sensitivity analysis of Feynman-Kac models

Pierre-Arnaud Coquelin*†, Romain Deguest*, Rémi Munos†

January 18, 2007

Abstract

The aim of this work is to provide efficient numerical methods to estimate the gradient of a Feynman-Kac flow with respect to a parameter of the model. The underlying idea is to view a Feynman-Kac flow as an expectation of a product of potential functions along a canonical Markov chain, and to use usual techniques of gradient estimation in Markov chains. Combining this idea with the use of interacting particle methods enables us to obtain two new algorithms that provide tight estimations of the sensitivity of a Feynman-Kac flow. Each algorithm has a linear computational complexity in the number of particles and is demonstrated to be asymptotically consistent. We also carefully analyze the differences between these new algorithms and existing ones. We provide numerical experiments to assess the practical efficiency of the proposed methods and explain how to use them to solve a parameter estimation problem in Hidden Markov Models. To conclude we can say that these algorithms outperform the existing ones in terms of trade-off between computational complexity and estimation quality.

Introduction

Let us consider a Hidden Markov Model (HMM) parameterized by $\theta \in \Theta$, where $\Theta \subset \mathbb{R}^{n_\theta}$ is an open subset of $\mathbb{R}^{n_\theta}$. The state process $(X_{p,\theta})_{p\geq 0}$ is an homogeneous Markov chain with initial probability measure $\mu_\theta(dx_0)$ and Markov transition kernel $K_\theta(dx_{p+1}|x_p)$. The observation process $(Y_{p,\theta})_{p\geq 1}$ is linked with the state process by the conditional probability measure $\mathbb{P}(Y_p \in dy_p|X_p = x_p) = g_\theta(y_p, x_p)\lambda(dy_p)$, where λ is a fixed probability measure. Suppose we are given a sequence of successive realizations of the observation process $(y_{p,\theta^*})_{p\geq 1}$ which were obtained using an unknown parameter denoted by θ^*. Our objective is to recover θ^* using the sequence of observations $(y_{p,\theta^*})_{p\geq 1}$. In the following, we make a slight abuse of notation by writing $g_{p,\theta}(x_p) = g_\theta(y_{p,\theta^*}, x_p)$ and we adopt

*CMAP, Ecole Polytechnique (France), Email: name@cmapx.polytechnique.fr
†SequeL team, INRIA futurs Lille (France), Email : firstname.name@inria.fr

the usual notation $z_{i:j} = (z_k)_{i \leq k \leq j}$.

It is known that, under some technical conditions on HMMs [8], the maximum likelihood estimator (MLE) provides an asymptotically consistent estimation of θ^*. Indeed, by defining the log-likelihood function at the time step n as $l_n(\theta) = \log \mathbb{P}(Y_{1:n,\theta} \in dy_{1:n,\theta^*})$, the sequence of estimated parameters $\theta_n = \text{argmax}_{\theta \in \Theta} \, l_n(\theta)$ converges to the true parameter θ^*. Thus, the problem of parameter estimation is transposed into the problem of maximizing the log-likelihood function. Except for some particular cases such as finite state space models or linear gaussian models, l_n cannot be analytically computed and requires to be numerically estimated (for example using particle filters [12]). In these situations, maximizing the log-likelihood function may be considered as a stochastic optimization problem. A natural approach to solve it is to use a stochastic gradient ascent on $l_n(\theta)$. This consists in applying a Robbin-Monroe procedure to search for a zero of the differential of the log-likelihood function ∇l_n with respect to the parameter θ. The crucial point for the practical success of such an optimization procedure is to provide a tight estimation $\widehat{\nabla l_n} \approx \nabla l_n$ of the gradient of the log-likelihood function.

The initial problem of parameter estimation in HMMs may be undergone by computing a tight estimator of the gradient of the log-likelihood function. Using elementary bayesian calculus (see [12] for details), one can show that:

$$l_n(\theta) = \sum_{p=1}^{n} \log[\pi_{p|p-1,\theta}(g_{p,\theta})], \tag{1}$$

where we define the predicted filtering distribution $\pi_{p|p-1,\theta}(dx_p) = \mathbb{P}(X_{p,\theta} \in dx_p | Y_{1:p-1,\theta} = y_{1:p-1,\theta^*})$ for $p \geq 1$, and where $\pi(g)$ refers to $\int g(x)\pi(dx)$ (all precise definitions will be detailed in the Notations Section). Differentiating Equation (1) leads to:

$$\nabla l_n(\theta) = \sum_{p=1}^{n} \frac{\nabla[\pi_{p|p-1,\theta}(g_{p,\theta})]}{\pi_{p|p-1,\theta}(g_{p,\theta})}. \tag{2}$$

In order to evaluate the gradient of the log-likelihood, we need to compute the flows $\pi_{p|p-1,\theta}(g_{p,\theta})$ and the gradient of these flows $\nabla[\pi_{p|p-1,\theta}(g_{p,\theta})]$. The flows $\pi_{p|p-1,\theta}(g_{p,\theta})$ can be evaluated using standard algorithms as particle filters. Nonetheless, the problem of efficiently estimating the gradient of the flows $\nabla[\pi_{p|p-1,\theta}(g_{p,\theta})]$ is still an open question. It is the subject of the present contribution.

Before going further, let us introduce some new notations. It is well known that $\pi_{p|p-1,\theta}$ satisfies the following bayesian recurrence relation (see for example [13]):

$$\pi_{p+1|p,\theta}(dx_{p+1}) \quad = \quad \int_{\mathcal{X}_p} \frac{K_\theta(dx_{p+1}|x_p) g_\theta(x_p) \pi_{p|p-1,\theta}(dx_p)}{\int_{\mathcal{X}_p} g_\theta(x_p) \pi_{p|p-1,\theta}(dx_p)}. \tag{3}$$

Using the Boltzman-Gibbs measure $\Psi_{g,\pi}(dx) = \frac{g(x)\pi(dx)}{\pi(g)}$, the relation (3) becomes

$$\pi_{p+1|p,\theta} = \Psi_{g_{p,\theta}, \pi_{p|p-1,\theta}} K_\theta. \tag{4}$$

More generally, given a sequence of positive measurable functions $(G_p)_{p\geq 0}$ and a Markov chain with initial probability distribution μ and transition kernels $(M_p)_{p\geq 1}$, the sequence of marginal predicted Feynman-Kac measures $(\eta_p)_{p\geq 0}$ is recursively defined by the following measure-valued equations:

$$\begin{aligned} \eta_0 &= \mu \\ \eta_{p+1} &= \Psi_{G_p,\eta_p} M_{p+1}. \end{aligned} \tag{5}$$

Equation (4) shows that the predicted filtering distribution $\pi_{p+1|p,\theta}$ is equal to the predicted Feynman-Kac measure η_p associated to the Markov chain with initial measure μ and transition kernels $M_p = K_\theta$, and to the potential functions $G_p = g_p$.

Similarly to what has been described earlier, in the Feynman-Kac framework, the initial parameter estimation problem may be turned into the problem of estimating the sensitivity $\nabla\eta_n(f)$ of a Feynman-Kac flow with respect to a parameter of a Feynman-Kac model. The framework of Feynman-Kac formulae [17] is a subject of high interest which appears in many applications besides Hidden Markov Models, such as directed polymer simulations, random excursion models or genetic optimization algorithms. In the remainder of this paper, we adopt the general Feynman-Kac formalism. It provides more general results, a better physical intuition of measure representations via interacting particle interpretation, and it allows to use remarkable convergence results.

To the best of our knowledge, the problem of gradient estimation of flows was never addressed directly in the Feynman-Kac framework, but it is the object of a lot of efforts in the field of HMMs. Previous works can be cast into three categories: joint particle estimation of the linear tangent filter and the filter by differentiating the update and prediction equations of the filter [15, 16, 3], joint particle estimation of the linear tangent filter and the filter by differentiating the prediction equation and using a likelihood ratio approach to deal with the update step [14, 19], and direct sensitivity analysis of the Markov chain followed by the particles [21, 20].
More precisely, [15, 16, 3] address the problem of jointly representing by a particle population the linear tangent measure of the filter and the measure of the filter. A representation of the linear tangent transition kernel from which one can sample is assumed to be given. It is used to compute some linear tangent update and prediction equations. The algorithm transports a joint particle approximation of the filter and its gradient using these linear tangent equations. An extensive compilation of experimental results on linear gaussian models is available in [3].
In [14] the authors address the same problem as above. The filter and its gradient are represented by the same particle population, but using different weights. As the population is selected using the weights associated to the filter, the weights associated to the linear tangent filter are multiplied by likelihood ratios during selection procedures. This operation prevents the algorithm from being efficient because of the resulting

additional variance term which explodes with the number of time-steps. In [19] the authors start with the same idea as in [14], but instead, they used a marginal particle filter to compute the weights associated to the particle approximation of the linear tangent filter. This approach solves the stability problem encountered in previous works. Unfortunately, the computational complexity is quadratic in the number of particles, thus preventing this method to be applied to real world applications.
In [21, 20], the authors proposed to approximate the sensitivity of the filter by the sensitivity of the Markov chain followed by its particles representation. Usual gradient evaluation techniques issued from controlled Markov chain theory are used. The obtained algorithms suffer from high variance and are not not of practical interest. As far as we know, at this time, [19] is the state of art method.

The aim of this work is to design efficient numerical methods for estimating the sensitivity of a Feynman-Kac flow with respect to a parameter of a the Feynman-Kac model. Our contribution consists in providing two sensitivity estimates, proven to be consistent, whose approximation quality is comparable to [19], but whose computational complexity is only linear in the number of particles.

The paper is divided into three parts. In a first section we give a formal introduction to Feynman-Kac models in order to precisely set the gradient estimation problem. The second part is the heart of this contribution. The basic idea is to view a Feynman-Kac flow as an expectation of the product of the potential functions G_p along the canonical chain $(X_p)_{p\geq 0}$. Using basic calculus on this expectation provides a formal expression of the gradient of a F-K flow in the form of an integral over the F-K path. Any interacting particle method can be used to estimate tightly this integral and in the same time, to provide a tight estimation of the sensitivity of the F-K flow. The proposed algorithms have a linear computational complexity in the number of particles and a Law of Large Numbers is stated. In the third part we provide numerical experiments, on two different gradient evaluation problems in HMMs, to assess the practical efficiency of the proposed algorithms. In particular, we compare the bias, variance and CPU time of these algorithms with existing methods. We experimentally show that the approximation error does not increase with the number of time steps and that this error tends to zero when the number of particles tends to infinity. We also show, since this was our initial motivation for this work, that these algorithms can be used to solve the parameter estimation problem in nonlinear HMMs.

Notations

All random variables are defined on a common probability space $(\Omega, \mathcal{F}, \mathcal{P})$. Most of random variables, measures and functions depend on a parameter $\theta \in \Theta$ where Θ is an open subset of $\mathbb{R}^{n_\theta}$. Nonetheless, when there is no ambiguity, we omit to note explicitly this dependancy.
A state space $(E, \sigma(E))$ is said to be general if its σ-field $\sigma(E)$ is countably generated. In the following all state spaces are assumed to be general and

for any $d > 0$, $\mathbb{R}^d$ is equipped with the Borel σ-algebra $\sigma(\mathbb{R}^d)$. We will use the following notations:

- The gradient operator ∇ denotes the derivative with respect to θ, and the symbol $'$ denotes the derivative with respect to the state variable.
- To each state space E, we associate the following sets: $\mathcal{D}(E)$, the set of probability measures on E, $B(E)$ the space of measurable functions from E to $\mathbb{R}$ and $B^+(E) = \{f \in B(E) | f \geq 0, f \neq 0\}$ the space of positive measurable functions from E to $\mathbb{R}$ that are not everywhere null. All along the paper we work with test functions defined from E to $\mathbb{R}$, but the results obtained remain the same if one works with functions defined from E to $\mathbb{R}^m$ where $m \geq 1$.
- Given two state spaces E and F, a Markov transition kernel M from E to F is an application $M : E \times \sigma(F) \to [0, 1]$ such that $\forall A \in \sigma(F), \{x \mapsto M(x, A)\} \in B(E)$ and $\forall x \in E, \{A \mapsto M(x, A)\} \in \mathcal{D}(F)$.
- Given $\mu \in D(E)$ a probability measure on E, M a Markov transition kernel from E to F, $f \in B(F)$ a measurable function on F and $x \in E$, we adopt the usual following notations: $\mu(f) = \int f(x)\mu(dx)$, $\mu M(dx') = \int_E \mu(dx)M(x, dx')$.
- Given a random variable X on E we denote by $\mathcal{L}_X$ its law. We have $\forall f \in B(E)$, $\mathbb{E}[f(X)] = \mathcal{L}_X(f)$.
- Given two measures ν and μ on E, ν is absolutely continuous with respect to μ (we write $\nu \gg \mu$) if $\nu(dx) = 0$ implies $\mu(dx) = 0$. Moreover if μ is sigma finite, then the Radon-Nikodym derivative of ν with respect to μ exists and is written $\frac{d\nu}{d\mu}$.
- For any $k \in \mathbb{N}^*$, $\mathbb{R}^k$ is considered as a topological vector space with norm $\|.\|$ and the topological structure $(\mathbb{R}^k, \mathcal{O}(\mathbb{R}^k))$ is induced by its norm. For each $\Theta \subset \mathbb{R}^{n_\theta}$ open subset of $\mathbb{R}^{n_\theta}$, one denotes by $\mathcal{V}_\Theta(\theta) = \{A \in \mathcal{O}(\Theta) | \theta \in A\}$ the set of all neighborhoods of θ in Θ.

