Bernoulli **11**(5), 2005, 893–932

Optimal quantization methods for nonlinear filtering with discrete-time observations

GILLES PAGÈS[1] and HUYÊN PHAM[2]

[1]*Laboratoire de Probabilités et Modèles Aléatoires CNRS, UMR 7599 Université Paris 6, e-mail: gpa@ccr.jussieu.fr*
[2]*Laboratoire de Probabilités et Modèles Aléatoires CNRS, UMR 7599 Université Paris 7, e-mail: pham@math.jussieu.fr and CREST, Laboratoire de Finance-Assurance*

We develop an optimal quantization approach for numerically solving nonlinear filtering problems associated with discrete-time or continuous-time state processes and discrete-time observations. Two quantization methods are discussed: a marginal quantization and a Markovian quantization of the signal process. The approximate filters are explicitly solved by a finite-dimensional forward procedure. A posteriori error bounds are stated, and we show that the approximate error terms are minimal at some specific grids that may be computed off-line by a stochastic gradient method based on Monte Carlo simulations. Some numerical experiments are carried out: the convergence of the approximate filter as the accuracy of the quantization increases and its stability when the latent process is mixing are emphasized.

Keywords: Euler scheme; Markov chain; nonlinear filtering; numerical approximation; stationary signal; stochastic gradient descent; vector quantization

1. Introduction

We address the following nonlinear discrete-time filtering problem. In this paper, all the random variables are defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. The signal process is an $\mathbb{R}^d$-valued Markov chain $\{X_k, k \in \mathbb{N}\}$ with known transition probability $P_k(x, \mathrm{d}x')$, $k \geqslant 1$ (i.e. the transition from time $k-1$ to time k). The initial law of X_0 is known and denoted by μ. We have noisy observations $\{Y_k, k \in \mathbb{N}^*\}$ valued in $\mathbb{R}^q$, and our aim is to compute at some time $n \geqslant 1$ the conditional law $\Pi_{Y,n}$ of X_n given the observations $Y = (Y_1, \ldots, Y_n)$. In other words, we wish to calculate the conditional expectations

$$\Pi_{Y,n} f = \mathbb{E}[f(X_n) | Y_1, \ldots, Y_n], \tag{1.1}$$

for all reasonable functions f on $\mathbb{R}^d$.

Throughout the paper, we fix the observations $Y = (Y_1, \ldots, Y_n)$ at $y = (y_1, \ldots, y_n)$ and we write $\Pi_{y,n}$ for $\Pi_{Y,n}$. The initial value Y_0 is assumed for simplicity to be non-random, equal to zero for convenience.

We consider an observation process (or design) where the pair $(X_k, Y_k)_{k \in \mathbb{N}}$ is a Markov chain and such that, for all $k \geqslant 1$:

(H) The law of Y_k conditional on (X_{k-1}, Y_{k-1}, X_k) admits a density

1350–7265 © 2005 ISI/BS

$$y' \mapsto g_k(X_{k-1}, Y_{k-1}, X_k, y').$$

Notice that the transition probability of the Markov chain $(X_k, Y_k)_{k\in\mathbb{N}}$ is then given by $P_k(x, \mathrm{d}x')g_k(x, y, x', y')\mathrm{d}y'$.

An example of the observation scheme considered above is the model

$$X_k = F_k(X_{k-1}, \varepsilon_k), \qquad k = 1, \ldots, n, \tag{1.2}$$
$$Y_k = G_k(X_{k-1}, Y_{k-1}, X_k, \eta_k), \qquad k = 1, \ldots, n, \tag{1.3}$$

where $(\varepsilon_k)_k$ and $(\eta_k)_k$ are independent sequences of independent and identically distributed (i.i.d.) random variables, and F_k, G_k are measurable functions. The pair $(X_k, Y_k)_k$ is then Markovian with respect to the filtration generated by $(\varepsilon_k, \eta_k)_k$. The basic assumption concerning G_k and $(\eta_k)_k$ is that for each k, $x, x' \in \mathbb{R}^d$, $y \in \mathbb{R}^q$, the variable $G_k(x, y, x', \eta_k)$ admits a density $y' \mapsto g_k(x, y, x', y')$.

An explicit solution to problem (1.1) can be found only in very special cases: essentially when the signal-observation model forms a linear Gaussian system, leading to the well-known Kalman–Bucy filter. In the general case, the nonlinear filtering problem (1.1) leads to a dynamical system in the infinite-dimensional space of measures, and we have to search for approximate solutions.

Actually, approximations to (1.1) have been studied by various authors. We refer for example to Kushner (1977), Di Masi and Runggaldier (1982) and Di Masi *et al.* (1985) for approaches related to the one followed here. In these papers, for an observation scheme in the particular form

$$Y_k = G(X_k) + \eta_k, \qquad k = 1, \ldots, n,$$

the method consists basically of approximating the signal Markov chain $(X_k)_k$ (or, in the continuous-time problem, the signal diffusion $(X_t)_t$) by a finite state space Markov chain. This reduces the nonlinear filtering problem to an approximate resolution by an iterative finite-dimensional system. In these methods, the space grid is fixed prior to any computation and regardless of the structure of the Markov chain. So, from a computational viewpoint, they are effective only for low dimensions of the signal state space. On the other hand, although some bounds are obtained in Di Masi and Runggaldier (1982) and Di Masi *et al.* (1985), they are not sharp. Moreover, they essentially yield the convergence of the approximate filter to the true filter but do not provide an estimate of the rate of convergence.

We propose in this paper an approximation of the filter based on an optimal quantization approach. Basically this means relying on a spatial discretization of the dynamics of the signal $(X_k)_{1\leqslant k\leqslant n}$ optimally 'fitted' to its probabilistic features. Let us be more specific by considering the case of a single random vector X. If we wish to approximate X by a random vector taking its values in a finite grid $\Gamma := \{x^1, \ldots, x^N\}$, we consider its projection $\mathrm{Proj}_\Gamma(X)$ on the grid according to the nearest-neighbour rule. Then the resulting mean L^p error ($p \geqslant 1$) is $\|X - \mathrm{Proj}_\Gamma(X)\|_p = \|\min_{1\leqslant i\leqslant N}|X - x^i|\|_p$. This only depends on the distribution $\mathbb{P}_X$ of X and the grid Γ. For historical reasons, $\mathrm{Proj}_\Gamma(X)$ is often called the *quantization* of the random variable X by the grid Γ and the induced error, the *L^p mean quantization error* (see Graf and Luschgy 2000). This quantity has been extensively investigated in signal processing and information theory since the early 1950s. Thus, the L^p mean quantization error is continuous as a function of the grid Γ and reaches a minimum over all the grids Γ with size at most N. Furthermore, following Zador's theorem (see Graf and Luschgy 2000),

$$\min_{\Gamma,|\Gamma|\leqslant N} \|X - \mathrm{Proj}_\Gamma(X)\|_p = c(\mathbb{P}_X, p) \times c_2(d)N^{-1/d} + o(N^{-1/d})$$

as N goes to infinity. If $\mathbb{P}_X$ has an absolutely continuous component, $c(\mathbb{P}_X, p) > 0$; its value is known, whereas that of the universal constant $c_2(d)$ remains unknown; see Graf and Luschgy (2000) for bounds and asymptotics.

On the other hand, except in some very specific cases such as the uniform distribution over the unit interval, no closed form is available for the optimal grids that achieve the minimal quantization error of a probability distribution. In fact, no rigorous result is available to describe precisely the geometric structure or 'shape' of such an optimal grid. However, using the integral representation of $\|X - \mathrm{Proj}_\Gamma(X)\|_p^p$ one derives a stochastic gradient descent that converges towards some (at least locally) optimal grids. The distribution of $\mathrm{Proj}_\Gamma(X)$ and the resulting quantization error can be obtained as by-products, especially when in the quadratic case $p = 2$ (see Pagès 1997). Simulations like those carried out in Pagès and Printems (2003) for the two-dimensional Gaussian distribution confirm what might be expected a priori: the more heavily an area is weighted by the quantized distribution, the more points it contains.

A first application to numerical probability is proposed in Pagès (1997) for numerical integration: if $\Gamma^* = \{x^{*,1}, \ldots, x^{*,N}\}$ is an optimal grid for the *quadratic* quantization of X and if $f : \mathbb{R}^d \to \mathbb{R}$ is $\mathcal{C}^1$ with Lipschitz continuous differential Df, then

$$\mathbb{E}f(\mathrm{Proj}_{\Gamma^*}(X)) = \sum_{1\leqslant i\leqslant N} f(x^{*,i})p_i, \qquad \textit{with } p_i = \mathbb{P}(\mathrm{Proj}_{\Gamma^*}(X) = x^i),$$

$$|\mathbb{E}f(\mathrm{Proj}_{\Gamma^*}(X)) - \mathbb{E}f(X)| \leqslant [Df]_{\mathrm{Lip}}\|X - \mathrm{Proj}_{\Gamma^*}(X)\|_2^2 = O\left(\frac{1}{N^{2/d}}\right).$$

This shows that weak approximation by quantization can be superior to strong approximation, so that the quantization method may outperform Monte Carlo simulation at least up to four dimensions. Numerical experiments carried out in Pagès and Printems (2003) even suggest that this naive approach is in fact pessimistic, in particular for not too large values of N.

These ideas can be transferred to Markovian dynamics $(X_k)_k$ in order to approximate efficiently the transition distributions $\mathcal{L}(X_k \,|\, X_{k-1})$ and the joint distributions (X_k, X_{k-1}). Two different methods can be implemented: one approach gives preference to the approximation of the marginal distributions of the signal X_k at every time $k = 0, \ldots, n$; the other enhances the preservation of the dynamics, namely the Markov property. In the first case, one approximates the signal X_k at each time k by its *marginal optimal quantization*,

$$\hat{X}_k := \mathrm{Proj}_{\Gamma_k}(X_k), \qquad k = 0, \ldots, n,$$

where the grids Γ_k minimize the L^p quantization error $\|X_k - \mathrm{Proj}_{\Gamma_k}(X_k)\|_p$ among the grids with size N_k for every $k = 0, \ldots, n$. The sequence $(\hat{X}_k)_{0\leqslant k\leqslant n}$ no longer has the Markov property. Then one defines the approximate quantized filter by simply replacing in the forward explicit formula for the nonlinear filter the conditional law of X_{k+1} given X_k by the conditional law of $\hat{X}_{k+1}$ given $\hat{X}_k$. This approach was originally introduced in Bally and Pagès (2003) and Bally *et al.* (2001) to discretize reflected backward stochastic differential equations.

In the Markovian approach, and for a signal-observation model of the form (1.2)–(1.3), one sets

$$\hat{X}_k = \mathrm{Proj}_{\Gamma_k}(F_k(\hat{X}_{k-1}, \varepsilon_k)), \qquad \hat{X}_0 = \mathrm{Proj}_{\Gamma_0}(X_0).$$

Thus, the sequence $(\hat{X}_k)_{0\leqslant k\leqslant n}$ remains a Markov chain with respect to the same filtration as (X_k) but is no longer the best L^p marginal approximation of $(X_k)_{0\leqslant k\leqslant n}$. Then one considers as an approximate quantized filter the nonlinear filter of $\hat{X}_n$ conditional on the process

$$\tilde{Y}_k = G_k(\hat{X}_{k-1}, \tilde{Y}_{k-1}, \hat{X}_k, \eta_k), \qquad k = 1, \ldots, n.$$

This Markovian quantization approach was introduced in Pagès *et al.* (2004) to approximate numerically some stochastic control problems for multidimensional Markov chains.

As far as filtering is concerned, both quantization approaches make possible the analysis of the error under some appropriate Lipschitz continuity assumptions on the underlying Markov dynamics. The a priori error bounds are expressed using the quantization errors $\|Z_k - \mathrm{Proj}_{\Gamma_k}(Z_k)\|_p$, where $Z_k = X_k$ in the marginal approach and $Z_k = F(\hat{X}_{k-1}, \varepsilon_k)$ in the Markovian approach. Although the methods of proof are significantly different, the a priori error bounds look quite similar for both methods, suggesting a balance between the positive and negative features of the two approximation methods. An extensive discussion and comparison of both quantization methods, marginal and Markovian, is carried out in Pagès *et al.* (2004b). For a detailed description of the algorithms we refer to Bally and Pagès (2003) and Pagès *et al.* (2004a) in which the methods were originally introduced.