1 Introduction to F-K models and gradient estimation of F-K flows

1.1 Feynman-Kac models

In this section we introduce relevant elements of Feynman-Kac theory which will be used all along this paper. For an extensive presentation of theoretical results related to F-K formulae, we refer the reader to the book [17], and for a good overview of possible applications in the area of particle methods, to the book [12].

Let $(E_p, \mathcal{E}_p)_{p \geq 0}$ be a sequence of measurable spaces, and $(X_p)_{p \geq 0}$ be a Markov chain from $(E_{p-1}, \mathcal{E}_{p-1})$ to $(E_p, \mathcal{E}_p)$ with initial measure μ defined on $(E_0, \mathcal{E}_0)$, verifying:

$$X_{p+1} = F_{p+1}(X_p, U_{p+1}), \text{ where } U_{p+1} \sim \nu$$

where $\nu \in \mathcal{D}(E)$ is a probability measure on the measurable space $(U, \sigma(U))$. The transition kernels of the Markov chain $(X_p)_{p\geq 0}$ are $(M_p)_{p\geq 1}$, and moreover $(X_p)_{p\geq 0}$ is called the canonical Markov chain. Consider a collection of bounded $\mathcal{E}_p$-measurable positive functions $G_p \in B^+(\mathcal{E}_p)$ such that:

$$\forall k \geq 0, \quad \mathbb{E}\left[\prod_{p=0}^{k} G_p(X_p)\right] > 0.$$

To a given Feynman-Kac model $(\mu, (M_p)_{p\geq 1}, (G_p)_{p\geq 0})$ and a time index n is associated three measures:

- the updated F-K path measure $\hat{\mathbb{Q}}_n$:

$$\hat{\mathbb{Q}}_n(dx_{0:n}) = \frac{\prod_{p=0}^{n} G_p(x_p) \prod_{p=1}^{n} M_p(x_{p-1}, dx_p)\mu(dx_0)}{\mathbb{E}\left[\prod_{p=0}^{n} G_p(X_p)\right]},$$

- the marginal updated F-K measure $\hat{\eta}_n$ is the marginalization of $\hat{\mathbb{Q}}_n$:

$$\hat{\eta}_n(dx_n) = \hat{\mathbb{Q}}_n(E_0 \times \cdots \times E_{n-1} \times dx_n),$$

- the unnormalized marginal updated F-K measure $\hat{\gamma}_n$:

$$\hat{\gamma}_n(dx_n) = \int_{E_0 \times \cdots \times E_{n-1}} \prod_{p=0}^{n} G_p(x_p) \prod_{p=1}^{n} M_p(x_{p-1}, dx_p)\mu(dx_0).$$

We immediately make a few elementary remarks. First of all, we have the relation $\hat{\eta}_n = \frac{\hat{\gamma}_n}{\hat{\gamma}_n(1)}$, where 1 denotes here the constant function equals to 1 everywhere. Secondly, as announced in the introduction, the time evolution equation associated to the marginal predicted Feynman-Kac measure (defined by $\eta_{n+1} = \hat{\eta}_n \; M_{n+1}$) can be syntactically expressed using the Boltzman-Gibbs measure: $\eta_{n+1} = \Psi_{G_{n+1},\eta_n} M_{n+1}$. Finally, given a test function $f \in B(E_n)$, the updated F-K flow $\hat{\eta}_n(f)$ trough f admits a simple expression as a ratio of two expectations along the canonical Markov chain $(X_p)_{p\geq 0}$:

$$\hat{\eta}_n(f) = \frac{\mathbb{E}\left[f(X_n)\prod_{p=0}^{n} G_p(X_p)\right]}{\mathbb{E}\left[\prod_{p=0}^{n} G_p(X_p)\right]}. \tag{6}$$

We adopt the following convention $X_0 = F_0(X_{-1}, U_0)$, where $U_0 \sim \nu$, and the initial transition kernel verifies $M_0(x_{-1}, .) = \mu$.

1.2 Interacting particle methods (IPM)

One of the main issue in F-K formulae theory is to compute an empirical measure that approximates the marginal F-K measure $\hat{\eta}_n$ by a weighted sum of Dirac measures $\sum_i w_{i,n}\delta_{\xi_{i,n}}$. The most famous methods for computing such an approximation are interacting particle methods (IPM) (also called particle filters or sequential monte carlo methods). IPM can

be interpreted from many different points of view. First of all, it can be seen as a McKean linearization of the non-linear updating Equation (5) (see [17] for details). Secondly, one can see IPM as a way to sample from a distribution $\frac{\varrho M g}{\varrho M(g)}$ when knowing only how to sample from ϱ and from the Markov transition kernel M (see [9] for details). Thirdly, it can be thought as a voting method that allows to track some signal in the spirit of general bootstrap methods. For our purpose we will think of IPM as stochastic numerical methods that allow to estimate efficiently integrals of the form (6).

A weighted sample $\{\Xi_p^N = (w_{i,p}^N, \xi_{i,p}^N)_{1\leq i\leq N}\}_{N\geq 0}$ is a triangular array of random variables such that $w_{i,p}^N > 0, \sum_i w_{i,p}^N = 1$ and $\sum_i (w_{i,p}^N)^2 \xrightarrow{\mathcal{P}} 0$ when $N \to \infty$. IPM propagate a weighted sample along a Feynman-Kac model in three steps:

Prediction. Define $\tilde{\Xi}_{p+1}^N = (\tilde{w}_{i,p+1}^N, \tilde{\xi}_{i,p+1}^N)_{1\leq i\leq N}$ with: $\tilde{w}_{i,p+1}^N = w_{i,p}^N$ and $\tilde{\xi}_{i,p+1}^N \overset{id}{\sim} M_{p+1}(\xi_{i,p}^N, \cdot)$,

Updating. Define $\hat{\Xi}_{p+1}^N = (\hat{w}_{i,p+1}^N, \hat{\xi}_{i,p+1}^N)_{1\leq i\leq N}$ with: $\hat{w}_{i,p+1}^N = \frac{\tilde{w}_{i,p+1}^N G_{p+1}(\tilde{\xi}_{i,p+1}^N)}{\sum_j \tilde{w}_{j,p+1}^N G_{p+1}(\tilde{\xi}_{j,p+1}^N)}$ and $\hat{\xi}_{i,p+1}^N = \tilde{\xi}_{i,p+1}^N$,

Selection. $\Xi_{p+1}^N = \mathcal{S}(\hat{\Xi}_{p+1}^N)$ where $\mathcal{S}$ is a selection operator.

The selection operator is designed to reduce the variance of the weights. In literature, selection operators $\mathcal{S}$ are often designed such that the weights have no variance after selection and such that the selection step adds no bias:

$$\begin{aligned} \mathcal{S}((\hat{w}_j, \hat{\xi}_j)_{1\leq j\leq N}) &= \left(\frac{1}{N}, \hat{\xi}_{k_{1:N}}\right) \\ \mathbb{E}\left[\frac{1}{N}\sum_i f(\xi_i) \,|\, \hat{\Xi}\right] &\overset{\mathcal{P}}{=} \sum_i \hat{w}_i f(\hat{\xi}_i), \end{aligned} \tag{7}$$

where $k_i \in \{1, \cdots, N\}$ denotes the i-th selection index used by the selection operator.
Many procedures have been built in order to add a minimum of variance in the selection process and to be computationally efficient. The most popular are the multinomial selection, the stratified selection and the residual selection. We refer the reader to [2] for a theoretical comparative study of the main selection schemes, and to [1] for an experimental analysis of them. In the remainder of this paper, we consider a selection scheme that satisfies Equation (7).

IPM generate a sequence of these three steps, which may be sketched in the following diagram:

$$\Xi_0^N \longrightarrow \ldots \longrightarrow \Xi_p^N \overset{\text{Prediction}}{\longrightarrow} \tilde{\Xi}_{p+1}^N \overset{\text{Update}}{\longrightarrow} \hat{\Xi}_{p+1}^N \overset{\text{Selection}}{\longrightarrow} \Xi_{p+1}^N \longrightarrow \ldots \tag{8}$$

where $\forall i \in \{1, \ldots, N\}, \xi_{i,0}^N \overset{iid}{\sim} \mu$ and $w_{i,0}^N = 1/N$.

Many asymptotic convergence results (when $N \to \infty$) of IPM-based estimate towards $\hat{\eta}_n$ can be derived. One can cite for instance: a law a large number, a L^q convergence, a central limit theorem, a large and a moderate deviation result. A review of these results is provided in [17]. The interested reader can find elementary proofs in [6], and some original approaches in [18, 9, 6, 7, 5]. For our purpose we will only use the following Law of Large Numbers derived from [9]:

Theorem 1 (Law of Large Number for Interacting Particle Methods)**.** *Using Procedure (8), for any $n \geq 0$ we have:*

$$\forall f \in B_n, \quad \tfrac{1}{N} \textstyle\sum_i f(\xi_{i,n}^N) \xrightarrow{\mathbb{P}} \hat{\eta}_n(f),$$

with $(B_p)_{p \geq 0}$ defined by induction as follows: $B_0 = L^1(\hat{\eta}_0)$, and $B_{p+1} = \{f_{p+1} \in L^1(\hat{\eta}_{p+1}) | M_{p+1}(\cdot, G_{p+1} f_{p+1}) \in B_p\}$.

In the prediction step, one may sample using a sequence of instrumental kernels $(Q_p)_{p \geq 0}$ satisfying $\forall p \geq 0, \forall x_p \in E_p, Q_{p+1}(x_p, .) \ll M_{p+1}(x_p, .)$, instead of sampling using the transition kernels $(M_p)_{p \geq 0}$ of the Feynman-Kac flow. If it is the case the weighting step is slightly different because the weights must be multiplied by the Radon-Nikodym derivative $\frac{dM_p}{dQ_p}$. Such importance sampling approaches can lead to a tighter estimate of (6), and can be applied to the methods that we will develop in next chapters of this paper. Nonetheless, in order to keep the clarity of the redaction we do not consider importance sampling via an instrumental kernel in the core of the text. We will give a deeper explanation of this remark in Section 3.1.1.

1.3 Gradient estimation of F-K flows

Let us now work with a Feynman-Kac model parameterized by a n_θ-dimensional parameter $\theta \in \Theta \subset \mathbb{R}^{n_\theta}$. For simplicity, we omit the θ index, writing M_p for $M_{\theta,p}$, μ for μ_θ, and G_p for $G_{\theta,p}$. We are interested in evaluating $\nabla \hat{\eta}_n(f)$, the gradient of the updated Feynman-Kac flow with respect to θ for f in a suitable class of functions. Recalling that $\hat{\eta}_n(f) = \frac{\hat{\gamma}_n(f)}{\hat{\gamma}_n(1)}$, we have:

$$\nabla[\hat{\eta}_n(f)] = \nabla \left[\frac{\hat{\gamma}_n(f)}{\hat{\gamma}_n(1)} \right] = \frac{\nabla[\hat{\gamma}_n(f)]}{\hat{\gamma}_n(1)} - \hat{\eta}_n(f) \frac{\nabla[\hat{\gamma}_n(1)]}{\hat{\gamma}_n(1)}. \tag{9}$$

Interacting particle methods are well suited to approximate $\hat{\eta}_n(f)$. Consequently the problem is now to estimate $\frac{\nabla[\hat{\gamma}_n(f)]}{\hat{\gamma}_n(1)}$.

As we have already noticed in Equation (6), $\hat{\gamma}_n(f)$ can explicitly be expressed in an integral form: $\hat{\gamma}_n(f) = \mathbb{E}\left[f(X_n) \prod_{p=0}^n G_p(X_p) \right]$. The term $\frac{\nabla \hat{\gamma}_n(f)}{\hat{\gamma}_n(1)}$ that remains to be evaluated in Equation (9) can be expressed as the gradient of an expectation along the canonical Markov chain:

$$\frac{\nabla \hat{\gamma}_n(f)}{\hat{\gamma}_n(1)} = \frac{\nabla \mathbb{E}\left[f(X_n) \prod_{p=0}^n G_p(X_p) \right]}{\hat{\gamma}_n(1)}. \tag{10}$$

This simple point of view allows us to apply, under quite general assumptions on $(X_p)_{p\geq 0}$, usual methods for gradient of expectations, such as the infinitesimal perturbation method or the score method.

2 Methods for gradient estimation of F-K flows

Let us state the following assumption:

Assumption 1. For any $0 \leq p \leq n$:

1. the transition function is smooth which means that for any $u \in U$ the function $(\theta, x_{p-1}) \mapsto F_{\theta,p}(x_{p-1}, u)$ is differentiable with respect to (θ, x_{p-1}).
2. the potential function G_p and the test function f are differentiable with respect to (θ, x_p),
3. for any $\theta \in \Theta$, the random variable $f(X_n)$ verifies:

$$\exists \mathcal{V} \in \mathcal{V}_\Theta(0), \exists C_\theta \in L^1(\mathcal{P}), \forall h \in \mathcal{V}, \|f_{\theta+h}(X_{\theta+h,n}) - f_\theta(X_{\theta,n})\| \overset{a.s.}{\leq} \|h\| C_\theta.$$

We notice that under Assumption 1, for any $0 \leq p \leq n$, X_p is differentiable with respect to θ. We write ∇X_p its gradient. We also remark that the random variable $f(X_p)$ is in $\mathcal{K}^1(\Theta, \mathcal{P})$ (see Appendix for the definition of $\mathcal{K}^1(\Theta, \mathcal{P})$). Moreover, as the functions G_p are all bounded, $f(X_n)\prod_{p=0}^n G_p(X_p)$ remains in $\mathcal{K}^1(\Theta, \mathcal{P})$, which is a result of Proposition 2 of Appendix. Therefore we can use Proposition 1 of Appendix to insert the gradient inside the expectation operator and deduce:

$$\begin{aligned}
\nabla\hat{\gamma}_n(f) &= \nabla\mathbb{E}\left[f(X_n)\prod_{p=0}^n G_p(X_p)\right] \\
&= \mathbb{E}\left[\nabla[f(X_n)]\prod_{p=0}^n G_p(X_p) + f(X_n)\nabla\left[\prod_{p=0}^n G_p(X_p)\right]\right] \\
&= \mathbb{E}\left[(\nabla f(X_n) + f'(X_n)\nabla X_n)\prod_{p=0}^n G_p(X_p)\right] \\
&+ \mathbb{E}\left[f(X_n)\left\{\sum_{p=0}^n \frac{G_p'(X_p)\nabla X_p + \nabla G_p(X_p)}{G_p(X_p)}\right\}\prod_{p=0}^n G_p(X_p)\right] \quad (11)
\end{aligned}$$

We define an augmented canonical Markov chain $(X_p, Z_p, R_p)_{p\geq 0}$ by the recursive relations:

$$\begin{aligned}
X_p &= F_p(X_{p-1}, U_p), \text{ where } U_p \sim \nu & (12)\\
Z_p &= \nabla F_p(X_{p-1}, U_p) + F_p'(X_{p-1}, U_p)Z_{p-1} & (13)\\
R_p &= R_{p-1} + \frac{G_p'(X_p)Z_p + \nabla G_p(X_p)}{G_p(X_p)}, & (14)
\end{aligned}$$

where we used the usual matrix notation for the derivatives, ie. Z_p is an $n_{E_p} \times n_\theta$ matrix. F_p' is an $n_{E_p} \times n_{E_p}$ matrix with i, j element $\frac{\partial F_p^i}{\partial x^j}$, and

∇F_p is an $n_{E_p} \times n_\theta$ matrix with i,j element $\frac{\partial F_p^i}{\partial \theta^j}$. Using this Markov chain we have:

$$\begin{aligned}\frac{\nabla \hat{\gamma}_n(f)}{\hat{\gamma}_n(1)} &= \mathbb{E}\left[(\nabla f(X_n) + f'(X_n)Z_n + f(X_n)R_n)\frac{\prod_{p=0}^n G_p(X_p)}{\hat{\gamma}_n(1)}\right] \\ &= \hat{\zeta}_n(\nabla f + A(f) + B(f)) \end{aligned} \tag{15}$$

where A and B are two linear operators such that $A(f)(x,z,r) = f'(x)z$ and $B(f)(x,z,r) = f(x)r$, and $\hat{\zeta}_n$ is the marginal Feynman-Kac measure associated to the potential functions $(G_p)_{p\geq 0}$ and to the augmented canonical Markov chain $(X_p, Z_p, R_p)_{p\geq 0}$.

Plugging together the Equation (15) and (9) leads to the following estimation of the gradient of Feynman-Kac flow:

$$\begin{aligned}\nabla[\hat{\eta}_n(f)] &= \nabla\left[\frac{\hat{\gamma}_n(f)}{\hat{\gamma}_n(1)}\right] = \frac{\nabla[\hat{\gamma}_n(f)]}{\hat{\gamma}_n(1)} - \hat{\eta}_n(f)\frac{\nabla[\hat{\gamma}_n(1)]}{\hat{\gamma}_n(1)} \\ &= \hat{\zeta}_n(\nabla f + A(f) + B(f)) - \hat{\eta}_n(f)\hat{\zeta}_n(A(1) + B(1)) \\ &= \hat{\zeta}_n(\nabla f + A(f) + B(f) - \hat{\eta}_n(f)B(1)). \end{aligned} \tag{16}$$

The right hand side of Equation (16) can be evaluated using a numerical method that approximates the marginal Feynman-Kac measure $\hat{\zeta}_n$. For example, in the following theorem we use Theorem 1 to estimate it using an Interacting Particle Method:

Theorem 2 (Law of Large Number for gradient estimation using IPA)**.**

Under Assumption 1 and for any $f \in B_n$, we have the following asymptotic estimator for the gradient of the Feynman-Kac flow $\hat{\eta}_n(f)$:

$$\frac{1}{N}\sum_{i=1}^N \left(\nabla f(x_{i,n}^N) + f'(x_{i,n}^N)z_{i,n}^N + f(x_{i,n}^N)\left(r_{i,n}^N - \frac{1}{N}\sum_j r_{j,n}^N\right)\right) \xrightarrow{\mathbb{P}} \nabla\hat{\eta}_n(f),$$

where $\Xi_n^N = ((x_{i,n}^N, z_{i,n}^N, r_{i,n}^N), w_{i,n}^N)$ is the weighted sample obtained using an IPM on the Feynman-Kac model associated to the potential functions $(G_p)_{p\geq 0}$ and to the augmented canonical Markov chain $(X_p, Z_p, R_p)_{p\geq 0}$.

Proof. Starting from Equation (16) and using two times Theorem 1 gives:

$$\begin{aligned}\nabla[\hat{\eta}_n(f)] &= \hat{\zeta}_n(\nabla f + A(f) + B(f) - \hat{\eta}_n(f)B(1)) \\ &\xleftarrow{\mathbb{P}} \frac{1}{N}\sum_{i=1}^N \nabla f(x_{i,n}^N) + \frac{1}{N}\sum_{i=1}^N f'(x_{i,n}^N)z_{i,n}^N \\ &\quad + \frac{1}{N}\sum_{i=1}^N f(x_{i,n}^N)r_{i,n}^N - \left(\frac{1}{N}\sum_{i=1}^N f(x_{i,n}^N)\right)\left(\frac{1}{N}\sum_j r_{j,n}^N\right) \\ &= \frac{1}{N}\sum_{i=1}^N \left(\nabla f(x_{i,n}^N) + f'(x_{i,n}^N)z_{i,n}^N + f(x_{i,n}^N)\left(r_{i,n}^N - \frac{1}{N}\sum_j r_{j,n}^N\right)\right) \end{aligned}$$

□

The estimator given in Theorem 2 admits a recursive implementation whose computational complexity is linear in the number of particles N. A pseudo-code for its computation is defined at any stage p by:

1. Sample independently $u_{i,p} \sim \nu$ and compute:

$$\begin{aligned} \tilde{x}_{i,p} &= F_p(x_{i,p-1}, u_{i,p}) \\ \tilde{z}_{i,p} &= \nabla F_p(\tilde{x}_{i,p}, u_{i,p}) + F_p'(\tilde{x}_{i,p}, u_{i,p}) z_{i,p-1} \\ \tilde{r}_{i,p} &= r_{i,p-1} + \frac{G_p'(\tilde{x}_{i,p})\tilde{z}_{i,p} + \nabla G_p(\tilde{x}_{i,p})}{G_p(\tilde{x}_{i,p})} \end{aligned}$$

2. Compute the weights $\hat{w}_{i,p} = \frac{G_p(\tilde{x}_{i,p})}{\sum_j G_p(\tilde{x}_{j,p})}$,
3. Define the selection indexes $(k_i)_{1 \leq i \leq N}$ using a consistent selection scheme based on the weights $(\hat{w}_{i,p})_{1 \leq i \leq N}$, and for any $i \in \{1, \ldots, N\}$ set $(x_{i,p}, z_{i,p}, r_{i,p}) = (\tilde{x}_{k_i,p}, \tilde{z}_{k_i,p}, \tilde{r}_{k_i,p})$.

To study more precisely the asymptotic behavior of the estimator given in Theorem 2, one can apply any of the asymptotic results about Interacting Particle Method (see for example [17]). For our purpose, we note that this estimator is unbiased and that the asymptotic variance is linked with the variance of the augmented canonical Markov chain $(X_p, Z_p, R_p)_{p \geq 0}$ (see [17] or [9] for details). In particular, to be efficient, this algorithm requires that the variances of Z_p and R_p do not increase too much as p increases. This is the only possible limitation to the practical efficiency of this algorithm.

The estimator given in Theorem 2, is new in the field of Feynman-Kac models. In [21] the authors use an Infinitesimal Perturbation Analysis (IPA) to compute the gradient of a filtering flow in the context of HMMs. Nonetheless, they do not apply IPA to the Markov chain $(X_p, Z_p, R_p)_{p \geq 0}$ but to the empirical Markov chain $(X_p, x_{1:N,p}, w_{1:N,p})_{p \geq 0}$. Their approach fails to provide an efficient estimator of the Feynman-Kac flow gradient because the variance of their estimator increases as the number of particles N increases (which is not the case for the estimator given in Theorem 2). In fact, in papers [15, 16, 3] the authors develop algorithms which are very close to the recursive implementation of the estimator given via IPA. Let us now explain precisely the relations.