In Section 6, we analyse the practical aspects of the algorithm in terms of complexity. The most interesting feature of the quantization approach is that, once an optimal quantization of the signal $(X_k)_{0\leqslant k\leqslant n}$ has been processed and kept off-line, it *instantaneously produces* for any set of observations some reproducible deterministic results. This underlines the fact that the set of observations is reasonably likely since the approximate filter distribution is structurally supported by the quantization of X. This can be implemented with multidimensional signal processes, at least up to four dimensions and possibly higher. This restriction on the dimension comes from the fact that, for a given size N of the quantization, the error does depend on the dimension d of the signal as for numerical integration.

In the special case of a stationary signal, the marginal quantization of the whole Markov chain reduces to that of its stationary distribution, so that the the quantization optimization phase – which is clearly the most demanding one – is divided by a factor n in terms of duration and storage.

In recent years, a technique for approximating the nonlinear filtering problem has received much attention: it is a Monte Carlo method based on interacting particle systems (see Del Moral 1998; Del Moral *et al.* 2001; Florchinger and LeGland 1992; Crisan and Lyons 1997); this is a typical 'on-line' method (the whole process involves the observation set y and the function f), while quantization is typically an off-line method (the demanding part of the computations can be stored, those depending on y and f being instantaneous). So it is rather difficult to define a comparison protocol since their fields of applications are quite different.

It may also be interesting to keep the same filter for close observation samples in order to avoid heavy computations. This can be processed by replacing the observations by some discrete values. This point of view is investigated for real-valued observations in Newton (2000a; 2000b): some functional weak convergence results toward the original filter are given. In a framework where both X and the observation process are quantized, some a priori error bounds are derived in Sellami (2004).

This paper is organized as follows. Some preliminaries on nonlinear filtering are provided in Section 2. In particular, we recall the well-known forward inductive formula for the filter, and also the (less well-known) backward formula. In Sections 3 and 4 we study the approximate filter by marginal and Markovian quantization respectively, including an explicit error analysis. In Section 5 the convergence of both quantized approximating filters is established. In Section 6 we show how to obtain optimal grids for both quantization methods, and we discuss their respective qualities and drawbacks from a practical viewpoint, especially concerning the optimization phase. We point out that the marginal quantization approach can be significantly simplified from a computational viewpoint in the important case of stationary signal processes. We discuss in Section 7 how our results may be applied to the case of discretely observed diffusions. Finally, in Section 8 we describe several numerical experiments. First, we compare our approximate filter with the explicit Kalman–Bucy filter: on the one hand its convergence – with some rate – as the size of the quantization increases is confirmed, and on the other hand its stability as n increases. Then we evaluate the approximate filter by quantization in a state model with multiplicative Gaussian noise arising in stochastic volatility models: a convergent behaviour is obtained as the quantization accuracy increases, although no reference value is available.

We close this introductory section with a couple of observations concerning notation. First, for every Borel function $f : \mathbb{R}^d \to \mathbb{R}$, set

$$\|f\|_\infty = \sup_{x\in\mathbb{R}^d} |f(x)| \quad \text{and} \quad [f]_{\mathrm{Lip}} = \sup_{x\neq x'} \frac{|f(x) - f(x')|}{|x - x'|}.$$

Second, we use the traditional notation for transition kernels: if P is a bounded transition kernel and f is a bounded measurable function, we write Pf for the bounded measurable function $Pf(x) := \int f(x')P(x, \mathrm{d}x')$.

2. Nonlinear filtering: preliminaries and remarks

In this section, we recall some useful facts about nonlinear filtering. Using the Markov property of the pair (X, Y) and the Bayes formula, one can derive the Kallianpur–Striebel (1968) formula for the the filter:

$$\Pi_{y,n}f := \frac{\pi_{y,n}f}{\pi_{y,n}\mathbf{1}}, \tag{2.1}$$

where $\pi_{y,n}$ is the so-called unnormalized filter defined by

$$\pi_{y,n}f = \int f(x_n)\mu(\mathrm{d}x_0)\prod_{k=1}^{n} g_k(x_{k-1}, y_{k-1}, x_k, y_k)P_k(x_{k-1}, \mathrm{d}x_k), \tag{2.2}$$

$$= \mathbb{E}[f(X_n)L_{y,n}], \tag{2.3}$$

with

$$L_{y,n} := \prod_{k=1}^{n} g_k(X_{k-1}, y_{k-1}, X_k, y_k) \tag{2.4}$$

(and by convention, $y_0 = 0$). Notice that

$$\pi_{y,n}\mathbf{1} = \mathbb{E}[L_{y,n}] = \phi_n(y), \tag{2.5}$$

where $\phi_n(y)$ is the value of the density function ϕ_n of $(Y_1, \ldots, Y_n)$ with respect to the Lebesgue measure at the observed values $y = (y_1, \ldots, y_n) \in (\mathbb{R}^q)^n$.

Henceforth, we shall write, for notational convenience,

$$g_{y,k}(x, x') = g_k(x, y_{k-1}, x', y_k), \qquad k \geqslant 1.$$

The unnormalized filter can be written using a family of bounded transition kernels $H_{y,k}$, $k = 1, \ldots, n$, defined on bounded measurable functions $f : \mathbb{R}^d \to \mathbb{R}$ by

$$H_{y,k}f(x) := \mathbb{E}[f(X_k)g_{y,k}(x, X_k)|X_{k-1} = x] = \int f(x')g_{y,k}(x, x')P_k(x, \mathrm{d}x'), \qquad x \in \mathbb{R}^d.$$

For convenience we also define

$$H_{y,0}f(x) := \pi_{y,0}f = \mathbb{E}[f(X_0)] = \int f(x_0)\mu(\mathrm{d}x_0), \qquad x \in \mathbb{R}^d.$$

Then, one can show that the unnormalized filter at time k, $\pi_{y,k} := \mathbb{E}[f(X_k)L_{y,k}]$, satisfies the inductive formula

$$\pi_{y,k}f = \pi_{y,k-1}H_{y,k}f, \qquad k = 1, \ldots, n \tag{2.6}$$

so that

$$\pi_{y,n} = H_{y,0}\circ H_{y,1}\circ \cdots \circ H_{y,n}. \tag{2.7}$$

Equation (2.6) is called the *forward expression* for the filter. One can also derive from the 'symmetric' expression (2.7) a *backward expression* for the filter which will turn out to be useful for our proofs, namely

$$\pi_{y,n}f = u_{y,-1}(f),$$

where $u_{y,-1}(f)$ is defined as the final value of the backward induction

$$\begin{aligned} &u_{y,n}(f)(x) = f(x), \\ &u_{y,k-1}(f) = H_{y,k}\, u_{y,k}(f), \qquad k = 0, \ldots, n. \end{aligned} \tag{2.8}$$

Note that in fact $u_{y,k}f = H_{y,k-1}\circ\cdots\circ H_{y,n}f$.

We shall replace the true filter by a computable approximate filter. This will follow from a spatial discretization of the signal process $(X_k)_k$ based on optimal quantization. We will propose two types of quantization – marginal and Markovian – leading to different approximations $\hat{H}_{y,k}$ of the transition kernel $H_{y,k}$ both based on (2.6). Then we will compute in both cases an approximate distribution $\hat{\pi}_{y,n,}$ of the unnormalized filter $\pi_{y,n}$ using a quantized form of the forward expression (2.6), $\hat{\pi}_{y,k}f = \hat{\pi}_{y,k-1}\hat{H}_{y,k}f$.

3. Approximate filter by marginal quantization

3.1. Method

In this section, we consider a marginal quantization of the Markov chain $(X_k)_k$, in temporal sequence, that is,

$$\hat{X}_k = \mathrm{Proj}_{\Gamma_k}(X_k), \qquad 0 \leqslant k \leqslant n, \tag{3.1}$$

where Γ_k, $k = 0, \ldots, n$, are grids consisting of N_k points x_k^i in $\mathbb{R}^d$, $i = 1, \ldots, N_k$, to be optimized later, and Proj_{Γ_k} denotes the nearest-neighbour projection on Γ_k. Notice that the process $(\hat{X}_k)_k$ is not a Markov chain. We construct an approximate filter based on an approximation of the transition probability $P_k(x_k, \mathrm{d}x_{k+1})$ of X_{k+1} given X_k by the transition probability matrix $\hat{P}_k := [\hat{P}_k^{ij}]$ of $\hat{X}_{k+1}$ given $\hat{X}_k$:

$$\hat{P}_k^{ij} = \mathbb{P}[\hat{X}_k = x_k^j \,|\, \hat{X}_{k-1} = x_{k-1}^i], \qquad i = 1, \ldots, N_{k-1}, j = 1, \ldots, N_k. \tag{3.2}$$

In other words, we approximate the transition kernel $H_{y,k}$ by the quantized transition kernel $\hat{H}_{y,k}$ given by

$$\hat{H}_{y,k} := \sum_{j=1}^{N_k} \hat{H}_{y,k}^{ij}\delta_{x_{k-1}^i}, \qquad k = 1, \ldots, n, \tag{3.3}$$

with

$$\hat{H}_{y,k}^{ij} = g_{y,k}(x_{k-1}^i, x_k^j)\hat{P}_k^{ij}, \qquad i = 1, \ldots, N_{k-1}, j = 1, \ldots, N_k, \tag{3.4}$$

for $k = 1, \ldots, n$, so that, for every function $f : \Gamma_k \to \mathbb{R}$,

$$\hat{H}_{y,k}f(\hat{X}_{k-1}) := \mathbb{E}[g_{y,k}(\hat{X}_{k-1}, \hat{X}_k)f(\hat{X}_k) \,|\, \hat{X}_{k-1}], \qquad k = 1, \ldots, n.$$

Finally, we set

$$\hat{H}_{y,0} = \sum_{i=1}^{N_0} \hat{P}_0^i\delta_{x_0^i} \qquad \text{with } \hat{P}_0^i := \mathbb{P}[\hat{X}_0 = x_0^i], \quad i = 1, \ldots, N_0.$$

We then define the approximate unnormalized filter $\hat{\pi}_{y,n} = \sum_{i=1}^{N_n}\hat{\pi}_{y,n}^i\delta_{x_n^i}$ by

$$\hat{\pi}_{y,n} = \hat{H}_{y,0}\circ\cdots\circ\hat{H}_{y,n}.$$

This is easily computed by the following forward induction:

$$\hat{\pi}_{y,0} = \hat{H}_{y,0},$$

$$\hat{\pi}_{y,k} = \pi_{y,k-1}\hat{H}_{y,k} := \left[\sum_{i=1}^{N_{k-1}} \hat{H}^{ij}_{y,k}\hat{\pi}^{i}_{y,k-1}\right]_{j=1,\dots,N_k}, \; k = 1, \dots, n. \tag{3.5}$$

The approximate filter $\hat{\Pi}_{y,n}$ is then given by

$$\hat{\Pi}_{y,n} = \sum_{i=1}^{N_n} \hat{\Pi}^{i}_{y,n}\delta_{x^i_n}$$

with

$$\hat{\Pi}^{i}_{Y,n} = \frac{\hat{\pi}^{i}_{y,n}}{\sum_{i=1}^{N_n}\hat{\pi}^{i}_{y,n}}, \qquad i = 1, \dots, N_n.$$

3.2 Error analysis

In this subsection, we will estimate the accuracy of the approximate filter $\hat{\Pi}_{y,n}$ in terms of the marginal quantization errors on the signal $\|\Delta_k\|_1$, $k = 0, \dots, n$, defined by

$$\Delta_k = X_k - \mathrm{Proj}_{\Gamma_k}(X_k). \tag{3.6}$$

Note that the process $(\hat{X}_k)$ is not a Markov chain. We shall impose some Lipschitz conditions on the Markov transition of X_k and on the conditional law Y_k given X_{k-1}, Y_{k-1}, X_k. We first recall some definitions. We say that a transition probability P on $\mathbb{R}^d$ is C-Lipschitz for some positive real constant C if, for every Lipschitz function φ on $\mathbb{R}^d$ with ratio $[\varphi]_{\mathrm{Lip}}$, $P\varphi$ is Lipschitz and $[P\varphi]_{\mathrm{Lip}} \leqslant C[\varphi]_{\mathrm{Lip}}$. Then we may define the Lipschitz ratio

$$[P]_{\mathrm{Lip}} = \sup\left\{\frac{[P\varphi]_{\mathrm{Lip}}}{[\varphi]_{\mathrm{Lip}}}, \varphi \text{ a non-zero Lipschitz continuous function}\right\} < +\infty.$$

(A1) The Markov transition operators $P_k(x, \mathrm{d}x')$, $k = 1, \dots, n$, are Lipschitz, so that

$$[P]_{\mathrm{Lip}} := \max_{k=0,\dots,n} [P_k]_{\mathrm{Lip}} < +\infty.$$

(A2) (i) For every $k = 1, \dots, n$, the functions g_k are bounded on $\mathbb{R}^d \times \mathbb{R}^q \times \mathbb{R}^d \times \mathbb{R}^q$ and we set $K_g := \max_{k=1,\dots,n}\|g_k\|_\infty$.