To compute $\nabla \hat{\gamma}_n(f)$ in Equation (11) one could have alternatively used a score method. Under suitable assumptions, one can apply Proposition 3 of Appendix and have:

$$\nabla \hat{\gamma}_n(f) = \mathbb{E}\left[\left(\nabla f(X_n) + f(X_n) \sum_{p=0}^{n} \left(\frac{\nabla m_p(X_{p-1}, X_p)}{m_p(X_{p-1}, X_p)} + \frac{\nabla G_p(X_p)}{G_p(X_p)} \right)\right) \prod_{p=0}^{n} G_p(X_p)\right]. \quad (17)$$

This leads to define a new augmented canonical Markov chain $(X_p, S_p)_{p \geq 0}$ by:

$$\begin{aligned} X_p &= F_p(X_{p-1}, U_p), \text{ where } U_p \sim \nu \\ S_p &= S_{p-1} + \frac{\nabla m_p(X_{p-1}, X_p)}{m_p(X_{p-1}, X_p)} + \frac{\nabla G_p(X_p)}{G_p(X_p)}. \end{aligned}$$

We immediately see that:

$$\frac{\nabla\hat{\gamma}_n(f)}{\hat{\gamma}_n(1)} \quad = \quad \hat{\varsigma}_n(\nabla f + C(f)),$$

where C is a linear operator verifying $C(f)(x,s) = f(x)\,s$, and $\hat{\varsigma}_n$ is the marginal Feynman-Kac measure associated to the potential functions $(G_p)_{p\geq 0}$ and to the augmented canonical Markov chain $(X_p, S_p)_{p\geq 0}$. Plugging together the equation above and Equation (9) leads to the following estimation of the Feynman-Kac flow gradient:

$$\begin{aligned}\nabla[\hat{\eta}_n(f)] &= \frac{\nabla[\hat{\gamma}_n(f)]}{\hat{\gamma}_n(1)} - \hat{\eta}_n(f)\frac{\nabla[\hat{\gamma}_n(1)]}{\hat{\gamma}_n(1)} \\ &= \hat{\varsigma}_n(\nabla f + C(f) - \hat{\eta}_n(f)C(1)).\end{aligned}$$

Using a particle approximation of $\hat{\varsigma}_n$ immediately leads to the following result:

Theorem 3 (Law of Large Number for gradient estimation using Score)**.**

Under the following assumptions, for any $0 \leq p \leq n$:

1. *the transition kernels $M_p(x_{p-1}, dx_p)$ admit a differentiable and bounded density $m_p(x_{p-1}, x_p)$ with respect to a probability measure $\rho_p \in \mathcal{D}(E_p)$,*
2. *the potential functions G_p are differentiable with respect to θ,*
3. *the test function $f : E_n \to \mathbb{R}$ is in $\mathcal{K}^1(\Theta, \rho_n) \cap B_n$,*

we have:

$$\frac{1}{N}\sum_{i=1}^{N}\left(\nabla f(x_{i,n}^N) + f(x_{i,n}^N)\left(s_{i,n}^N - \frac{1}{N}\sum_{j=1}^{N} s_{j,n}^N\right)\right) \xrightarrow{\mathcal{P}} \nabla\hat{\eta}_n(f),$$

where $\Xi_n^N = ((x_{i,n}^N, s_{i,n}^N), w_{i,n}^N)$ is the weighted sample obtained using an IPM on the Feynman-Kac model associated to the potential functions $(G_p)_{p\geq 0}$ and to the augmented canonical Markov chain $(X_p, S_p)_{p\geq 0}$.

The estimator given in Theorem 3 admits a recursive implementation with computational complexity linear in the number of particles N. A pseudo-code for its computation is defined at any stage p by:

1. Sample independently $u_{i,p} \sim \nu$ and compute:

$$\begin{aligned}\tilde{x}_{i,p} &= F_p(x_{i,p-1}, u_{i,p}) \\ \tilde{s}_{i,p} &= s_{i,p-1} + \frac{\nabla m_p(x_{i,p-1}, \tilde{x}_{i,p})}{m_p(x_{i,p-1}, \tilde{x}_{i,p})} + \frac{\nabla G_p(\tilde{x}_{i,p})}{G_p(\tilde{x}_{i,p})}\end{aligned}$$

2. Compute the weights $\hat{w}_{i,p} = \frac{G_p(\tilde{x}_{i,p})}{\sum_j G_p(\tilde{x}_{j,p})}$,
3. Define the selection indexes $(k_i)_{1\leq i\leq N}$ using a consistent selection scheme based on the weights $(\hat{w}_{i,p})_{1\leq i\leq N}$, and for any $i \in \{1, \ldots, N\}$ set $(x_{i,p}, s_{i,p}) = (\tilde{x}_{k_i,p}, \tilde{s}_{k_i,p})$.

This last Algorithm is nearly the same as the one presented in [15, 16, 3]. The weak point of this approach is that it uses a score method to estimate the sensitivity of the Markov chain $(X_p)_{p\geq 0}$ with respect to the parameter θ, and score methods are known to have more variance than IPA methods. A second remark is that our presentation of this algorithm differs from the presentation given in articles [15, 16, 3]. In particular, the authors of [15, 16, 3] wonder how analyze their algorithm (for example how to obtain a Central Limit Theorem), whereas it is straightforward if one follows the presentation of this article.

3 Numerical experiments and applications

3.1 Remarks about practical implementations and possible improvements of the proposed algorithms

In this subsection we discuss some variants and extensions of the Algorithm 2. In a first subsection we explain two variance reduction techniques. Then we explain how to compute differentials of Feynman-Kac flow gradients of orders higher than one. In particular we explicitly show how to adapt the Algorithm 2 to evaluate the Hessian matrix of a Feynman-Kac flow. In a third subsection, we explain how to use the Algorithm 2 if one cannot compute analytically the gradient of functions of the model.

3.1.1 Variance reduction methods

Variance reduction techniques for Interacting Particle Methods have been intensively studied during the last years. See for example the Rao-Blackwellised particle filters [11] or the use of auto adaptive minimal variance instrumental kernels [4]. These methods can be used to accelerate the convergence of the algorithms described in this paper, but their exposition is out the scope of this contribution. In the following, we explain how to use an instrumental kernel and give a slight improvement (in term of variance) of the estimator given in Theorem 2.

To reduce the variance of an estimator obtained using an Interacting Particle Method, one can sample using a family of instrumental transition kernels $(Q_p)_{p\geq 1}$ instead of sampling from the family of canonical transition kernels $(M_p)_{p\geq 1}$:

$$x_{i,p+1}^N \overset{i.i.d.}{\sim} Q_{p+1}(x_{i,p}^N, .). \tag{18}$$

Indeed, if the instrumental transition kernels are absolutely continuous with respect to the canonical transition kernels:

$$\forall p \geq 0,\ \forall x_p \in E_p, Q_{p+1}(x_p, .) \ll M_{p+1}(x_p, .),$$

then the Radon-Nikodym derivative $\frac{dM_p}{dQ_p}$ of $(M_p)_{p\geq 0}$ with respect to $(Q_p)_{p\geq 0}$ is well defined. Under this assumption, the estimator of Theorem 2 and the procedure explained in Section 2 are still valid if one uses

the prediction rule given in Equation (18) and the following updating rule:

$$\hat{w}_{i,p} = \frac{\frac{dM_p}{dQ_p}(x_{i,p-1}, \tilde{x}_{i,p})G_p(\tilde{x}_{i,p})}{\sum_j \frac{dM_p}{dQ_p}(x_{j,p-1}, \tilde{x}_{j,p})G_p(\tilde{x}_{j,p})}.$$

A second way of reducing the variance could be to use a particle approximation available before the selection step. Indeed, the estimator of $\nabla\hat{\eta}_n(f)$ given in Theorem 2 is obtained by evaluating the right hand side of Equation (16) after the selection step. Nonetheless one can also estimate expectancies by a weighed particle approximation, which means before the selection step. As the selection step is unbiased, using selected particles adds a variance term to the estimator. Consequently, a more accurate estimator of $\hat{\eta}_n(f)$ could be provided by its weighed particle approximation. Therefore, Theorem 1 can be rewritten as

$$\forall f \in B_n, \quad \sum_i \hat{w}^N_{i,n} f(\hat{\xi}^N_{i,n}) \xrightarrow{\mathbb{P}} \hat{\eta}_n(f),$$

with the same $(B_p)_{p\geq 0}$ spaces, and Theorem 2 can also be written with a weighed particle approximation:

Theorem 4 (Law of Large Number for gradient estimation using IPA)**.** *Under Assumption 1 and for any $f \in B_n$, we have the following asymptotic estimator for the gradient of the Feynman-Kac flow $\hat{\eta}_n(f)$:*

$$\sum_{i=1}^N \hat{w}^N_{i,n}\left(\nabla f(\hat{x}^N_{i,n}) + f'(\hat{x}^N_{i,n})\hat{z}^N_{i,n} + f(\hat{x}^N_{i,n})\left(\hat{r}^N_{i,n} - \sum_j \hat{w}^N_{j,n}\hat{r}^N_{j,n}\right)\right) \xrightarrow{\mathbb{P}} \nabla\hat{\eta}_n(f),$$

where $\hat{\Xi}^N_n = ((\hat{x}^N_{i,n}, \hat{z}^N_{i,n}, \hat{r}^N_{i,n}), \hat{w}^N_{i,n})$ is the weighted sample obtained using an IPM on the Feynman-Kac model associated to the potential functions $(G_p)_{p\geq 0}$ and to the augmented canonical Markov chain $(X_p, Z_p, R_p)_{p\geq 0}$.

Proof. Starting from Equation (16) and using two times the weighted version of Theorem 1 (see above) gives:

$$\begin{aligned}
\nabla[\hat{\eta}_n(f)] &= \hat{\zeta}_n(\nabla f + A(f) + B(f) - \hat{\eta}_n(f)B(1)) \\
&\xleftarrow{\mathbb{P}} \sum_{i=1}^N \hat{w}^N_{i,n}\nabla f(\hat{x}^N_{i,n}) + \sum_{i=1}^N \hat{w}^N_{i,n} f'(\hat{x}^N_{i,n})\hat{z}^N_{i,n} \\
&\quad + \sum_{i=1}^N \hat{w}^N_{i,n} f(\hat{x}^N_{i,n})\hat{r}^N_{i,n} - \left(\sum_{i=1}^N \hat{w}^N_{i,n} f(\hat{x}^N_{i,n})\right)\left(\sum_j \hat{w}^N_{j,n}\hat{r}^N_{j,n}\right) \\
&= \sum_{i=1}^N \hat{w}^N_{i,n}\left(\nabla f(\hat{x}^N_{i,n}) + f'(\hat{x}^N_{i,n})\hat{z}^N_{i,n} + f(\hat{x}^N_{i,n})\left(\hat{r}^N_{i,n} - \sum_j \hat{w}^N_{j,n}\hat{r}^N_{j,n}\right)\right)
\end{aligned}$$

□

The IPA Algorithm is not deeply modified. One has to use predicted particles and weights instead of using selected particles and all the same weights equal to $\frac{1}{N}$.

3.1.2 Estimation of higher order differentials. A case study with the Hessian matrix.

The method developed in Section 2 allows to estimate $\nabla\hat{\eta}_n(f)$. One should be interested in estimating higher order differentials such as the Hessian matrix $H[\hat{\eta}_n(f)]$ of the Feynman-Kac flow. This is an interesting issue in optimization context, because it allows the use of second order optimization methods, as the Newton-Raphson algorithm, that achieve a quadratic rate of convergence to a locally optimal solution. The Algorithm 2 can be adapted to compute differential of $\hat{\eta}_n(f)$ of any order. As an illustration, we explain in the following how to modify the Algorithm 2 in order to approximate the Hessian matrix $H[\hat{\eta}_n(f)]$.

The Hessian matrix $H[\hat{\eta}_n(f)]$ is a $n_\theta \times n_\theta$ matrix whose element at the position i, j is $\frac{\partial^2 \hat{\eta}_n(f)}{\partial_{\theta_i}\partial_{\theta_j}}$. To simplify the explanation, we suppose in the following that the test function f doesn't depend on the parameter θ, $n_\theta = 1$ and for any $p \geq 0, n_{E_p} = 1$. We also suppose that all the quantities of interest are sufficiently smooth. Under these assumptions, we have:

$$\begin{aligned}
H[\hat{\eta}_n(f)] &= \nabla\left[\nabla\hat{\eta}_n(f)\right] \\
&= \nabla\left[\frac{\nabla[\hat{\gamma}_n(f)]}{\hat{\gamma}_n(1)} - \hat{\eta}_n(f)\frac{\nabla[\hat{\gamma}_n(1)]}{\hat{\gamma}_n(1)}\right] \\
&= \frac{H[\hat{\gamma}_n(f)]}{\hat{\gamma}_n(1)} - \frac{\nabla\hat{\gamma}_n(f)}{\hat{\gamma}_n(1)}\frac{\nabla\hat{\gamma}_n(1)}{\hat{\gamma}_n(1)} - \nabla\hat{\eta}_n(f)\frac{\nabla\hat{\gamma}_n(1)}{\hat{\gamma}_n(1)} \\
&\quad -\hat{\eta}_n(f)\left(\frac{H[\hat{\gamma}_n(1)]}{\hat{\gamma}_n(1)} - \left(\frac{\nabla\hat{\gamma}_n(1)}{\hat{\gamma}_n(1)}\right)^2\right) \\
&= \frac{H[\hat{\gamma}_n(f)]}{\hat{\gamma}_n(1)} - 2\frac{\nabla\hat{\gamma}_n(f)}{\hat{\gamma}_n(1)}\frac{\nabla\hat{\gamma}_n(1)}{\hat{\gamma}_n(1)} \\
&\quad -\hat{\eta}_n(f)\left(\frac{H[\hat{\gamma}_n(1)]}{\hat{\gamma}_n(1)} - 2\left(\frac{\nabla\hat{\gamma}_n(1)}{\hat{\gamma}_n(1)}\right)^2\right) \qquad (19)
\end{aligned}$$

To evaluate the right hand side of Equation (19), we have to evaluate $H[\hat{\gamma}_n(f)]$. For this purpose, we can apply two times Proposition 1:

$$\begin{aligned}
H[\hat{\gamma}_n(f)] &= \nabla\left[\nabla\mathbb{E}\left[f(X_n)\prod_{p=0}^{n} G_p(X_p)\right]\right] \\
&= \nabla\mathbb{E}\left[(f'(X_n)\nabla X_n + f(X_n)R_n)\prod_{p=0}^{n} G_p(X_p)\right] \\
&= \mathbb{E}\left[\left\{\nabla(f'(X_n)\nabla X_n + f(X_n)R_n) + (f'(X_n)\nabla X_n + f(X_n)R_n)R_n\right\}\prod_{p=0}^{n} G_p(X_p)\right] \\
&= \mathbb{E}\left[\left\{f''(X_n)\nabla X_n^2 + f'(X_n)(H[X_n] + 2\nabla X_n R_n) + f(X_n)(\nabla R_n + R_n^2)\right\}\prod_{p=0}^{n} G_p(X_p)\right]
\end{aligned}$$

The above relation can be seen as an expectation along the Markov chain $(X_p, Z_p, H_p, R_p, T_p)_{p\geq 0}$ where the recurrence relations verified by X_p, Z_p and R_p have already been explained in Section 2 and the recurrence relations verified by H_p is given by:

$$\begin{aligned} H_p &= H[X_p] \\ &= \nabla[\nabla X_p] \\ &= \nabla[\nabla F_p(X_{p-1}, U_p) + F_p'(X_{p-1}, U_p) Z_{p-1}] \\ &= H[F_p](X_{p-1}, U_p) + 2\nabla F_p'(X_{p-1}, U_p) Z_{p-1} \\ &\quad + F_p'(X_{p-1}, U_p) H_{p-1} + F_p''(X_{p-1}, U_p) Z_{p-1}^2 \end{aligned} \tag{20}$$

and by T_p verifies:

$$\begin{aligned} T_p &= \nabla R_p \\ &= \nabla \left[R_{p-1} + \frac{G_p'(X_p)}{G_p(X_p)} Z_p + \frac{\nabla G_p(X_p)}{G_p(X_p)} \right] \\ &= T_{p-1} + \nabla \left[\frac{G_p'(X_p)}{G_p(X_p)} \right] Z_p + \frac{G_p'(X_p)}{G_p(X_p)} H_p] + \nabla \left[\frac{\nabla G_p(X_p)}{G_p(X_p)} \right] \\ &= T_{p-1} + \frac{\nabla[G_p'(X_p)]}{G_p(X_p)} Z_p - \frac{G_p'(X_p) \nabla[G_p(X_p)]}{G_p(X_p)^2} Z_p + \frac{G_p'(X_p)}{G_p(X_p)} H_p \\ &\quad + \frac{\nabla[\nabla G_p(X_p)]}{G_p(X_p)} - \frac{\nabla G_n(X_p) \nabla[G_p(X_p)]}{G_p(X_p)^2} \\ &= T_{p-1} + \frac{\nabla G_p'(X_p) + G_p''(X_n) Z_p}{G_p(X_p)} Z_p - \frac{G_p'(X_p) \nabla G_p(X_p) + G_p'(X_p)^2 Z_p}{G_p(X_p)^2} Z_p \\ &\quad + \frac{G_p'(X_p)}{G_p(X_p)} H_p + \frac{H[G_p](X_p) + (\nabla G_p)'(X_p) Z_p}{G_p(X_p)} \\ &\quad - \frac{\nabla G_p(X_p)^2 + \nabla G_p(X_p) G_p'(X_p) Z_p}{G_p(X_p)^2} \end{aligned} \tag{21}$$

Using an Interacting Particle Method on the Feynman-Kac model associated to the Markov chain $(X_p, Z_p, H_p, R_p, T_p)_{p\geq 0}$ and to the potential functions $(G_p)_{p\geq 0}$ allows to numerically estimate the right hand side of Equation (19). A sketch of the pseudo code is:

1. Sample independently $u_{i,p} \sim \nu$ and compute $\tilde{\xi}_{1:N,p} = (\tilde{x}_{1:N,p}, \tilde{z}_{1:N,p}, \tilde{h}_{1:N,p}, \tilde{r}_{1:N,p}, \tilde{t}_{1:N,p})$ using the recurrence relations (12),(13),(14),(20) and (21),
2. Compute the weights $\hat{w}_{1:N,p} = \frac{G_p(\tilde{x}_{1:N,p})}{\sum_j G_p(\tilde{x}_{j,p})}$,
3. Define the selection indexes $(k_i)_{1\leq i\leq N}$ using a consistent selection scheme based on the weights $(\hat{w}_{i,p})_{1\leq i\leq N}$, and for any $i \in \{1, \ldots, N\}$ sets $\xi_{1:N,p} = \tilde{\xi}_{k_{1:N},p}$.

3.1.3 What to do if one cannot compute analytically gradients of functions needed for algorithms?

If transition functions are smooth (they verify Assumption 1) but one cannot (or doesn't want to) compute analytically the expression of their

differentials, it is possible to use finite difference estimation. For example this is the case when transition functions are the result of a full computer program. This situation is quite frequent in industrial applications.

In such a situation the Algorithm 2 still works, but we need to evaluate ∇F_p or F_p' by a finite difference method:

$$\nabla F_p \approx \left(\frac{F_{\theta+h_i,p} - F_{\theta-h_i,p}}{2h} \right)_{1 \leq i \leq n_\Theta}$$

where h_i is a vector for which all coordinates are null except the i-th which is equal to $0 < h \ll 1$. Note that this evaluation requires the computation of $2n_\Theta$ transition evaluations and this could be too computationally expensive in situations where $n_\Theta \gg 1$ and where sampling a transition is time consuming.

3.2 Application to gradient evaluation and parameter identification in HMM

In this section we briefly recall the notations and definitions used in Section 1.1 of this paper. We show how to apply the estimators given in Theorem 2 to the problem of likelihood evaluation and maximization in HMMs.

A HMM is built from a state process and an observation process. The state process $(X_p)_{p \geq 0}$ is an homogeneous Markov chain with initial probability measure $\mu(dx_0)$ and Markov transition kernel $K(dx_{p+1}|x_p)$. The observation process $(Y_p)_{p \geq 1}$ is linked with the state process by the conditional probability measure $\mathbb{P}(Y_p \in dy_p | X_p = x_p) = g(y_p, x_p)\lambda(dy_p)$ and is conditionally independent given the state process: $i \neq j \Rightarrow \mathbb{P}(Y_i \in dy_i, Y_j \in dy_j | X_i = x_i, X_j = x_j) = \mathbb{P}(Y_i \in dy_i | X_i = x_i)\mathbb{P}(Y_j \in dy_j | X_j = x_j)$. Moreover, we consider HMMs for which the state and the observation processes are parameterized by $\theta \in \Theta$, where $\Theta \subset \mathbb{R}^{n_\theta}$ is an open subset of $\mathbb{R}^{n_\theta}$.

We present two classical examples of parameterized HMMs. They are frequently used to compare performances of algorithms dealing with identification and filtering in HMMs.

Example 3.1 (Autoregressive model)**.** The autoregressive model of order 1 (AR_1) is the simplest representation of the class of linear-Gaussian HMMs:

$$\begin{aligned} X_p &= \phi X_{p-1} + \sigma U_p, \quad X_0 \sim \mathcal{N}\left(0, \frac{\sigma^2}{1-\phi^2}\right) \\ Y_p &= \rho X_p + \beta V_p, \end{aligned} \tag{22}$$

where $U_p \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$ and $V_p \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$ are two mutually independent and independent of the initial state X_0 sequences of random variables. In this example, $\theta = (\phi, \sigma, \rho, \beta)$ is a four-dimensional parameter and $\Theta = (\mathbb{R}_+)^4$. For an AR model the filtering density (and therefore the likelihood

function and its gradient) can be computed exactly using a Kalman filter. This model will allow us to measure the bias of algorithms given in Section 2, by comparing them with the true estimation.

Example 3.2 (Stochastic Volatility)**.** Hull and al. generalized the Black-Scholes option pricing formula to allow for stochastic volatility. Their formula has emerged as the dominant approach. In fact it is a typical nonlinear and nongaussian HMM. The dynamics are:

$$\begin{aligned} X_p &= \phi X_{p-1} + \sigma U_p, \;\; X_0 \sim \mathcal{N}\left(0, \frac{\sigma^2}{1-\phi^2}\right) \\ Y_p &= \beta \exp\left(\frac{X_p}{2}\right) V_p, \end{aligned} \tag{23}$$

where $U_p \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$ and $V_p \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$ are two mutually independent and independent of the initial state X_0 sequences of random variables. It deals with a three-dimensional parameter $\theta = (\phi, \sigma, \beta)$ and $\Theta = \mathbb{R}_+^3$. This model will allow us to assess the capacity of the Algorithm 2 to deal with non linear and non gaussian dynamics.

Let us now suppose we are given a sequence of successive realizations of the observation process $(y_{p,\theta^*})_{p\geq 1}$ that were obtained using an unknown parameter denoted θ^*. In the following, we make a slight abuse of notation by denoting $g_{p,\theta}(x_p) = g_\theta(y_{p,\theta^*}, x_p)$. We recall that the objective of parameter identification in HMMs is to recover θ^* using the sequence of observations $(y_{p,\theta^*})_{p\geq 1}$. We have seen in the introduction that the parameter estimation problem can be turned into maximizing the log-likelihood l_n. In practice, this can be achieved with a stochastic gradient ascent with respect to θ on l_n. We also recall that the gradient of the log-likelihood function may be written as:

$$\nabla l_n(\theta) = \sum_{p=1}^{n} \frac{\nabla[\pi_{p|p-1,\theta}(g_{p,\theta})]}{\pi_{p|p-1,\theta}(g_{p,\theta})}. \tag{24}$$

We have also seen in Section 1.1 that a HMM can be viewed as a particular case of Feynman-Kac model. Indeed, taking for any $p \geq 0$, $M_p = K$ and $G_p = g_p$, implies that $\eta_{p+1} = \pi_{p+1|p}$ is the marginal predicted measure of the Feynman-Kac model associated to $(G_p, M_p)_{p\geq 0}$. Consequently we have:

$$\nabla[\pi_{p|p-1}(g_{p,\theta})] = \nabla\eta_p(G_p). \tag{25}$$

Theorem 2 has been designed to evaluate the gradient of the updated Feynman-Kac flow $\nabla\hat{\eta}_p(f)$ and not the gradient of the predicted Feynman-Kac flow $\nabla\eta_p(f)$. Nonetheless, Theorem 2 can be easily adapted to evaluate the gradient of the predicted Feynman-Kac flow $\nabla\eta_p(f)$. Indeed the algorithm remains the same, but the evaluation is done using the predicted particles system:

$$\frac{1}{N}\sum_{i=1}^{N}\left(\nabla f(\tilde{x}_{i,p}^N) + f'(\tilde{x}_{i,p}^N)\tilde{z}_{i,p}^N + f(\tilde{x}_{i,p}^N)\left(r_{i,p-1}^N - \frac{1}{N}\sum_j r_{j,p-1}^N\right)\right) \xrightarrow{\mathbb{P}} \nabla\eta_n(f). \tag{26}$$