(ii) For every $k = 1, \dots, n$, there exist two Borel functions $[g^1_k]_{\mathrm{Lip}}$, $[g^2_k]_{\mathrm{Lip}} : \mathbb{R}^q \times \mathbb{R}^q \to \mathbb{R}_+$ such that, for all x, x', $\hat{x}$, $\hat{x}' \in \mathbb{R}^d$ and y, $y' \in \mathbb{R}^q$,

$$|g_k(x, y, x', y') - g_k(\hat{x}, y, \hat{x}', y')| \leqslant [g^1_k]_{\mathrm{Lip}}(y, y')|x - \hat{x}| + [g^2_k]_{\mathrm{Lip}}(y, y')|x' - \hat{x}'|.$$

An essential device for the proof of Theorem 3.1 below is to introduce the sequence of functions $(\hat{u}_{y,k}(f))_{-1\leqslant k\leqslant n}$ which is the quantized counterpart of the sequence

$(u_k(f))_{-1\leqslant k\leqslant n}$ defined in (2.8) as the backward expression of the filter: mimicking this backward dynamic formula, we recursively define the $\hat{u}_{y,k}(f)$ on Γ_k, $k=0,\ldots,n$, by

$$\begin{aligned}
&\hat{u}_{y,n}(f)=f, && \text{on the grid } \Gamma_n,\\
&\hat{u}_{y,k}(f)=\hat{H}_{y,k+1}\hat{u}_{y,k+1}(f), && \text{on the grid } \Gamma_{k-1},\ k=-1,\ldots,n-1.
\end{aligned}$$

The approximate unnormalized filter $\hat{\pi}_{y,n}$ is then given by

$$\hat{\pi}_{y,n}f=\hat{u}_{y,-1}(f),$$

so that $|\pi_{y,n}f-\hat{\pi}_{y,n}f|=|u_{y,-1}(f)-\hat{u}_{y,-1}(f)|$.

Theorem 3.1. *Assume that* (A1) *and* (A2) *hold. Then, for every bounded Lipschitz continuous function f on $\mathbb{R}^d$ and each n-tuple of observations $y=(y_1,\ldots,y_n)$, we have, for every $p\geqslant 1$,*

$$|\Pi_{y,n}f-\hat{\Pi}_{y,n}f|\leqslant\frac{K_g^n}{\phi_n(y)\vee\hat{\phi}_n(y)}\sum_{k=0}^{n}B_k^n(f,y,p)\|\Delta_k\|_p. \tag{3.7}$$

with

$$\hat{\phi}_n(y):=\hat{\pi}_{y,n}\mathbf{1}, \tag{3.8}$$

$$\begin{aligned}
B_k^n(f,y,p):=&(2-\delta_{2,p})[P]_{\text{Lip}}^{n-k}[f]_{\text{Lip}}+2\Bigg(\frac{\|f\|_\infty}{K_g}\left([g_{k+1}^1]_{\text{Lip}}(y_k,y_{k+1})+[g_k^2]_{\text{Lip}}(y_{k-1},y_k)\right)\\
&+(2-\delta_{2,p})\frac{\|f\|_\infty}{K_g}\sum_{j=k+1}^{n}[P]_{\text{Lip}}^{j-(k+1)}\Big([g_j^1]_{\text{Lip}}(y_{j-1},y_j)+[P]_{\text{Lip}}[g_j^2]_{\text{Lip}}(y_{j-1},y_j)\Big)\Bigg).
\end{aligned} \tag{3.9}$$

(By convention, $g_0=g_{n+1}\equiv 0$, and $\delta_{r,p}$ is the usual Kronecker delta.)

Remark 3.1. Note that

$$\hat{\phi}_n(y)=\hat{\pi}y,\ n\mathbf{1}=\sum_{i=1}^{N_n}\hat{\pi}_{y,n}^i$$

is the normalizing factor of the approximate filter distribution so that (3.7) produces a completely computable error bound.

Remark 3.2. The interesting case for the general L^p bounds is the case $p=2$ where the coefficients $B_n^k(f,y,p)$ are smaller than in the L^1 case (other bounds are trivial since the L^p norm is non-decreasing as a function of p).

Remark 3.3. If we introduce

$$[g]_{\mathrm{Lip}} := \max_{k=1,\ldots,n} \sup_{y,y'\in\mathbb{R}^q} ([g_k^1]_{\mathrm{Lip}}(y,\,y') \vee [g_k^2]_{\mathrm{Lip}}(y,\,y')),$$

then $B_k^n(f,\,y,\,p)$ is upper-bounded by the simpler coefficient

$$\tilde{B}_k^n(f,\,p) := (2-\delta_{2,p})[P]_{\mathrm{Lip}}^{n-k}[f]_{\mathrm{Lip}} + 2\frac{\|f\|_\infty}{K_g}[g]_{\mathrm{Lip}}\left(2 + (2-\delta_{2,p}\frac{[P]_{\mathrm{Lip}}+1}{[P]_{\mathrm{Lip}}-1}(\mathrm{p}P]_{\mathrm{Lip}}^{n-k} - 1)\right).$$

with the usual convention

$$\frac{1}{u-1}\times(u^m - 1) = m, \qquad \text{when } u = 1 \text{ and } m \in \mathbb{N}.$$

Remark 3.4. Suppose the Lipschitz condition (A2)(ii) is weakened into a local Lipschitz one:

(A2) (ii′) For every $k = 1,\ldots,n$, there exist two Borel functions $[g_k^1]_{\mathrm{Liploc}}$, $[g_k^2]_{\mathrm{Liploc}} : \mathbb{R}^q \times \mathbb{R}^q \mapsto \mathbb{R}_+$ such that, for all $x,\,x',\,\hat{x},\,\hat{x}' \in \mathbb{R}^q$,

$$|g_k(x,\,y,\,x',\,y') - g_k(\hat{x},\,y,\,\hat{x}',\,y')| \leqslant [g_k^1]_{\mathrm{Liploc}}(y,\,y')(1+|x|+|x'|+|\hat{x}|+\hat{x}'|)|x-\hat{x}|$$
$$+ [g_k^2]_{\mathrm{Liploc}}(y,\,y')(1+|x|+x'|+|\hat{x}|+|\hat{x}'|)|x'-\hat{x}'|.$$

Then we may state an estimate for the approximate filter similar to that in Theorem 3.1: for every $p \geqslant 1$ and every $p',\,q' \in (1,\infty)$, $1/p' + 1/q' = 1$,

$$|\Pi_{y,n}f - \hat{\Pi}_{y,n}f| \leqslant \frac{K_g^n}{\phi_n(y)}\sum_{k=0}^{n}\overline{B}_k^n(p,\,pq',\,f)\|\Delta_k\|_{pp'},$$

with

$$\overline{B}_k^n(p,\,r,\,f) = (2-\delta_{2,p})[P]_{\mathrm{Lip}}^{n-k}[f]_{\mathrm{Lip}}$$
$$+ 2\frac{\|f\|_\infty}{K_g}[g]_{\mathrm{Lip}}\left(2 + (2-\delta_{2,p})\frac{[P]_{\mathrm{Lip}}+1}{[P]_{\mathrm{Lip}}-1}([P]_{\mathrm{Lip}}^{n-k}-1)\right)M_n(r),$$

$$M_n(r) = 1 + 2\|X\|_r + 2\|\hat{X}\|_r, \qquad r \geqslant 1.$$

Here we have set

$$[g]_{\mathrm{Liploc}} := \max_{k=1,\ldots,n}\sup_{y,y'\in\mathbb{R}^q}([g_k^1]_{\mathrm{Liploc}}(y,\,y') \vee [g_k^2]_{\mathrm{Liploc}}(y,\,y')),$$

$$\|X\|_q = \max_{k=0,\ldots,n}\|X_k\|_q, \qquad \|\hat{X}\|_q = \max_{k=0,\ldots,n}\|\hat{X}_k\|_q.$$

Furthermore, when $\hat{X}$ is an *optimal quadratic quantization* it can be shown (see Graf and Luschgy 2000, or (B.7) in Appendix B) that $\hat{X}_k = \mathbb{E}(X_k|\hat{X}_k)$ so that $\|\hat{X}_k\|_r \leqslant \|X_k\|_r$ for every $r \geqslant 1$. Hence, one may take

$$M_n(r) = 1 + 4\|X\|_r, \qquad r \geqslant 1.$$

We will see in Section 7 (and Appendix A) that assumption (A2) is satisfied by the conditional density of certain discretely observed diffusion models.

To obtain the announced error bound, three steps are required. The first one, in Lemma 3.1, makes an abstract connection between errors in the unnormalized world and the normalized world (it is used in next section too). The second one, in Lemma 3.2, yields a bound for the Lipschitz coefficient of the functions $u_{y,k}(f)$ defined in (2.8). In the third step – which is the proof of the theorem itself – we will bound $\|u_{y,k}(f)(X_k) - \hat{u}_{y,k}(f)(\hat{X}_k)\|_p$ by a backward induction, bearing in mind that $|\pi_{y,n}f - \hat{\pi}_{y,n}f| = |u_{y,-1}(f)(X_k) - \hat{u}_{y,-1}(f)(\hat{X}_k)|$.

Lemma 3.1. *Let* (μ_y) *and* (ν_y) *two families of finite positive measures on a measurable space* $(E, \mathcal{E})$. *Assume that there exist two symmetric functions* R *and* S *defined on the set of positive finite measures such that, for every bounded Lipschitz function* f,

$$\left|\int f \mathrm{d}\mu_y - \int f \mathrm{d}\nu_y\right| \leqslant \|f\|_\infty R(\mu_y, \nu_y) + [f]_{\mathrm{Lip}} S(\mu_y, \nu_y). \tag{3.10}$$

Then

$$\left|\frac{\int d\mu_y}{\mu_y(E)} - \frac{\int f d\nu_y}{\nu_y(E)}\right| \leqslant \frac{1}{\mu_y(E) \vee \nu_y(E)} \big(2\|f\|_\infty R(\mu_y, \nu_y) + [f]_{\mathrm{Lip}} S(\mu_y, \nu_y)\big).$$

Proof. We have

$$\begin{aligned}\left|\frac{\int \mathrm{d}\mu_y}{\mu_y(E)} - \frac{\int f \mathrm{d}\nu_y}{\nu_y(E)}\right| &\leqslant \frac{|\int \mathrm{d}\mu_y - \int f \mathrm{d}\nu_y|}{\mu_y(E)} + \int |f| \mathrm{d}\nu_y \left|\frac{1}{\mu_y(E)} - \frac{1}{\nu_y(E)}\right| \\ &\leqslant \frac{\|f\|_\infty R(\mu_y, \nu_y) + [f]_{\mathrm{Lip}} S(\mu_y, \nu_y)}{\mu_y(E)} + \|f\|_\infty \left|\frac{\nu_y(E)}{\mu_y(E)} - 1\right| \\ &\leqslant \frac{\|f\|_\infty R(\mu_y, \nu_y) + [f]_{\mathrm{Lip}} S(\mu_y, \nu_y) + \|f\|_\infty |\nu_y(E) - \mu_y(E)|}{\mu_y(E)}.\end{aligned}$$

Now $|\mu_y(E) - \nu_y(E)| \leqslant 1 \times R(\mu_y, \nu_y)$, so that

$$\left|\frac{\int f \mathrm{d}\mu_y}{\mu_y(E)} - \frac{\int f \mathrm{d}\nu_y}{\nu_y(E)}\right| \leqslant \frac{1}{\mu_y(E)} \big(2\|f\|_\infty R(\mu_y, \nu_y) + [f]_{\mathrm{Lip}} S(\mu_y, \nu_y)\big).$$

A symmetry argument completes the proof. □

Lemma 3.2. *Assume that* (A1) *and* (A2) *hold. Let* $(y_k)_{k=1,\dots,n}$ *be a generic observation. Then, for every bounded Lipschitz continuous function* f, *the functions* $u_{y,k}(f)$ *defined by* (2.8) *are bounded Lipschitz continuous as well, with Lipschitz coefficient* $[u_{y,k}(f)]_{\mathrm{Lip}}$ *satisfying*

$$[u_{y,k}(f)]_{\mathrm{Lip}} \leqslant [P_{k+1}]_{\mathrm{Lip}}\Big(K_g[u_{y,k+1}(f)]_{\mathrm{Lip}} + \|u_{y,k+1}(f)\|_\infty[g^2_{y,k+1}]_{\mathrm{Lip}}\Big)$$
$$+ \|u_{y,k+1}(f)\|_\infty[g^1_{y,k+1}]_{\mathrm{Lip}}, \qquad k = 0, \ldots, n-1,$$

and

$$\|u_{y,k}(f)\|_\infty \leqslant K_g^{n-k}\|f\|_\infty, \qquad k = 0, \ldots, n.$$

In particular, for every $k \in \{0, \ldots, n\}$,

$$[u_{y,k}(f)]_{\mathrm{Lip}} \leqslant ([P]_{\mathrm{Lip}}K_g)^{n-k}[f]_{\mathrm{Lip}} + \|f\|_\infty K_g^{n-(k+1)}\sum_{\ell=1}^{n-k}[P]^{\ell-1}_{\mathrm{Lip}}([g^1_{k+\ell}]_{\mathrm{Lip}} + [P]_{\mathrm{Lip}}[g^2_{k+\ell}]_{\mathrm{Lip}}).$$

For notational convenience, we will temporarily drop the dependency in the function f and in the observation sequence y in the proofs below.