To evaluate the right hand side of Equation (25), we can apply the evaluation provided in Equation (26) taking $f = G_p$, this leads to the following estimator:

$$\frac{1}{N}\sum_{i=1}^{N}\left(\nabla G_p(\tilde{x}_{i,p}^N) + G_p'(\tilde{x}_{i,p}^N)\tilde{z}_{i,p}^N + G_p(\tilde{x}_{i,p}^N)\left(r_{i,p-1}^N - \frac{1}{N}\sum_j r_{j,p-1}^N\right)\right).$$

Consequently, we have:

$$\frac{\sum_{i=1}^N\left(\nabla G_p(\tilde{x}_{i,p}^N) + G_p'(\tilde{x}_{i,p}^N)\tilde{z}_{i,p}^N + G_p(\tilde{x}_{i,n}^N)\left(\tilde{r}_{i,p}^N - \frac{1}{N}\sum_j \tilde{r}_{j,p}^N\right)\right)}{\sum_j G_p(\tilde{x}_{j,p}^N)} \xrightarrow{\mathbb{P}} \frac{\nabla[\pi_{p|p-1,\theta}(g_{p,\theta})]}{\pi_{p|p-1,\theta}(g_{p,\theta})}. \quad (27)$$

The above procedure shows how to estimate the gradient of the log-likelihood ∇l_n. If we want to recover the unknown parameter θ^*, we have seen that it is possible to compute a stochastic gradient ascent using ∇l_n. A pseudo-code of the application of Algorithm 2 to the parameter estimation in HMMs when n observations are available, is given through the repetition for increasing time step k of these two points:

1. estimate $\nabla l_n(\theta_k)$ using Equation 27 and Equation 24
2. update the parameter $\theta_{k+1} = \theta_k + \epsilon_k \nabla l_n(\theta_k)$

where $\sum_k \epsilon_k = \infty$ and $\sum_k \epsilon_k^2 < \infty$. A main advantage of this algorithm is that it can be used on-line.

3.3 Numerical validation on an AR_1 model

In this section we present results of numerical experiments on a gradient estimation problem of the log-likelihood associated to the linear gaussian Hidden Markov Model defined in Example 3.1 with $\theta^* = (\phi^*, \sigma^*, \rho^*, \beta^*) = (0.8, 0.5, 1.0, 1.0)$. The Kalman filter allows to compute the exact value of the filtering distribution, and therefore the exact value of the log-likelihood function $l_n(\theta)$ and its gradient $\nabla l_n(\theta)$ for any values of the parameter of the model θ. We know that for linear gaussian models the Algorithms obtained from the two Theorems 2 and 3 have no interest because the gradient estimation problem is exactly solved by using the Kalman filter. Nonetheless, as the exact value of the log-likelihood gradient $\nabla l_n(\theta)$ is known, it allows us to evaluate the mean quadratic error e_n^N of the approximation of $\nabla l_n(\theta)$ by the value $\widehat{\nabla l_n}^N(\theta)$ computed with Algorithms 2 and 3 using N particles.

During numerical experiments, we evaluate separately the bias and the variance of the gradient estimator $\widehat{\nabla l_n}^N(\theta)$ by computing independently m-times the quantity $\widehat{\nabla l_n}^N(\theta)$ and by using the following bias/variance

decomposition:

$$\begin{aligned} e_n^N(\theta) &= \mathbb{E}\left[\left(\nabla l_n(\theta) - \widehat{\nabla l_n}^N(\theta)\right)^2\right] \\ &= \mathbb{E}\left[\nabla l_n(\theta) - \widehat{\nabla l_n}^N(\theta)\right]^2 + var\left(\widehat{\nabla l_n}^N(\theta)\right) \\ &\approx \left(\nabla l_n(\theta) - \frac{1}{m}\sum_{k=1}^{m} \widehat{\nabla l_{n,k}}^N(\theta)\right)^2 \\ &\quad + \left(\frac{1}{m}\sum_{k=1}^{m} \widehat{\nabla l_{n,k}}^N(\theta)^2 - \left(\frac{1}{m}\sum_{k=1}^{m} \widehat{\nabla l_{n,k}}^N(\theta)\right)^2\right), \quad (28) \end{aligned}$$

where $\mathbb{E}\left[\nabla l_n(\theta) - \widehat{\nabla l_n}^N(\theta)\right]$ is called the bias of the estimator and $var\left(\widehat{\nabla l_n}^N(\theta)\right)$ the variance.

In the following three subsections, we evaluate the performances of Algorithms 2,3 and [19]. The experiments were realized with a 1.6 Ghz PC under the software MatLab. We have not optimized the implementation of the three algorithms, in particular we have not implemented the acceleration method proposed in [10] for the marginal particle filter as used in [19]. The acceleration method adds a numerical approximation through the use of a tree-based numerical method to compute quickly the linear combination of the kernel evaluation. Moreover we study the behavior of Algorithms 2 and 3 along the two variables N and n: when the number of particle N tends to infinity and when the number of times step n tends to infinity.

3.3.1 Comparisons with the state of art method [19]

In this section we present a comparative study of performances of Algorithms 2, 3 and [19] in terms of bias, variance and CPU time. We evaluate the log-likelihood at time step $n = 50$ and at point $\theta = (\phi, \sigma, \rho, \beta) = (0.7, 0.4, 0.9, 0.9)$. For each algorithm, we compute $m = 500$ log-likelihood estimates $(\widehat{\nabla l_{50,k}}^N(\theta))_{1\leq k\leq 500}$, and evaluate empirically the bias and the variance using Equation (28). For Algorithms 2 and 3, we use successively $N = 5\ 10^2$ particles and $N = 1\ 10^4$ particles, and for the Algorithm [19] we only use $N = 5\ 10^2$ particles as it is practically impossible to apply it using more than $N = 10^3$ particles. We also measure the computational time (in second) needed to evaluate one log-likelihood $\widehat{\nabla l_{50,k}}^N(\theta)$. The results are reported in Table 1.

As it is expected, the first remark is that the bias is null. In these numerical simulations it is not exactly null because we evaluate it using a finite number of simulations $m = 500$. But we see that the bias is always inferior to the standard deviation in one or two orders of dimension. A second conclusion is that the all three Algorithms 2, 3 and [19] achieve a rather small quadratic error $e^{0.5} < 10^{-2}$, thus they are likely to be asymptotically consistent. A third conclusion is that the method [19] is very time consuming when one doesn't use any acceleration technique.

		N	IPA	Score	[19]
ϕ	bias	$5\ 10^2$	$4.5\ 10^{-3}$	$2.5\ 10^{-3}$	$9.0\ 10^{-4}$
		10^4	$2.0\ 10^{-4}$	$4.0\ 10^{-4}$	X
	standard deviation	$5\ 10^2$	$4.7\ 10^{-2}$	$6.0\ 10^{-2}$	$5.3\ 10^{-2}$
		10^4	$8.8\ 10^{-3}$	$6.0\ 10^{-2}$	X
	CPU time (s.)	$5\ 10^2$	$1.8\ 10^{-1}$	$1.9\ 10^{-1}$	21
		10^4	2.1	2.3	X
σ	bias	$5\ 10^2$	$1.3\ 10^{-3}$	$4.0\ 10^{-3}$	$1.0\ 10^{-4}$
		10^4	$2.0\ 10^{-4}$	$1.1\ 10^{-3}$	X
	standard deviation	$5\ 10^2$	$2.3\ 10^{-2}$	$6.6\ 10^{-2}$	$2.2\ 10^{-2}$
		10^4	$7.9\ 10^{-3}$	$2.0\ 10^{-2}$	X
	CPU time (s.)	$5\ 10^2$	$1.7\ 10^{-1}$	$1.9\ 10^{-1}$	23
		10^4	2.1	2.3	X
ρ	bias	$5\ 10^2$	$6.0\ 10^{-4}$	$1.2\ 10^{-3}$	$8.0\ 10^{-4}$
		10^4	$3.0\ 10^{-4}$	$1.0\ 10^{-4}$	X
	standard deviation	$5\ 10^2$	$1.5\ 10^{-2}$	$1.5\ 10^{-2}$	$6.9\ 10^{-2}$
		10^4	$6.0\ 10^{-3}$	$5.7\ 10^{-3}$	X
	CPU time (s.)	$5\ 10^2$	$1.9\ 10^{-2}$	$2.1\ 10^{-1}$	19
		10^4	2.2	2.3	X
ρ	bias	$5\ 10^2$	$1.7\ 10^{-4}$	$3.9\ 10^{-3}$	$4.5\ 10^{-3}$
		10^4	$2.0\ 10^{-4}$	$1.0\ 10^{-4}$	X
	standard deviation	$5\ 10^2$	$4.4\ 10^{-2}$	$4.3\ 10^{-2}$	$2.7\ 10^{-2}$
		10^4	$6.2\ 10^{-3}$	$5.7\ 10^{-3}$	X
	CPU time (s.)	$5\ 10^2$	$2.0\ 10^{-1}$	$2.1\ 10^{-1}$	20
		10^4	2.3	2.5	X

Table 1: Results computed at $\phi = 0.7$ whereas $\phi^\star = 0.8$ with 500 simulations, 500 particles and 50 time steps.

Indeed, even for a small number of particles $N = 500$, the CPU time needed to compute a log-likelihood approximation is 10^2-times the CPU time needed for Algorithms 2, 3. The fourth conclusion is that for a fixed number of particles, the method [19] performs slightly better than the two others algorithms. However, as we can notice in experiments using $N = 10^4$ particles, for a fixed CPU time the two others algorithms strongly outperform the Algorithm [19].

3.3.2 Estimation behavior when the number of particles increases

We have seen in Subsection 3.3.1 that the estimators of Algorithms 2 and 3 achieve a rather small quadratic error for $N = 10^4$ particles. Moreover, we have noticed in Section 2 that it is possible to deduce a Central Limit Theorem for these estimators. It is interesting to study experimentally the dependance of the quadratic error e_n^N in the number of particles N. To evaluate the quadratic error of each algorithm, we

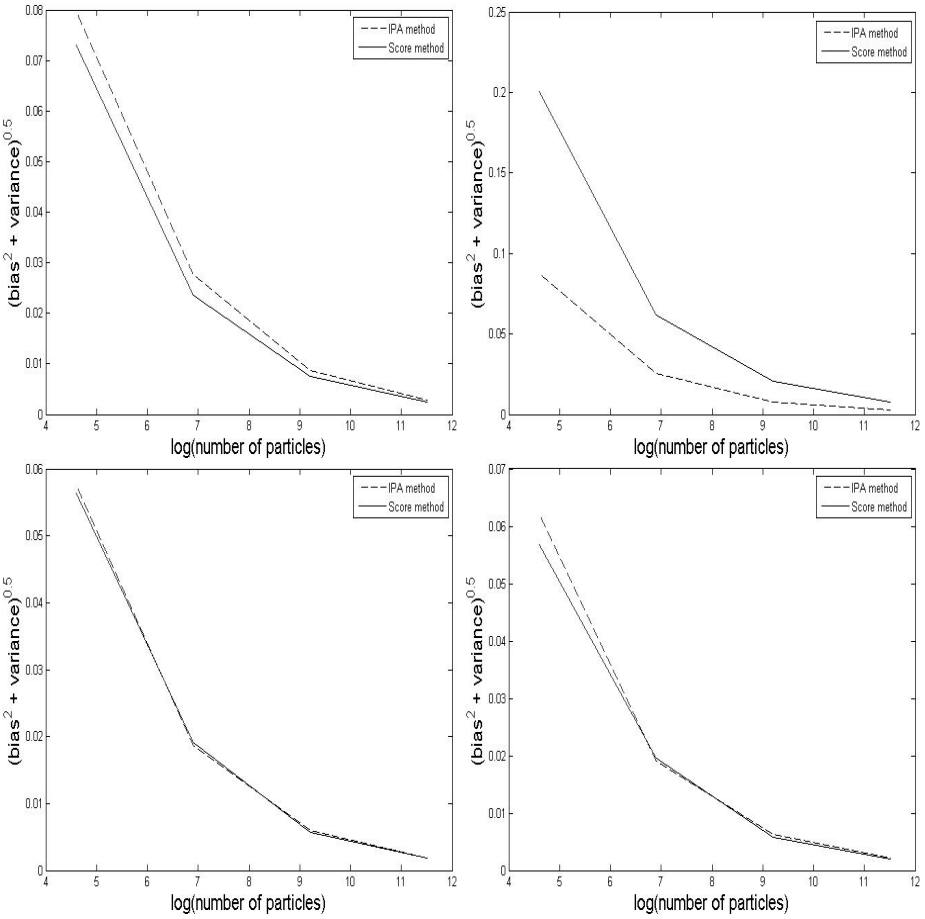

Figure 1: Asymptotic behavior along the number of particles N of Algorithms 2 and 3 by computing the log-likelihood derivative of Example 3.1. We use $10^2, 10^3, 10^4$, and 10^5 particles, and 50 time steps. The values of the true parameters are $(0.8, 0.5, 1.0, 1.0)$, and the log-likelihood derivatives are computed at $(0.7, 0.4, 0.9, 0.9)$. Left up: ϕ. Right up: σ. Left down : ρ. Right down : β.

compute $m = 500$ log-likelihood estimates $(\widehat{\nabla l_{50,k}}^N(\theta))_{1\leq k\leq 500}$ at point $\theta = (\phi, \sigma, \rho, \beta) = (0.7, 0.4, 0.9, 0.9)$ and at time step $n = 50$ for different number of particles $N \in \{10^2, 10^3, 10^4, 10^5\}$. The results are reported in Figure 1.

The main conclusion is that the experimental results strengthen the consistence result given in Section 2: the error tends to zero as the number of particle tends to infinity. A second remark is that the two methods have the same performances for coefficients in the observable equation. In fact, this is obvious because the algorithms are the same for coefficients

belonging only to the potentials functions $(G_p)_{p\geq 0}$. A third remark is that the IPA estimator (Theorem 2) performs slightly better than the Score estimator (Theorem 3) for coefficients in the state equation (in F-K context, it means coefficients belonging to the transition kernels $(M_p)_{p\geq 0}$). In this example it is particulary true for the coefficient σ which is the noise of the state equation.

3.3.3 Estimation behavior when the number of time steps increases

		IPA			Score		
	n	10	10^2	10^3	10	10^2	10^3
ϕ	bias	$2.0\ 10^{-4}$	$1.7\ 10^{-3}$	$6.0\ 10^{-4}$	$3.0\ 10^{-4}$	$4.0\ 10^{-4}$	$2.3\ 10^{-3}$
	standard deviation	$1.8\ 10^{-2}$	$3.3\ 10^{-2}$	$2.7\ 10^{-2}$	$2.7\ 10^{-2}$	$2.6\ 10^{-2}$	$2.5\ 10^{-2}$
	CPU time (s)	$5.0\ 10^{-2}$	$5.0\ 10^{-1}$	5.2	$5.0\ 10^{-2}$	$5.5\ 10^{-1}$	5.3
σ	bias	$9.0\ 10^{-4}$	$2.5\ 10^{-3}$	$1.3\ 10^{-3}$	$4.7\ 10^{-3}$	$7.0\ 10^{-4}$	$1.2\ 10^{-3}$
	standard deviation	$2.3\ 10^{-2}$	$2.2\ 10^{-2}$	$2.2\ 10^{-2}$	$6.5\ 10^{-2}$	$6.1\ 10^{-2}$	$5.6\ 10^{-2}$
	CPU time (s)	$4.0\ 10^{-2}$	$5.0\ 10^{-1}$	5.2	$6.0\ 10^{-2}$	$5.3\ 10^{-1}$	5.6
ρ	bias	$7.0\ 10^{-4}$	$3.0\ 10^{-4}$	$4.0\ 10^{-4}$	$8.0\ 10^{-4}$	$6.0\ 10^{-4}$	$7.0\ 10^{-4}$
	standard deviation	$1.5\ 10^{-2}$	$9.7\ 10^{-3}$	$1.2\ 10^{-2}$	$1.6\ 10^{-2}$	$8.9\ 10^{-3}$	$1.3\ 10^{-2}$
	CPU time (s)	$5.0\ 10^{-2}$	$5.3\ 10^{-1}$	5.2	$5.0\ 10^{-2}$	$5.8\ 10^{-1}$	5.5
β	bias	$7.0\ 10^{-4}$	$1.9\ 10^{-3}$	$2.6\ 10^{-3}$	$8.4\ 10^{-6}$	$2.3\ 10^{-3}$	$8.0\ 10^{-4}$
	standard deviation	$1.7\ 10^{-2}$	$1.9\ 10^{-2}$	$2.3\ 10^{-2}$	$1.8\ 10^{-2}$	$1.9\ 10^{-2}$	$2.0\ 10^{-2}$
	CPU time (s)	$6.0\ 10^{-2}$	$5.3\ 10^{-1}$	5.7	$5.0\ 10^{-2}$	$5.8\ 10^{-1}$	5.9

Table 2: Asymptotic behavior along the number of time steps n of Algorithms 2 and 3 by computing the log-likelihood derivative of Example 3.1. We use $N = 10^3$ particles. The values of the true parameters are $(0.8, 0.5, 1.0, 1.0)$, and the log-likelihood derivatives are computed at $(0.7, 0.4, 0.9, 0.9)$.

One of the main issue in Feynman-Kac flow estimation is the stability of the numerical approximation when the number of time steps n increases. In the HMM framework this corresponds to an increasing number of observations $y_{1:n}$. To assess the stability of the estimator of Algorithms 2 and 3, we evaluate the quadratic error for each algorithm by computing $m = 500$ log-likelihood estimates $(\widehat{\nabla l_{50,k}}^N(\theta))_{1\leq k\leq 500}$ at point $\theta = (\phi, \sigma, \rho, \beta) = (0.7, 0.4, 0.9, 0.9)$ with $N = 10^3$ particles and for different values of the number of time steps $n \in \{10, 10^2, 10^3\}$. The results are presented in Table 2.

The conclusion of this last study is that the quadratic error e_n^N is independent of the number of time steps n, which experimentally confirms the stability of the two Algorithms 2 and 3.

3.4 Numerical experiments on a stochastic volatility model

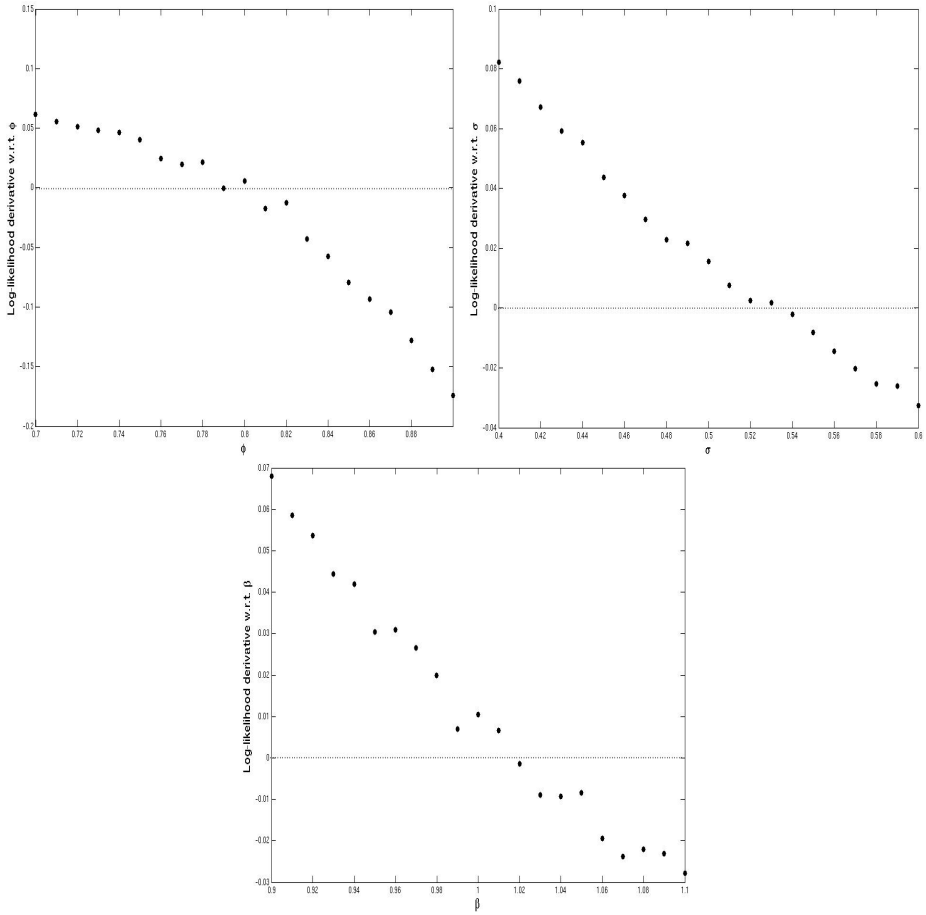

Figure 2: Graphical representation, for some values of θ, of the evaluation of the log-likelihood gradient function $\widehat{\nabla l_n}^N(\theta)$ associated to the stochastic volatility model presented in Example 3.2 at time step $n = 5\ 10^3$ and using the estimator given in Algorithm 2. The true parameter is $\theta^* = (\phi^\star, \sigma^\star, \beta^\star) = (0.8, 0.5, 1)$ and we use $N = 5\ 10^4$ particles. Left up: ϕ. Right up: σ. Down : β.

In this subsection, we are interested in evaluating the log-likelihood function for the stochastic volatility model presented in Example 3.2. As this model is neither linear nor gaussian, the log-likelihood function cannot be evaluated using a Kalman filter. We have seen that parameter identification can be reduced to searching the zero of the log-likelihood gradient function. Consequently, we have plotted the value of one evaluation of $\widehat{\nabla l_n}^N(\theta)$ for $n = 5\ 10^3$, for different values of parameter θ, with

$\theta^* = (\phi^\star, \sigma^\star, \beta^\star) = (0.8, 0.5, 1)$, and using $N = 5\ 10^4$ particles. Results are represented in Figure 2.

These plots show that the estimation is a little bit noisy but the estimated derivatives are equal to zero at a point belonging to a neighborhood of the true parameter, which confirms that the derivative of log-likelihood function is well evaluated.

4 Appendix

Let $(\Omega, \mathcal{F}, \mathcal{P})$ be a probability space. For any $k \in \mathbb{N}^*$, $\mathbb{R}^k$ is considered as a topological vector space with norm $\| \ \|$ and the topological structure $(\mathbb{R}^k, \mathcal{O}(\mathbb{R}^k))$ induced by its norm. In the following $\{(\omega, \theta) \mapsto X_\theta(\omega)\}$ is a family of $\mathbb{R}^m$-valued random variables indexed by a parameter $\theta \in \Theta$, where $\Theta \in \mathcal{O}(\mathbb{R}^{n_\theta})$ is an open subset of $\mathbb{R}^{n_\theta}$. For any $\theta \in \Theta$, one denotes by $\mathcal{V}_\Theta(\theta) = \{A \in \mathcal{O}(\Theta) | \theta \in A\}$ the set of all neighborhoods of θ in Θ.

We adopt the following notation:

Notation 1. $\mathcal{K}^q(\Theta, \mathcal{P})$ denotes the set of all functions $\{(\omega, \theta) \mapsto X_\theta(\omega)\}$ such that for any $\theta \in \Theta$:

1. the function $\{\omega \mapsto X_\theta(\omega)\}$ is measurable,
2. for almost all $\omega \in \Omega$ the function $\{\theta \mapsto X_\theta(\omega)\}$ is differentiable at θ. We write ∇X_θ its gradient,
3. $\exists \mathcal{V} \in \mathcal{V}_\Theta(0), \exists C_\theta \in L^q(\mathcal{P}), \forall h \in \mathcal{V}, \|X_{\theta+h} - X_\theta\| \overset{a.s.}{\leq} \|h\| C_\theta$.