Proof. One can derive the first two formulae from the recursive definition (2.8) of the u_k:

$$u_k(x) = \mathbb{E}[g_{k+1}(x, X_{k+1})u_{k+1}(X_{k+1})|X_k = x] = \int g_{k+1}(x, x')u_{k+1}(x')P_{k+1}(x, \mathrm{d}x')$$

and from the Lipschitz property of the transitions $P_k(x, \mathrm{d}x')$:

$$[u_k]_{\mathrm{Lip}} \leqslant [P_{k+1}]_{\mathrm{Lip}} \sup_{x\in\mathbb{R}^d}[x' \mapsto g_{k+1}(x, x')u_{k+1}(x')]_{\mathrm{Lip}} + \|u_{k+1}\|_\infty[g^1_{k+1}]_{\mathrm{Lip}}.$$

Now $\|u_n\|_\infty = \|f\|_\infty$ and $\|u_k\|_\infty \leqslant K_g\|u_{k+1}\|_\infty$ so that $\|u_k\|_\infty \leqslant K_g^{n-k}\|f\|_\infty$. Hence,

$$[u_k]_{\mathrm{Lip}} \leqslant A[u_{k+1}]_{\mathrm{Lip}} + B\,K_g^{-(k+1)}[g^2_{k+1}]_{\mathrm{Lip}} + CK_g^{-(k+1)}[g^1_{k+1}]_{\mathrm{Lip}}$$

with $A = [P]_{\mathrm{Lip}}K_g$, $B := [P]_{\mathrm{Lip}}\|f\|_\infty K_g^n$ and $C := \|f\|_\infty K_g^n$. Standard computations complete the proof. □

Proof of Theorem 3.1. To obtain an upper bound for $\|u_k(X_k) - \hat{u}_k(\hat{X}_k)\|_p$, we proceed by induction. Temporarily set, for every $k \leqslant n-1$ and every $x_k, x_{k+1}, x'_{k+1} \in \mathbb{R}^d$,

$$\varphi(x_k, x_{k+1}, x'_{k+1}) := g_{k+1}(x_k, x_{k+1})u_{k+1}(x'_{k+1}).$$

Then,

$$\|u_k(X_k) - \hat{u}_k(\hat{X}_k)\|_p = \|\mathbb{E}(\varphi(X_k, X_{k+1}, X_{k+1})\,|\,X_k) - \mathbb{E}(g_{k+1}(\hat{X}_k, \hat{X}_{k+1})\hat{u}_{k+1}(\hat{X}_{k+1})\,|\,\hat{X}_k)\|_p.$$

Using the fact that $\mathbb{E}(\,.\,|\,\hat{X}_k)$ is an L^p contraction, we obtain, for every $k \in \{0, \ldots, n-1\}$,

$$\begin{aligned}\|u_k(X_k) - \hat{u}_k(\hat{X}_k)\|_p &\leqslant \|\mathbb{E}(\varphi(X_k, X_{k+1}, X_{k+1})\,|\,X_k) - \mathbb{E}(\varphi(\hat{X}_k, \hat{X}_{k+1}, X_{k+1})\,|\,\hat{X}_k)\|_p \\ &\quad + \|\varphi(\hat{X}_k, \hat{X}_{k+1}, X_{k+1}) - g_{k+1}(\hat{X}_k, \hat{X}_{k+1})\hat{u}_{k+1}(\hat{X}_{k+1})\|_p \\ &\leqslant \|\mathbb{E}(\varphi(X_k, X_{k+1}, X_{k+1})\,|\,\mathcal{F}_k) - \mathbb{E}(\varphi(\hat{X}_k, \hat{X}_{k+1}, X_{k+1})\,|\,\hat{X}_k)\|_p \\ &\quad + K_g\|u_{k+1}(X_{k+1}) - \hat{u}_{k+1}(\hat{X}_{k+1})\|_p. \end{aligned} \tag{3.11}$$

Now

$$\|\mathbb{E}(\varphi(X_k, X_{k+1}, X_{k+1}) \,|\, X_k) - \mathbb{E}(\varphi(\hat{X}_k, \hat{X}_{k+1}, X_{k+1}) \,|\, \hat{X}_k)\|_p$$
$$\leqslant \|u_k(X_k) - \mathbb{E}(u_k(X_k) \,|\, \hat{X}_k)\|_p + \|\mathbb{E}(u_k(X_k) \,|\, \hat{X}_k) - \mathbb{E}(\varphi(\hat{X}_k, \hat{X}_{k+1}, X_{k+1}) \,|\, \hat{X}_k)\|_p. \quad (3.12)$$

Then

$$\|u_k(X_k) - \mathbb{E}(u_k(X_k) \,|\, \hat{X}_k)\|_p \leqslant \|u_k(X_k) - u_k(\hat{X}_k)\|_p + \|\mathbb{E}(u_k(\hat{X}_k) - u_k(X_k) \,|\, \hat{X}_k)\|_p$$
$$\leqslant 2\|u_k(X_k) - u_k(\hat{X}_k)\|_p$$

since conditional expectation is an L^p contraction. When $p = 2$,

$$\|u_k(X_k) - \mathbb{E}(u_k(X_k) \,|\, \hat{X}_k)\|_2 = \min\big\{\|u_k(X_k) - \psi(\hat{X}_k)\|_2,\, \psi(\hat{X}_k) \in L^2\big\} \leqslant \|u_k(X_k) - u_k(\hat{X}_k)\|_2.$$

Now $\hat{X}_k$ being $\sigma(X_k)$-measurable, $\mathbb{E}(u_k(X_k) \,|\, \hat{X}_k) = \mathbb{E}(\varphi(X_k, X_{k+1}, X_{k+1}) \,|\, \hat{X}_k)$, so that

$$\|\mathbb{E}(u_k(X_k) \,|\, \hat{X}_k) - \mathbb{E}(\varphi(\hat{X}_k, \hat{X}_{k+1}, X_{k+1}) \,|\, \hat{X}_k)\|_p \leqslant \|\varphi(X_k, X_{k+1}, X_{k+1}) - \varphi(\hat{X}_k, \hat{X}_{k+1}, X_{k+1})\|_p$$
$$\leqslant \|u_{k+1}\|_\infty \|g_{k+1}(X_k, X_{k+1}) - g_{k+1}(\hat{X}_k, \hat{X}_{k+1})\|_p.$$

Consequently, substitution into (3.12) yields

$$\|\mathbb{E}(\varphi(X_k, X_{k+1}, X_{k+1}) \,|\, X_k) - \mathbb{E}(\varphi(\hat{X}_k, \hat{X}_{k+1}, X_{k+1}) \,|\, \hat{X}_k)\|_p$$
$$\leqslant (2 - \delta_{p,2})[u_k]_{\mathrm{Lip}} \|X_k - \hat{X}_k\|_p$$
$$+ \|u_{k+1}\|_\infty \big([g^1_{k+1}]_{\mathrm{Lip}} \|X_k - \hat{X}_k\|_p + [g^2_{k+1}]_{\mathrm{Lip}} \|X_{k+1} - \hat{X}_{k+1}\|_p\big).$$

Since we are dealing with marginal quantization $\Delta_k = X_k - \hat{X}_k$, substitution into (3.11) yields the following induction: for every $k \in \{0, \ldots, n-1\}$,

$$\|u_k(X_k) - \hat{u}_k(\hat{X}_k)\|_p \leqslant K_g \|u_{k+1}(X_{k+1}) - \hat{u}_{k+1}(\hat{X}_{k+1})\|_p + \alpha_k \|\Delta_k\|_p + \beta_{k+1} \|\Delta_{k+1}\|_p,$$

with

$$\alpha_k := (2 - \delta_{p,2})[u_k]_{\mathrm{Lip}} + \|u_{k+1}\|_\infty [g^1_{k+1}]_{\mathrm{Lip}}, \qquad 0 \leqslant k \leqslant n-1,$$
$$\beta_k := [g^2_k]_{\mathrm{Lip}} \|u_k\|_\infty,\ 1 \leqslant k \leqslant n.$$

For notational convenience, we set $\alpha_n := (2 - \delta_{p,2})[f]_{\mathrm{Lip}}$ (in fact $\alpha_n = [f]_{\mathrm{Lip}}$ would always be suitable) and $\beta_0 := 0$. Then, standard computations using Lemma 3.2 yield

$$|\pi_{y,n} f - \hat{\pi}_{y,n} f| = \|u_0(f)(X_0) - \hat{u}_0(f)(\hat{X}_0)\|_1$$
$$\leqslant \|u_0(f)(X_0) - \hat{u}_0(f)(\hat{X}_0)\|_p \leqslant \sum_{k=0}^{n} C^n_k(f, y, p) \|\Delta_k\|_p$$

where, for every $0 \leqslant k \leqslant n$,

$$
\begin{aligned}
&C_k^n(f, y, p)\\
&\quad := K_g^{k-1}(\alpha_k K_g + \beta_k)\\
&\quad = (2 - \delta_{2,p}) K_g^k [u_k]_{\mathrm{Lip}} + K_g^{n-1} \|f\|_\infty ([g_{k+1}^1]_{\mathrm{Lip}} + [g_k^2]_{\mathrm{Lip}})\\
&\quad \leqslant K_g^n \Big[(2 - \delta_{2,p}) [P]_{\mathrm{Lip}}^{n-k} [f]_{\mathrm{Lip}}\\
&\qquad + \frac{\|f\|_\infty}{K_g} \left([g_{k+1}^1]_{\mathrm{Lip}} + [g_k^2]_{\mathrm{Lip}} + (2 - \delta_{2,p}) \sum_{m=1}^{n-k} [P]_{\mathrm{Lip}}^{m-1} \big([g_{k+m}^1]_{\mathrm{Lip}} + [P]_{\mathrm{Lip}} [g_{k+m}^2]_{\mathrm{Lip}} \big) \right) \Big].
\end{aligned}
$$

An application of Lemma 3.1 concludes the proof. □

4. Approximate filter by Markovian quantization

4.1. Method

This method is based on the Markovian quantization developed in Pagès *et al*. (2004a). We assume that the signal-observation model is given by (1.2)–(1.3). At each time $k = 0, \ldots, n$, we are given a grid Γ_k consisting of N_k points x_k^i in $\mathbb{R}^d$, $i = 1, \ldots, N_k$, to be optimized later on. We then consider the Markovian process $(\hat{X}_k, \tilde{Y}_k)_k$ defined by

$$
\begin{aligned}
\hat{X}_k &= \mathrm{Proj}_{\Gamma_k}\big(F_k(\hat{X}_{k-1}, \varepsilon_k)\big), \qquad k = 1, \ldots, n, && (4.1)\\
\tilde{Y}_k &= G_k(\hat{X}_{k-1}, \tilde{Y}_{k-1}, \hat{X}_k, \eta_k), \quad k = 1, \ldots, n, && (4.2)
\end{aligned}
$$

with $\hat{X}_0 = \mathrm{Proj}_{\Gamma_0}(X_0)$ and $\tilde{Y}_0 = Y_0 = 0$. Here Proj_{Γ_k} still denotes the nearest-neighbour projection on Γ_k. The idea is now to approximate the filter $\Pi_{y,n}$ by the discrete conditional law $\hat{\Pi}_{y,n}$ of $\hat{X}_n$ given that the observations $\tilde{Y} = (\tilde{Y}_1, \ldots, \tilde{Y}_n)$ are fixed at $y = (y_1, \ldots, y_n)$.