Proposition 1 (Dominated convergence for gradient)**.**
If $X \in \mathcal{K}^1(\Theta, \mathcal{P})$ then for any $\theta \in \Theta$, the function $\{\omega \mapsto \nabla_\theta X(\omega)\}$ is in $L^1(\mathcal{P})$, the function $\{\theta \mapsto \mathbb{E}[X_\theta]\}$ is differentiable and verifies $\nabla \mathbb{E}[X_\theta] = \mathbb{E}[\nabla X_\theta]$.

Proof. Given a sequence h_n of vectors, we set $F_{\theta,n}(\omega) = \frac{X_{\theta+h_n}(\omega) - X_\theta(\omega)}{\|h_n\|}$. For any sequence $h_n \to 0$, we have:

$$\begin{aligned} \nabla \mathbb{E}[X_\theta] &= \lim_n \frac{\mathbb{E}[X_{\theta+h_n}] - \mathbb{E}[X_\theta]}{\|h_n\|} \\ &= \lim_n \mathbb{E}\left[\frac{X_{\theta+h_n} - X_\theta}{\|h_n\|}\right] \\ &= \lim_n \mathbb{E}[F_{\theta,n}]. \end{aligned}$$

For almost all $\omega \in \Omega$, $\|F_{\theta,n}(\omega)\| \leq g_\theta(\omega)$, using the dominated convergence theorem we have $\lim_n \mathbb{E}[F_{\theta,n}] = \mathbb{E}[\lim_n F_{\theta,n}]$. And, by differentiability of the function $\{\theta \mapsto X_\theta(\omega)\}$, we have $\lim_n F_{\theta,n} \overset{a.s.}{=} \nabla X_\theta$. □

Example 4.1. Take $\Omega = [0,1], \mathcal{P} = \mathbb{U}([0,1]), \Theta =]0,1[$ and $X_\theta(\omega) = \mathbf{1}_{[\theta,1]}$. For any $\theta \in \Theta$ and any $\omega \neq \theta$, the function $\{\omega \mapsto X_\theta(\omega)\}$ is differentiable with value $\nabla X_\theta(\omega) = 0$. Note that $X_{\theta+h}(\omega) - X_\theta(\omega) = \mathbf{1}_{[\theta+h,1]} - \mathbf{1}_{[\theta,1]} = -\mathbf{1}_{[\theta,\theta+h[}$ implies $\sup_h \frac{|X_{\theta+h}(\omega) - X_\theta(\omega)|}{|h|} = \frac{1}{|\omega - \theta|}$, consequently X **is not** in $\mathcal{K}^1(\Theta, \mathcal{P})$.

Example 4.2. Using the same spaces, now take $X_\theta(\omega) = \frac{\omega-\theta}{b}\mathbf{1}_{[\theta,\theta+b[}(\omega) + \mathbf{1}_{[\theta+b,1[}(\omega)$. X is in $\mathcal{K}^1(\Theta,\mathcal{P})$ because $|X_{\theta+h}(\omega) - X_\theta(\omega)| \leq \frac{|h|}{b}$.

Proposition 2 (Stability of $\mathcal{K}^q(\Theta,\mathcal{P})$ spaces)**.**
The following assertions are true:

1. *For any $q \in \mathbb{N}\cup\{\infty\}$, a differential and locally contracting function is in $\mathcal{K}^q(\Theta,\mathcal{P})$. In particular $C^1(\Theta) \subset \mathcal{K}^q(\Theta,\mathcal{P})$.*
2. *For any $q \in \mathbb{N}\cup\{\infty\}$, $\mathcal{K}^q(\Theta,\mathcal{P})$ is a vectorial space.*
3. *For any $q \in \mathbb{N}\cup\{\infty\}$, if $X \in \mathcal{K}^q(\Theta,\mathcal{P})$ and $(Y_\theta)_{\theta\in\Theta}$ is locally bounded:*

$$\forall\theta\in\Theta, \exists B_\theta > 0, \exists\mathcal{V}\in\mathcal{V}\Theta(0), \forall h\in\mathcal{V}, \|Y_{\theta+h}\|_\infty < B_\theta,$$

then $XY \in \mathcal{K}^q(\Theta,\mathcal{P})$.

Proof. 1. Given a family of function $(X_\theta)_{\theta\in\Theta}$ differentiable with respect to θ and locally contracting with local contraction constant C_θ, for any $q > 0$ we have $C_\theta(\omega) = C_\theta \in L^q(\mathcal{P})$.
As any continuously differentiable function taking its values in a locally compact vectorial space is locally contracting, the conclusion follows from the remark above.

2. For any $q > 0$, take $\lambda \in \mathbb{R}, X \in \mathcal{K}^q(\Theta,\mathcal{P})$ and $Y \in \mathcal{K}^q(\Theta,\mathcal{P})$, for almost all $\omega\in\Omega$ we have:

$$\begin{aligned} & \|\lambda X_{\theta+h}(\omega) + Y_{\theta+h}(\omega) - \lambda X_\theta(\omega) - Y_\theta(\omega)\| \\ \leq\ & |\lambda|\|X_{\theta+h}(\omega) - X_\theta(\omega)\| + \|Y_{\theta+h}(\omega) - Y_\theta(\omega)\| \\ \leq\ & (|\lambda|C_\theta^X(\omega) + C_\theta^Y(\omega))\|h\|. \end{aligned}$$

As $L^q(\mathcal{P})$ is a vectorial space $(|\lambda|C_\theta^X(\omega) + C_\theta^Y(\omega)) \in L^q(\mathcal{P})$.

3. For almost all $\omega\in\Omega$ we have:

$$\begin{aligned} & \|X_{\theta+h}(\omega)Y_{\theta+h}(\omega) - X_\theta(\omega)Y_\theta(\omega)\| \\ \leq\ & \|X_{\theta+h}(\omega) - X_\theta(\omega)\|B_\theta \\ \leq\ & \|h\|B_\theta^Y C_\theta^X(\omega) \end{aligned}$$

□

Proposition 3 (Score Method applied to a Markov chain)**.**
Let $n \in \mathbb{N}$ be an integer, $(X_p)_{p\geq 0}$ be a E_p-valued Markov chain and f a function $f : \times_{p=0}^n E_p \to \mathbb{R}$. If for any $0 \leq p \leq n$, the transition kernels $K_p(x_{p-1}, dx_p) = k(x_{p-1}, x_p)\nu_p(dx_p)$ admit a differentiable and bounded density relative to a probability measure $\nu_p \in D(E_p)$, and $f \in \mathcal{K}^1(\Theta, \wedge_{p=0}^n \nu_p)$, then:

$$\nabla\mathbb{E}[f(X_n,\ldots,X_0)] = \mathbb{E}\left[f(X_n,\ldots,X_0)\sum_{p=0}^n \frac{\nabla k_p}{k_p}(X_{p-1},X_p)\right].$$

Proof. First we note that $\mathbb{E}[f(X_n,\ldots,X_0)] = \int f(x_n,\ldots,x_0)\prod_{p=0}^{n} K_p(x_{p-1},dx_p)$. As $K_p(x_{p-1},dx_p) = k(x_{p-1},x_p)\nu_p(dx_p)$, we have $\int f(x_n,\ldots,x_0)\prod_{p=0}^{n} K_p(x_{p-1},dx_p) = \int f(x_n,\ldots,x_0)\prod_{p=0}^{n} k_p(x_{p-1},x_p)\nu_p(dx_p)$. The functions $k_p(x_{p-1},x_p)$ are bounded and differentiable, and $f \in \mathcal{K}^1(\Theta,\wedge_{p=0}^{n}\nu_p)$, then thanks to the result 3 of Proposition 2, we have: $f\prod_{p=0}^{n} k_p \in \mathcal{K}^1(\Theta,\wedge_{p=0}^{n}\nu_p)$. Consequently, using Proposition 1, we have:

$$\begin{aligned}
& \nabla \int f(x_n,\ldots,x_0)\prod_{p=0}^{n} k_p(x_{p-1},x_p)\nu_p(dx_p) \\
= & \int f(x_n,\ldots,x_0)\nabla[\prod_{p=0}^{n} k_p(x_{p-1},x_p)]\prod_{p=0}^{n}\nu_p(dx_p) \\
= & \int f(x_n,\ldots,x_0)\sum_{p=0}^{n}\frac{\nabla k_p(x_{p-1},x_p)}{k_p(x_{p-1},x_p)}\prod_{p=0}^{n} k_p(x_{p-1},x_p)\nu_p(dx_p) \\
= & \mathbb{E}\left[f(X_n,\ldots,X_0)\sum_{p=0}^{n}\frac{\nabla k_p}{k_p}(X_{p-1},X_p)\right]
\end{aligned}$$

Remarking $(k(x_{p-1},x_p)=0) \Rightarrow (\nabla k(x_{p-1},x_p)=0)$, we see that the ratio $\frac{\nabla k_p(x_{p-1},x_p)}{k_p(x_{p-1},x_p)}$ is well defined. □

References

[1] M. Bolic, P. M. Djuric, and S. Hong. Resampling algorithms for particle filters: A computational complexity perspective. *EURASIP Journal on Applied Signal Processing*, 15:2267–2277, 2004.

[2] O. Cappé, R. Douc, and E. Moulines. Comparaison of resampling schemes for particle filtering. In *4th ISPA*, 2005.

[3] F. Cérou, F. LeGland, and N.J. Newton. *Stochastic particle methods for linear tangent filtering equations*, pages 231–240. IOS Press, Amsterdam, 2001.

[4] B. L. Chan, A. Doucet, and V. B. Tadic. Optimisation of particle filters using simultaneous perturbation stochastic approximation. In *ICASSP*, pages 681–685, 2003.

[5] N. Chopin. Central limit theorem for sequential monte carlo methods and its application to bayesian inference. *Ann. Statist.*, 32:2385–2411, 2004.

[6] D. Crisan and A. Doucet. A survey of convergence results on particle filtering for practitioners. *IEEE Trans. Signal Processing*, 50:736–746, 2002.

[7] R. Douc, A. Guillin, and J. Najim. Moderate deviation in particle filtering. *Ann. Appl. Probab.*, 15:587–614, 2004.

[8] R. Douc and C. Matias. Asymptotics of the maximum likelihood estimator for general hidden markov models. *Bernouilli*, 7:381–420, 2001.

[9] R. Douc and E. Moulines. Limit theorems for weighted samples with applications to sequential monte carlo methods. *Ann. Appl. Probab*, Submitted 2006.

[10] A. Doucet, N. de Freitas, and M. Klaas. Toward practicle n^2 monte carlo: the marginal particle filter. In *ICML*, 2005.

[11] A. Doucet, N. de Freitas, K. Murphy, and S. Russell. Rao-blackwellised particle filtering for dynamic bayesian networks. In *16th CUAI*, pages 176–183, 2000.

[12] A. Doucet, N. De Freitas, and N. Gordon. *Sequential Monte Carlo Methods in Practice*. Springer, 2001.

[13] A. Doucet and S. Godsill. On sequential monte carlo sampling methods for bayesian filtering. *Statistics and Computing*, 10:197–208, 2000.

[14] A. Doucet and V.B. Tadic. Parameter estimation in general state-space models using particle methods. *Ann. Inst. Stat. Math*, 2003.

[15] J. Fichoud, F. LeGland, and L. Mevel. Particle-based methods for parameter estimation and tracking : numerical experiments. Technical Report 1604, IRISA, 2003.

[16] A. Guyader, F. LeGland, and N. Oudjane. A particle implementation of the recursive mle for partially observed diffusions. In *13th IFAC Symposium on System Identification*, pages 1305–1310, 2003.

[17] P. Del Moral. *Feynman-Kac Formulae Genealogical and Interacting Particle Systems with Applications*. Springer, 2004.

[18] P. Del Moral and L. Miclo. Branching and interacting particle systems. approximations of feynman-kac formulae with applications to non-linear filtering. *Séminaire de probabilités de Strasbourg*, 34:1–145, 2000.

[19] G. Poyadjis, A. Doucet, and S.S. Singh. Particle methods for optimal filter derivative: Application to parameter estimation. In *IEEE ICASSP*, 2005.

[20] G. Poyiadjis. *Particle Method for Parameter Estimation in General State Space Models*. PhD thesis, University of Cambridge, 2006.

[21] G. Poyiadjis, S.S. Singh, and A. Doucet. Particle filter as a controlled markov chain for on-line parameter estimation in general state space models. In *IEEE ICASSP*, 2006.