Since $\hat{X}_n$ is valued in the finite grid Γ_n consisting of N_n points x_n^i, $i = 1, \ldots, N_n$, the discrete probability measure $\hat{\Pi}_{y,n}$ is characterized by its weights $\hat{\Pi}_{y,n}^i = \mathbb{P}[\hat{X}_n = x_n^i | \tilde{Y} = y]$, $i = 1, \ldots, N_n$: for any bounded measurable function f on $\mathbb{R}^d$, we have

$$
\hat{\Pi}_{y,n} f = \sum_{i=1}^{N_n} f(x_n^i) \hat{\Pi}_{y,n}^i.
$$

In other words, $\hat{\Pi}_{y,n} = \sum_{i=1}^{N_n} \hat{\Pi}_{y,n}^i \delta_{x_n^i}$, where δ_x is the Dirac mass at x. By same arguments as in Section 2, using the Bayes rule and Markov property of $(\hat{X}_k, \tilde{Y}_k)_k$, we have

$$
\hat{\Pi}_{y,n}^i = \frac{\hat{\pi}_{y,n}^i}{\sum_{i=1}^{N_n} \hat{\pi}_{y,n}^i}, \qquad i = 1, \ldots, N_n, \tag{4.3}
$$

where

$$
\hat{\Pi}_{y,n}^i = \mathbb{E}\Big[\mathbf{1}_{\hat{X}_n = x_n^i} \hat{L}_{y,n}\Big], \qquad i = 1, \ldots, N_n, \tag{4.4}
$$

in which

$$\hat{L}_{y,n} = \prod_{k=1}^{n} g_k(\hat{X}_{k-1}, y_{k-1}, \hat{X}_k, y_k). \tag{4.5}$$

From an algorithmic viewpoint, the unnormalized approximate filter $\hat{\pi}_{y,n}$ may be computed either in a forward or backward induction in view of (2.1) or (2.2). We describe here the forward procedure which is less costly in terms of complexity. We denote by $\hat{P}_0 = (\hat{P}_0^i)_{i=1,\dots,N_0}$ the probability law of $\hat{X}_0$, that is, $\hat{P}_0^i = \mathbb{P}[\hat{X}_0 = x_0^i]$, $i = 1, \dots, N_0$, and by $(\hat{P}_k)_k$, $k = 1, \dots, n$, the transition probability matrix of the finite state space Markovian process $(\hat{X}_k)_k$, that is,

$$\hat{P}_k^{ij} = \mathbb{P}[\hat{X}_k = x_k^j | \hat{X}_{k-1} = x_{k-1}^i], \qquad i = 1, \dots, N_{k-1}, j = 1, \dots, N_k. \tag{4.6}$$

We introduce the transition matrix $\hat{H}_{y,k}$ given by

$$\hat{H}_{y,k}^{ij} = g_{y,k}(x_{k-1}^i, x_k^j)\hat{P}_k^{ij}, \qquad i = 1, \dots, N_{k-1}, j = 1, \dots, N_k, \tag{4.7}$$

for $k = 1, \dots, n$. We then compute explicitly $\hat{\pi}_{y,n} = \sum_{i=1}^{N_n} \hat{\pi}_{y,n}^i \delta_{x_n^i}$ by the following forward algorithm:

$$\hat{\pi}_{y,0} = \hat{P}_0,$$

$$\hat{\pi}_{y,k}^j = \sum_{i=1}^{N_{k-1}} \hat{H}_{y,k}^{ij} \hat{\pi}_{y,k-1}^i, \qquad j = 1, \dots, N_k, \qquad k = 1, \dots, n. \tag{4.8}$$

4.2. Error analysis

In this subsection, we estimate the quality of the approximate filter $\hat{\Pi}_{y,n}$ in terms of the Markovian quantization errors on the signal $\|\Delta_k\|_1$, $k = 0, \dots, n$, defined by

$$\Delta_k = F_k(\hat{X}_{k-1}, \varepsilon_k) - \mathrm{Proj}_{\Gamma_k}\big(F_k(\hat{X}_{k-1}, \varepsilon_k)\big), \qquad k \geqslant 1 \tag{4.9}$$

$$\Delta_0 = X_0 - \mathrm{Proj}_{\Gamma_0}(X_0) \tag{4.10}$$

We make the following Lipschitz assumptions on the model (1.2)–(1.3):

(A1′) For each $k = 1, \dots, n$, there exists a positive constant $[F_k]_{\mathrm{Lip}}$ such that

$$\forall x, \hat{x} \in \mathbb{R}^d, \qquad \|F_k(x, \varepsilon_k) - F_k(\hat{x}, \varepsilon_k)\|_1 \leqslant [F_k]_{\mathrm{Lip}}|x - \hat{x}|.$$

We then set $[F]_{\mathrm{Lip}} = \max_{k=1,\dots,n}[F_k]_{\mathrm{Lip}}$. Note that $[P_k]_{\mathrm{Lip}} \leqslant [F_k]_{\mathrm{Lip}}$. In fact it may happen for some models that

$$[P_k]_{\mathrm{Lip}} < \infty = [F_k]_{\mathrm{Lip}}$$

which means that the field of application of the marginal quantization is wider than that of Markovian quantization, at least in theory. For example, set

$$X_{k+1} = \mathrm{sign}(X_k - \varepsilon_{k+1})G(X_k, \varepsilon_{k+1}),$$

where $(\varepsilon_k)_k$ is an i.i.d. sequence, $P_{\varepsilon_1}(\mathrm{d}u) = g(u)\lambda_q(\mathrm{d}u)$ (λ_q being Lebesgue measure on $\mathbb{R}^q$) and $(x, u) \mapsto G(x, u)$ is Lipschitz continuous in x uniformly in u with ratio $[G]_{\mathrm{Lip}}$. Then, it can easily be shown that

$$[P]_{\mathrm{Lip}} \leqslant [G]_{\mathrm{Lip}} < +\infty$$

whereas $x \mapsto F(x, \varepsilon_1) = \mathrm{sign}(x - \varepsilon_1)G(x, \varepsilon_1)$ is not even continuous so that (A1′) is not satisfied in general (e.g. when ε_1 has atoms).

We also rely on assumption (A2) introduced in Section 3 for the marginal quantization.

Remark 4.1. Beyond natural discrete-time dynamics, we notice that assumption (A1′) is satisfied by a Gaussian diffusion discretization scheme such as the Euler scheme. On the other hand, assumption (A2) often needs to be slightly strengthened to encompass diffusion discretization schemes. This is the aim of Remarks 4.3 and 3.4 which provide a setting often fulfilled by the Euler scheme of non-degenerate diffusions.

Theorem 4.1. *Assume that* (A1′) *and* (A2) *hold. Then, for every bounded Lipschitz continuous function* $f : \mathbb{R}^d \to \mathbb{R}$ *and any sequence of observed values* $y = (y_1, \ldots, y_n) \in (\mathbb{R}^q)^n$, *we have the a posteriori estimator*

$$\left|\Pi_{y,n}f - \hat{\Pi}_{y,n}f\right| \leqslant \frac{K_g^n}{\phi_n(y) \vee \hat{\phi}_n(y)} \sum_{k=0}^{n} A_k^n(f, y)\|\Delta_k\|_1. \tag{4.11}$$

with

$$\hat{\phi}_n(y) := \hat{\pi}_{y;n}\mathbf{1}, \tag{4.12}$$

$$A_k^n(f, y) = [F]_{\mathrm{Lip}}^{n-k}[f]_{\mathrm{Lip}} + 2\frac{\|f\|_\infty}{K_g}[g_k^2]_{\mathrm{Lip}}(y_{k-1}, y_k) \tag{4.13}$$
$$+ 2\frac{\|f\|_\infty}{K_g}\sum_{j=k+1}^{n}[F]_{\mathrm{Lip}}^{j-k-1}\Big([g_j^1]_{\mathrm{Lip}}(y_{j-1}, y_j) + [F]_{\mathrm{Lip}}[g_j^2]_{\mathrm{Lip}}(y_{j-1}, y_j)\Big),$$

(with the convention that the sum in (4.13) *is zero for* $k = n$*).*

Remark 4.2. By introducing

$$[g]_{\mathrm{Lip}} := \max_{k=1,\ldots,n}\ \sup_{y,y'\in\mathbb{R}^q}([g_k^1]_{\mathrm{Lip}}(y, y') \vee [g_k^2]_{\mathrm{Lip}}(y, y')),$$

we obtain that $A_k^n(f, y)$ is bounded by the simpler quantity

$$\tilde{A}_k^n(f) = [F]_{\mathrm{Lip}}^{n-k}[f]_{\mathrm{Lip}} + \frac{2\|f\|_\infty}{K_g}[g]_{\mathrm{Lip}}\left(\frac{[F]_{\mathrm{Lip}} + 1}{[F]_{\mathrm{Lip}} - 1}([F]_{\mathrm{Lip}}^{n-k} - 1) + 1\right)$$

with the usual convention that

$$\frac{1}{u-1}(u^m - 1) = m \qquad \text{if } u = 1 \text{ and } m \in \mathbb{N}.$$

Remark 4.3. Suppose that the Lipschitz condition (A2)(ii) is weakened into (A2)(ii′) as in Section 3 and that (A1) is strengthened into the following slightly more stringent condition than (A1′):

(A1$'_p$) For each $k = 1, \ldots, n$, there exists a positive constant $[F_k]_{\mathrm{Lip}}$ such that

$$\|F_k(x, \varepsilon_k) - F_k(\hat{x}, \varepsilon_k)\|_p \leqslant [F_k]_{\mathrm{Lip}}|x - \hat{x}|,$$

for all $p \in (1, \infty)$ and $x, \hat{x} \in \mathbb{R}^d$.

Then we may state a similar estimate for the approximate filter as in Theorem 4.1: for each $p \in (1, \infty)$ (and $1/p + 1/q = 1$),

$$\left|\Pi_{y,n}f - \hat{\Pi}_{y,n}f\right| \leqslant \frac{K_g^n}{\phi_n(y) \vee \hat{\phi}_n(y)} \sum_{k=0}^{n} \bar{A}_k^n(q, f)\|\Delta_k\|_p,$$

with

$$\bar{A}_k^n(q, f) = [F]_{\mathrm{Lip}}^{n-k}[f]_{\mathrm{Lip}} + \frac{2\|f\|_\infty}{K_g}[g]_{\mathrm{Lip}}\left(\frac{[F]_{\mathrm{Lip}} + 1}{[F]_{\mathrm{Lip}} - 1}([F]_{\mathrm{Lip}}^{n-k} - 1) + 1\right)N_n(q),$$

$$N_n(q) = 1 + 4\|X\|_q + (1 + [F]_{\mathrm{Lip}})([F]_{\mathrm{Lip}} \vee 1)^{n-1} \sum_{l=0}^{n} \|\Delta_l\|_q.$$

Here we have set

$$[g]_{\mathrm{Lip}} := \max_{k=0,\ldots,n} \sup_{y,y' \in \mathbb{R}^q} ([g_k^1]_{\mathrm{Liploc}}(y, y') \vee [g_k^2]_{\mathrm{Liploc}}(y, y'))$$

and $\|X\|_q = \max_{k=0,\ldots,n}\|X_k\|_q$.

In order to prove Theorem 4.1, we need the following two lemmas.

Lemma 4.1. *Assume that* (A2) *holds. Then, for all* $y_1, \ldots, y_n \in \mathbb{R}^q$*, we have*

$$|L_{y,n} - \hat{L}_{y,n}| \leqslant K_g^{n-1} \sum_{k=1}^{n} [g_k^1]_{\mathrm{Lip}}(y_{k-1}, y_k)|X_{k-1} - \hat{X}_{k-1}| + [g_k^2]_{\mathrm{Lip}}(y_{k-1}, y_k)|X_k - \hat{X}_k|.$$

Proof. For notational convenience, we omit the dependence of L_n and $\hat{L}_n$ on $y_1, \ldots, y_n$. From (2.4) and (4.5), we have, for all $k = 1, \ldots, n$,

$$L_k - \hat{L}_k = \big(g_k(X_{k-1}, y_{k-1}, X_k, y_k) - g_k(\hat{X}_{k-1}, y_{k-1}, \hat{X}_k, y_k)\big)L_{k-1}$$
$$+ g_k(\hat{X}_{k-1}, y_{k-1}, \hat{X}_k, y_k)(L_{k-1} - \hat{L}_{k-1}).$$

From the boundedness condition (A2)(i) on g_k, we have $L_{k-1} \leqslant K_g^{k-1}$. Hence, by Assumption (A2)(ii), we obtain

$$|L_k - \hat{L}_k| \leqslant K_g^{k-1}\big([g_k^1]_{\mathrm{Lip}}(y_{k-1}, y_k)|X_{k-1} - \hat{X}_{k-1}| + [g_k^2]_{\mathrm{Lip}}(y_{k-1}, y_k)|X_k - \hat{X}_k|\big)$$
$$+ K_g|L_{k-1} - \hat{L}_{k-1}|.$$

Noting that $L_0 = \hat{L}_0 = 1$, we obtain the required result by induction. □

Lemma 4.2. *Assume that* (A1′) *holds. Then, for each* $k = 0, \ldots, n$, *we have*

$$\|X_k - \hat{X}_k\|_1 \leqslant \sum_{j=0}^{k} [F]_{\mathrm{Lip}}^{k-j} |\Delta_j|_1.$$

Proof. From the definitions (1.2) and (4.1) of X_k and $\hat{X}_k$, and (4.9) of Δ_k, we obviously obtain, for each $k \geqslant 1$,

$$\|X_k - \hat{X}_k\|_1 \leqslant \|F_k(X_{k-1}, \varepsilon_k) - F_k(\hat{X}_{k-1}, \varepsilon_k)\|_1 + \|\Delta_k\|_1.$$

By assumption (A1′) and since ε_k is independent of X_{k-1} and $\hat{X}_{k-1}$, we then obtain

$$\|X_k - \hat{X}_k\|_1 \leqslant [F_k]_{\mathrm{Lip}} \|X_{k-1} - \hat{X}_{k-1}\|_1 + \|\Delta_k\|_1.$$

Recalling that $\|X_0 - \hat{X}_0\|_1 = \|\Delta_0\|_1$, we conclude by backward induction. □

Proof of Theorem 4.1. From expressions (2.3) and (4.4), we derive that

$$\begin{aligned} |\pi_{y,n} f - \hat{\pi}_{y,n} f| &= |\mathbb{E}[f(X_n) L_{y,n}] - \mathbb{E}[f(\hat{X}_n)\hat{L}_{y,n}]| \\ &\leqslant \|f\|_\infty \mathbb{E}|L_{y,n} - \hat{L}_{y,n}| + [f]_{\mathrm{lip}} \mathbb{E}[|X_n - \hat{X}_n| \hat{L}_{y,n}] \\ &\leqslant \|f\|_\infty \mathbb{E}|L_{y,n} - \hat{L}_{y,n}| + [f]_{\mathrm{lip}} K_g^n \|X_n - \hat{X}_n\|_1. \end{aligned}$$

Lemmas 3.1, 4.1 and 4.2 complete the proof. □

5. Convergence of the quantized filters

In both the marginal and Markovian approaches the error analysis leads to a priori error bounds (4.11) and (3.7) with the same structure, from which one derives a slightly looser upper bound given by

$$|\Pi_{y,n} f - \hat{\Pi}_{y,n} f| \leqslant \|f\|_\infty \vee [f]_{\mathrm{Lip}} \frac{K_g^n}{\phi_n(y)} \sum_{k=0}^{n} D_k^n(y, p) \|\Delta_k\|_p,$$

for all $p \in [1, \infty)$, where $\Delta_k = Z_k - \mathrm{Proj}_{\Gamma_k}(Z_k)$ is the difference between a simulated random variable Z_k and its projection by the nearest-neighbour rule onto the grid Γ_k with size $|\Gamma_k| = N_k$ (in the marginal quantization method $Z_k = X_k$, in the Markovian quantization approach $Z_k = F(\hat{X}_{k-1}, \varepsilon_k)$). Here

$$D_k^n(y, p) = \begin{cases} B_k^n(f_0, y, p), & \text{for the marginal quantization,} \\ A_k^n(f_0, y), & \text{for the Markovian quantization } (p = 1), \end{cases}$$

where $f_0(x) = |x|/(1 + |x|)$ (so that $\|f_0\|_\infty = [f_0]_{\mathrm{Lip}} = 1$). If one assumes that, at every time $k = 0, 1, \ldots, n$, the grid Γ_k is L^p optimal, that is, minimizes the mean L^p quantization error among all grids with size N_k, then the asymptotic behaviour of the optimal quantization stated in Zador's theorem (see Theorem B.1 in Appendix B) implies that, for every $k = 0, \ldots, n$, there is a positive real constant $\theta_k := \theta(p, \mathbb{P}_{Z_k}, d)$ such that

$$\|\Delta_k\|_p \leqslant \theta_k N_k^{-1/d},$$

so that

$$|\Pi_{y,n}f - \hat{\Pi}_{y,n}f| \leqslant \|f\|_\infty \vee [f]_{\mathrm{Lip}} \frac{K_g^n}{\hat{\phi}_n(y)} \sum_{k=0}^{n} \theta_k\, D_k^n(y, p) N_k^{-1/d}. \tag{5.1}$$

Theorem 5.1. *Let $n \geqslant 1$. In both marginal and Markovian settings, the optimally quantized approximate filters converge toward the true filter as* $\min_{1 \leqslant k \leqslant n} N_k$ *goes to infinity.*

5.1. Application to optimal dispatching

If some numerical bounds $\bar{d}_k^n$ and $\bar{\theta}_k$ are available for $D_k^n(y, p)$ (locally) uniformly in the observations $y = (y_1, \ldots, y_n)$ and for θ_k (which requires some information on the density of Z_k), then one may easily solve numerically the optimal allocation problem

$$\min_{N_0+\ldots+N_n=N} \sum_{k=0}^{n} \bar{\theta}_k\, \bar{d}_k^n(y, p) N_k^{-1/d} \tag{5.2}$$

to optimally dispatch the N points among the $n+1$ time steps. For more details we refer to Bally and Pagès (2003) and Pagès *et al*. (2004a) in which this phase has been carried out in different settings.

In some situations, such as the marginal quantization of a stationary signal, it may happen that alternative approaches turn out to be more efficient: optimizing only one huge grid and its transition parameters and then replicating it at every time k produces better results.

5.2. Application to discretized diffusions

We now discuss the convergence of the quantized filter when n also goes to infinity. This asymptotic behaviour is relevant especially when the signal $(X_k)_{0 \leqslant k \leqslant n}$ is a time discretization with step $h = T/n$ of a continuous-time signal $(X_t)_{0 \leqslant t \leqslant T}$. For example, if X_t follows a diffusion process

$$\mathrm{d}X_t = b(X_t)\mathrm{d}t + \sigma(X_t)\mathrm{d}W_t,$$

with W a standard Brownian motion, one may discretize it by an Euler scheme

$$X_{k+1} = F(X_k, \varepsilon_{k+1}) := X_k + b(X_k)\frac{T}{n} + \sigma(X_k)\sqrt{\frac{T}{n}}\varepsilon_{k+1},$$

where $(\varepsilon_k)_k$ is a Gaussian white noise process. Standard computations show that when b and σ are Lipschitz, condition (A1′) is satisfied with

$$[F]_{\mathrm{Lip}} = 1 + \frac{c}{n},$$

for some positive constant independent of n. We can then easily see that $D_k^n(y, p)$, $k = 0, \ldots, n$, is bounded by a constant independent of k, n. Therefore if one simply assigns $N_k = \overline{N} := N/(n+1)$ points at each grid Γ_k, $k = 0, \ldots, n$, (5.1) provides a rate of convergence for the approximate filters of order

$$\frac{K_g^n}{\phi_n(y)} \frac{n+1}{\overline{N}^{1/d}}.$$

This has to be compared with the rate of convergence obtained by particle Monte Carlo methods using $\overline{N}$ interacting particles (see Del Moral *et al*. 2001):

$$\left(\frac{K_g^n}{\phi_n(y)}\right)^n \frac{1}{\overline{N}^{1/2}}.$$

6. On practical implementation

6.1. General features and complexity

At this stage, it is important to mention when and how a quantization method can be implemented. First, one must bear in mind that it is an off-line method: a significant part of the computations can be carried out and kept off-line. In fact, things need to be done that way to make the method fully competitive.

A natural framework for implementing the quantization approach is to assume that the probabilistic features of the state process (X_k) do not change too fast or too often. This is not a real restriction in typical applications such as sea surge prediction, satellite tracking, and financial modelling based on stochastic volatility.

Moreover, the more functions f one needs to estimate for a given set of observations, the more efficient the method becomes. This can be easily understood when one describes in more detail the three phases of the filter approximation.

1. *Off-line optimization*. This phase is devoted to the construction of the weighted optimal *quantization tree* of the Markov process (X_k), given that it contains a total of N points. This means

- specifying the sizes N_k of the grids Γ_k, $1 \leqslant k \leqslant n$ (a priori dispatching);
- optimizing every grid $\Gamma_k := \{x_k^i, 1 \leqslant i \leqslant N_k\}$, $1 \leqslant k \leqslant n$, that is, solving the optimization problem

$$\min_{|\Gamma_K| \leqslant N_k} \|X_k - \mathrm{Proj}_{\Gamma_k}(X_k)\|_2;$$

- computing the transition weights $\hat{P}_k^{ij} = \mathbb{P}(\hat{X}_k = x^j \mid \hat{X}_{k-1} = x_{k-1}^i)$, $1 \leqslant i \leqslant N_k$, $1 \leqslant j \leqslant N_{k+1}$, $k = 0, 1, \ldots n$.

The dispatching is done a priori, based either on the minimization of the theoretical error bounds (see Bally and Pagès 2003; or Pagès *et al.* 2004b) by solving (5.2) or on more specific features of the state process (see the stationary case below). The optimization of the grids results from a stochastic gradient descent called *competitive learning vector quantization* based on Monte Carlo simulations of the (X_k). The computation of the transition weights and of the quantization error is carried out either simultaneously or by a new Monte Carlo simulation (see Bally and Pagès 2003). This has been extensively investigated in earlier papers, to which we refer for a precise description of the procedure (see Bally and Pagès 2003; Pagès *et al.* 2004b). This optimization phase is computationally the most demanding, requiring nearly 10 minutes of CPU time to compute the whole quantization tree ('height' $n = 20$, size $N = 20\,000$) of an (asymptotically) non-stationary four-dimensional process using a 1 GHz microprocessor.

However, in many cases this phase can be significantly shortened: when (X_k) is a stationary process only one grid is necessary (see Section 6.2), which drastically reduces the procedure by a factor n. Furthermore, if X is a Gaussian process, a library of optimal grids with various sizes is now available for the d-dimensional normal distributions $\mathcal{N}(0; I_d)$ (see Pagès and Printems 2003; the files are available at www.proba.jussieu.fr/pageperso/pages.html or www.univ-paris12.fr/www/labos/cmup/homepages/printems). The crucial fact is that the optimization phase does not depend on the observations, which explains why its results can be kept off-line.

2. *Computation of the quantized filter distribution*. One computes the weight vector $(\hat{\pi}_{y,n}^j)_{1 \leqslant j \leqslant N_n}$ by plugging the observation vector $y = (y_1, \ldots, y_n)$ and the transition weights $\hat{P}_k^{ij}$ into the *forward* representations of the approximate filter, that is, (3.5) or (4.8).

The theoretical complexity of the quantization tree descent is $\sum_{k=0}^{n-1} N_k N_{k+1}$, which is at least $nN^2/(n+1)^2 \approx N^2/n$ (when $N_k = N/(n+1)$, $k = 0, \ldots, n$). In practice, many transitions are 0 (when x_{k-1}^i and x_k^j are remote) and every node x_k^i of the tree has approximately the same number ν of 'active connections' with nodes at time $k+1$. The resulting complexity, after an appropriate pruning of the quantization tree, is thus approximately $\nu \times n \times \overline{N}$, where $\overline{N} = N/(n+1)$ is the average number of points per time step. In all our numerical experiments, this phase is almost instantaneous (less than 0.1 second with $N = 20\,000$ points in dimension $d = 4$ with the same 1 GHz microprocessor). This distribution approximation phase does not depend on the function f.

3. *Computation of* $\hat{\Pi}_{y,n} f$. One computes for every (required) function f

$$\int f(x) \hat{\Pi}_{y,n}(\mathrm{d}x) = \sum_{i=1}^{N_n} f(x_n^i) \hat{\Pi}_n^i.$$

The complexity of this phase is proportional to N_n and is negligible (although dependent on f).

In particle methods, at every time step, one needs to simulate $\overline{N}$ new particles following a (weighted) empirical measure (with a support of size $\overline{N}$). This requires one first to compute the weights of the empirical measure, secondly to generate $\overline{N}$ random numbers and then to simulate by an inverse distribution function method $\overline{N}$ appropriately distributed numbers. The average complexity of the last phase cannot be lower than $O(\overline{N}\log(\overline{N}))$ comparisons (see Devroye 1986) – indeed, a naive approach yields the almost surely worst possible complexity, which is $O(\overline{N}^2)$ – which makes an average total of $(n+1)\overline{N}\log(\overline{N})$ comparisons and $O(\overline{N})$ multiplications.

It may happen for some observation vectors that the optimal filter and the prior distribution of the process X assign some masses to significantly different areas of the space, making the algorithm less efficient. One way to prevent this problem is to quantize the observation process to evaluate the likelihood of an observation vector (see Sellami 2004).

6.2. A special case: marginal quantization of a stationary signal

In the case where the signal $(X_k)_{0\leqslant k\leqslant n}$ is a stationary Markov chain with distribution ν, the optimal L^p quantization of the whole chain clearly amounts to that of its stationary distribution ν. Let $\hat{\Gamma} := \{\hat{x}^1, \ldots, \hat{x}^{\overline{N}}\}$ be an $\overline{N} := N/(n+1)$-optimal grid, that is, such that

$$\|X - \mathrm{Proj}_{\hat{\Gamma}}(X)\|_p = \min_{\Gamma\subset\mathbb{R}^d,\, |\Gamma|\leqslant\overline{N}} \|X - \mathrm{Proj}_{\Gamma}(X)\|_p.$$

Then the $\hat{\Gamma}_k := \hat{\Gamma}$, $0 \leqslant k \leqslant n$, make up the optimal quantization of the chain. The companion parameters are the quantization of the distribution ν induced by $\hat{\Gamma}$, that is, $\hat{P}_0 = \mathrm{Proj}_{\hat{\Gamma}}(X_0)$, and a single transition matrix

$$\hat{P}_k^{ij} = \hat{P}_1^{ij} := \mathbb{P}(\hat{X}_1 = \hat{x}^j \,|\, \hat{X}_0 = \hat{x}^i), \qquad 0 \leqslant i, j \leqslant \overline{N}.$$

The size of the parameters to be stored is obviously divided by a factor n (or the possible quantization size for the distribution ν and the transition matrix is multiplied by n). This ability to take into account the stationarity of the signal process is an interesting feature of the optimal quantization approach which seems not to be shared by other numerical methods.

7. Discretely observed diffusions

In this section, we discuss how our previous results can be applied when the signal-observation process evolves according to a stochastic differential equation of the form

$$\mathrm{d}X_t = b(X_t)\mathrm{d}t + \sigma(X_t)\mathrm{d}W_t, \qquad X_0 \rightsquigarrow \mu, \tag{7.1}$$

$$\mathrm{d}Y_t = \beta(X_t, Y_t)\mathrm{d}t + \gamma(X_t, Y_t)\mathrm{d}B_t, \qquad Y_0 = 0, \tag{7.2}$$

where W is a d-dimensional Brownian motion independent of the q-dimensional Brownian motion B, μ is a known distribution, and b, β, σ, γ are known functions. We set $\sum(x) = \sigma(x)\sigma(x)^{\mathrm{T}}$ and $\Lambda(x, y) = \gamma(x, y)\gamma(x, y)^{\mathrm{T}}$. We assume that $x \mapsto \sum(x)$ and $(x, y) \mapsto \Lambda(x, y)$ are uniformly non-degenerate functions and we denote by $x \mapsto \sum^{1/2}(x)$

and $(x, y) \mapsto \Lambda^{1/2}(x, y)$ their square roots functions[1] which are clearly uniformly non-degenerate. The process (X, Y) is Markov with a transition semigroup denoted by $(R_t)_t$.

We suppose here that the sample path (Y_t) is observed at n discrete times with regular sampling interval, say 1. Our aim is then to compute the filter $\Pi_{y,n}$ of X_n conditional on the observations $(Y_1, \ldots, Y_n)$ set at $y = (y_1, \ldots, y_n)$.

The sequences $(X_k, Y_k)_{k \in \mathbb{N}}$ and $(X_k)_{k \in \mathbb{N}}$ are (homogeneous) Markov chains with transitions $R_1(x, y, dx', dy')$ and $P(x, dx') = R_1(x, y, dx', \mathbb{R}^q)$. Under suitable conditions on the coefficients of the diffusion (7.1)–(7.2), for example if the functions b, σ, β, γ are twice differentiable with bounded derivatives of all orders up to 2, the transition $R_1(x, y, dx', dy')$ admits a density $(x', y') \mapsto r(x, y, x', y')$. Hence, we are in the situation of (H): the law of Y_k conditional on $(X_{k-1}, Y_{k-1}, X_k) = (x, y, x')$ admits a density $y' \mapsto g(x, y, x', y')$ given by

$$g(x, y, x', y') = \frac{r(x, y, x', y')}{p(x, x')}$$

where $p(x, x') = \int r(x, y, x', y')dy'$ is the density of the transition $P(x, dx')$.

But we do not know explicitly the density r (and so g) and we have to approximate it by an Euler scheme. We follow closely here the arguments of Del Moral *et al.* (2001). For a step size $1/m$, and given a starting point $(x, y) \in \mathbb{R}^d \times \mathbb{R}^q$, we define by induction the variables

$$\begin{aligned}
&\bar{X}(x)_0^{(m)} = x,\\
&\bar{X}(x)_{i+1}^{(m)} = \bar{X}(x)_i^{(m)} + b(\bar{X}(x)_i^{(m)})\frac{1}{m} + \sigma(\bar{X}(x)_i^{(m)})\frac{\varepsilon_{i+1}}{\sqrt{m}},\\
&\bar{Y}(x, y)_0^{(m)} = y,\\
&\bar{Y}(x, y)_{i+1}^{(m)} = \bar{Y}(x, y)_i^{(m)} + \beta(\bar{X}(x)_i^{(m)}, \bar{Y}(x, y)_i^{(m)})\frac{1}{m} + \gamma(\bar{X}(x)_i^{(m)}, \bar{Y}(x, y)_i^{(m)})\frac{\eta_{i+1}}{\sqrt{m}},
\end{aligned}$$

for $i = 0, \ldots, m-1$, where the $(\varepsilon_i)_i$ and $(\eta_i)_i$ are independent sequences of i.i.d centred Gaussian vectors with unit covariance matrices. We denote by $R_1^{(m)}(x, y, dx', dy')$ the law of $(\bar{X}(x)_m^{(m)}, \bar{Y}(x, y)_m^{(m)})$. Then $R_1^{(m)}(x, y, dx', dy')$ has a density $(x', y') \to r^{(m)}(x, y, x', y')$ explicitly given by

$$r^{(m)}(x, y, x', y') = \int \prod_{i=0}^{m-1} \phi(x_i, x_{i+1})\psi(x_i, y_i, y_{i+1})dx_1 \ldots dx_{m-1}dy_1 \ldots dy_{m-1},$$

with $(x_0, y_0) = (x, y)$, $(x_m, y_m) = (x', y')$ and

[1] Every non-negative symmetric S matrix admits a unique square root $S^{1/2}$ which is non-negative, symmetric, satisfies $S^{1/2}S^{1/2} = S$ and commutes with S.

$$\phi(x, x') = \frac{m^{d/2}}{(2\pi)^{d/2}\det(\sum^{1/2}(x))} \exp\left[-\frac{m}{2}\left|(\sum^{1/2}(x))^{-1}\left(x' - x - \frac{b(x)}{m}\right)\right|^2\right],$$

$$\psi(x, y, y') = \frac{m^{q/2}}{(2\pi)^{q/2}\det(\Lambda^{1/2}(x, y))} \exp\left[-\frac{m}{2}\left|(\Lambda^{1/2}(x))^{-1}\left(y' - y\frac{\beta(x, y)}{m}\right)\right|^2\right].$$

The density $r^{(m)}(x, y, x', y')$ is an approximation of the density $r(x, y, x', y')$. More precisely, we have from (Bally and Talay 1996) the existence of constants C and C' depending only on the coefficients b, β, σ, γ such that

$$r(x, y, x', y') + r^{(m)}(x, y, x', y') \leqslant C\exp\left(-C'(|x - x'|^2 + |y - y'|^2)\right), \tag{7.3}$$

$$|x - x'| + |y - y'| > \frac{2}{m} \Rightarrow |r(x, y, x', y') - r^{(m)}(x, y, x', y')| \leqslant \frac{C}{m}\exp(-C(|x - x'|^2 + |y - y'|^2)). \tag{7.4}$$

The law $P^{(m)}(x, \mathrm{d}x')$ of $X(x)_m^{(m)}$ has a density $x' \to p^{(m)}(x, x') = \int r^{(m)}(x, y, x', y')\mathrm{d}y'$. We then have an approximation of $g(x, y, x', y')$ given by

$$g^{(m)}(x, y, x', y') = \frac{r^{(m)}(x, y, x', y')}{p^{(m)}(x, x')}. \tag{7.5}$$

We then approximate $\Pi_{y,n}$ by $\hat{\Pi}_{y,n}^{(m)}$ defined by the marginal quantization algorithm in Section 3, where we replace the unknown function g by $g^{(m)}$. The estimation error is measured via

$$|\Pi_{y,n}f - \hat{\Pi}_{y,n}^{(m)}f| \leqslant |\Pi_{y,n}f - \overline{\Pi}_{y,n}^{(m)}f| + |\overline{\Pi}_{y,n}^{(m)}f - \hat{\Pi}_{y,n}^{(m)}f|, \tag{7.6}$$

where $\overline{\Pi}_{y,n}^{(m)}$ is the filter given by formulae (2.1)–(2.2), with the transition probability distribution $P(x, \mathrm{d}x') = p(x, x')\mathrm{d}x'$ replaced by $P^{(m)}(x, \mathrm{d}x') = p^{(m)}(x, x')\mathrm{d}x'$ and the conditional density g replaced by $g^{(m)}$. Actually, from the preliminaries in Section 2, we recall that the true filter $\Pi_{y,n}$ is given in an inductive form by

$$\Pi_{y,0} = \mu,$$

$$\Pi_{y,k} = \frac{\Pi_{y,k-1}H_{y,k}f}{\Pi_{y,k-1}H_{y,k}1}, \qquad k = 1, \ldots, n,$$

with

$$H_{y,k}f(x) = \int f(x')g(x, y_{k-1}, x', y_k)P(x, \mathrm{d}x') = \int f(x')r(x, y_{k-1}, x', y_k)\mathrm{d}x',$$

while the approximate probability measure $\Pi_{y,n}^{(m)}$ is given by

$$\overline{\Pi}^{(m)}_{y,0} = \mu,$$

$$\overline{\Pi}^{(m)}_{y,k} = \frac{\overline{\Pi}^{(m)}_{y,k-1} H^{(m)}_{y,k} f}{\overline{\Pi}^{(m)}_{y,k-1} H^{(m)}_{y,k} 1}, \qquad k = 1, \ldots, n,$$

with

$$H^{(m)}_{y,k} f(x) = \int f(x') g^{(m)}(x, y_{k-1}, x', y_k) P^{(m)}(x, \mathrm{d}x') = \int f(x') r^{(m)}(x, y_{k-1}, x', y_k) \mathrm{d}x'.$$

By using (7.3)–(7.4), it can easily be checked that for any bounded function f,

$$\| H_{y,k} f - H^{(m)}_{y,k} f \|_\infty \leqslant \frac{C}{m} \| f \|_\infty,$$

for some positive constant C independent of m and y. Therefore, by Proposition 2.1 in Del Moral *et al*. (2001), the first term in (7.6) is estimated by

$$\left| \Pi_{y,n} f - \overline{\Pi}^{(m)}_{y,n} f \right| \leqslant \frac{C}{m} \| f \|_\infty \frac{\rho_n(y)^{n+1} - \rho_n(y)}{K_g(\rho_n(y) - 1)},$$

with $\rho_n(y) = 2K_g^n / \phi_n(y)$. The second term in (7.6) is given by Theorem 3.1 provided one can check some Lipschitz condition for $g^{(m)}$. Actually, it is proved in Appendix A that when the functions b, σ and γ are constant, there exists a positive constant C (independent of m) such that, for all $x, x', \hat{x}, \hat{x}' \in \mathbb{R}^d$ and $y, y' \in \mathbb{R}^q$,

$$|g^{(m)}(x, y, x', y') - g^{(m)}(\hat{x}, y, \hat{x}', y')| \leqslant Cm^{(q+1)/2}(1 + |x| + |x'| + |\hat{x}| + |\hat{x}'|)|x - \hat{x}| + Cm^{(q+3)/2}(1 + |x| + |x'| + |\hat{x}| + |\hat{x}'|)|x' - \hat{x}'|. \quad (7.7)$$

8. Numerical illustrations

Two numerical illustrations are presented: one with the Kalman–Bucy model derived from the noisy observation of the discretization of an Ornstein–Uhlenbeck model, the other with a stochastic volatility model arising in financial time series.

8.1. The Kalman–Bucy model

The Kalman–Bucy model is given by

$$X_k = AX_{k-1} + \tau \varepsilon_k \in \mathbb{R}^d, \quad (8.8)$$

$$Y_k = BX_k + \theta \eta_k \in \mathbb{R}^q, \quad (8.9)$$

for $k \in \mathbb{N}$, with X_0 normally distributed with mean $m_0 = 0$ and covariance matrix $\sum_0^2$. Here A, B, τ and θ are matrices of appropriate dimensions, and $(\varepsilon_k)_{k \geqslant 1}$, $(\eta_k)_{k \geqslant 1}$ are independent centred Gaussian processes, $\varepsilon_k \rightsquigarrow \mathcal{N}(0, I_d)$, $\eta_k \rightsquigarrow \mathcal{N}(0, I_q)$. In this case, we have, assuming that θ is invertible,

$$g_k(x, y) = g(x, y) = \frac{1}{(2\pi)^{d/2}\sqrt{\det(\theta\theta^{\mathrm{T}})}} \exp\left(-\frac{1}{2}\left|\theta^{-1}(y - Bx)\right|^2\right).$$

If $|||A||| < 1$, then $(X_k)_{k\geqslant 0}$ is stationary if and only if $\sum_0^2 = \sum_{n\geqslant 0}(A^n\tau)(A^n\tau)^{\mathrm{T}}$ (unique solution of $\sum_0^2 = A\sum_0^2 A^{\mathrm{T}} + \tau\tau^{\mathrm{T}}$). Of course, the filter $\Pi_{y,n}$ is explicitly known (see Elliot *et al*. 1995): it is a Gaussian distribution of mean m_n and covariance matrix C_n given by the inductive equations

$$C_{k+1} = (I_d - K_{k+1}B)(\tau\tau^{\mathrm{T}} + A\,C_k\,A^{\mathrm{T}}), \qquad C_0 := 0,$$

$$m_{k+1} = A\,m_k + K_{k+1}(y_{k+1} - BA\,m_k), \qquad m_0 := 0,$$

where

$$K_{k+1} = (\tau\tau^{\mathrm{T}} + A\,C_k A^{\mathrm{T}})B^{\mathrm{T}}(B(\tau\tau^{\mathrm{T}} + A\,C_k A^{\mathrm{T}})B' + \theta\theta^{\mathrm{T}})^{-1}.$$

Note that m_n depends on the observation vector y, whereas C_n does not.

The above system can be seen as the Euler scheme with step Δt of the (linear) Gaussian diffusion system

$$\mathrm{d}X(t) = -\alpha X(t)\mathrm{d}t + \sigma_X\,\mathrm{d}W^X(t),$$

$$dY(t) = B\,\mathrm{d}X(t) + \sigma_{_Y}\,\mathrm{d}W^Y(t),$$

with $\langle W^x, W^Y\rangle \equiv 0$, if one sets $A = I_d - \Delta t\,\alpha$, $\tau = \sqrt{\Delta t}\sigma_X$, $\theta = \sqrt{\Delta t}\sigma_Y$.

A numerical experiment has been carried out as follows: the (stationary) process X is quantized by marginal quantization. However, we decided to use grids which are not optimal for the stationary distribution $\mathcal{N}(0, \sum_0^2)$. Instead, we selected some grids of the form

$$\Gamma_0 = \textstyle\sum_0 \Gamma^* := \{\sum_0 \xi,\ \xi \in \Gamma^*\}$$

where Γ^* is L^2-optimal for $\mathcal{N}(0, I_d)$. This induces slightly less accurate results but illustrates the robustness of the method and the desirability of keeping some tabulations off-line.

Two kinds of tests were carried out with the Kalman–Bucy filter in order to track the behaviour of the quantized filter as a function of N (or $\overline{N}$) and n. The choice of a stationary setting is motivated by the possibility of detecting more simply the dependency in these parameters. However, some simulations carried out using the same model, but starting at some deterministic value $X_0 = 0$, yield quite similar results for the appropriate architecture of the 'quantization tree' (see Sellami 2004). This tree was made up of (non-optimal) grids suitably scaled from optimal grids for the normal distribution.

TEST 1: *Convergence of the filter at a fixed instant n as a function of the size* $\overline{N} = N/(n+1)$ *of the grid* Γ_0. We set $\Delta t = 1/250$, $n = 15$, and considered three functions

$$f_1(x) = x^d, \qquad f_2(x) = |x|^2, \qquad f_3(x) = \mathrm{e}^{-|x^d|},$$

implemented in dimensions $d = 1$ and $d = 3$ (closed forms exist for $\hat{\Pi}_n f_i$, $i = 1, 2, 3$). In both dimensions the results are summarized in a diagram showing for every function f_i, $i = 1, 2, 3$, the graphs $\overline{N} \mapsto \hat{\Pi}_n(f_i)$ (or $\overline{N} \mapsto |\Pi_n(f_i) - \hat{\Pi}_n(f_i)|$) and $\log\overline{N} \mapsto \log|\Pi_n(f_i) - \hat{\Pi}_n(f_i)|$ (i.e. a log scale) with its least-squares regression line denoted by '$y = -ax + b$' (a and b appearing as numerical values). This means that

$$|\Pi_n(f_i) - \hat{\Pi}_n(f_i)| \approx \frac{\mathrm{e}^b}{\overline{N}^a}.$$

For $d = 1$, $\alpha = B = 1$, $\sigma^X = 0.5$ and $\sigma^Y = 1$, we have $A = 0.996$, $\tau = 0.0316$ (so that $\sum_0 = \tau/\sqrt{1 - A^2} \approx 0.354$) and $\theta = 0.0663$. The grid size $\overline{N}$ ranges over the interval [50, 400]. Figure 1 shows the results. Furthermore, in Figure 2, for every function f_i, $i = 1, 2, 3$, we have added a graph $\overline{N} \mapsto |\Pi_n(f_i) - \hat{\Pi}_n(f_i)|$ for three different observation vectors.

Turning to $d = 3$, let

$$\alpha := \begin{bmatrix} 1.4445 & 0.5556 & 0.7778 \\ 0.5556 & 0.9445 & 0.2222 \\ 0.7778 & 0.2222 & 1.6110 \end{bmatrix}$$

so that

$$A = \begin{bmatrix} 0.9942 & -0.00222 & -0.0031 \\ -0.0022 & 0.9962 & -0.0009 \\ -0.00311 & -0.0009 & 0.9936 \end{bmatrix}.$$

Set

$$\tau = \begin{bmatrix} 0.1079 & 0.0317 & 0.0444 \\ 0.0317 & 0.0793 & 0.0127 \\ 0.0444 & 0.0127 & 0.1173 \end{bmatrix}$$

so that

$$\sum\nolimits_0 = \begin{bmatrix} 1.0118 & 0.0900 & 0.1349 \\ 0.0900 & 0.9219 & 0.0449 \\ 0.1349 & 0.0449 & 1.0343 \end{bmatrix}.$$

Then set $B = I_3$, $\theta = 0.5\, I_3$. The grid size $\overline{N}$ ranges over the interval [50, 600]. Figure 3 shows the results.

TEST 2: *Stability of the filter for a fixed grid size $\overline{N} = \overline{N}_{\max}$ as n grows.* We now consider a model in $d = 2$ dimensions with

$$\alpha = \begin{bmatrix} 1.1625 & -0.8488 \\ -0.8488 & 1.5875 \end{bmatrix},$$

so that

$$A := \begin{bmatrix} 0.99535 & 0.00340 \\ 0.003340 & 0.99365 \end{bmatrix}, \qquad \tau := \begin{bmatrix} 0.08830 & -0.02419 \\ -0.02419 & 0.10041 \end{bmatrix},$$

$B = I_2$, $\theta := 0.5 I_2$, and

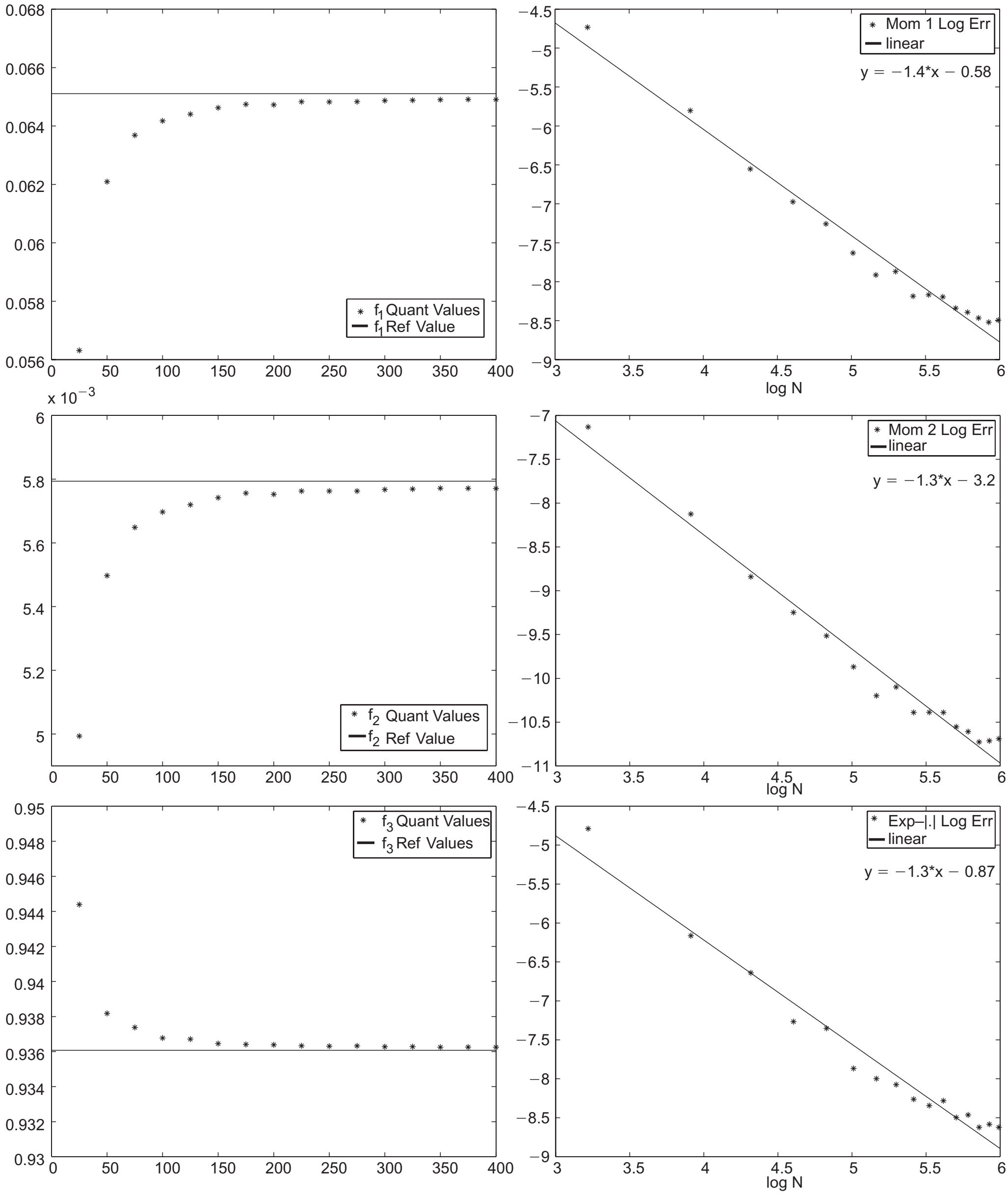

Figure 1. $d = 1$, $n = 15$. Convergence and convergence rate (on a log scale) as $\overline{N}$ grows.

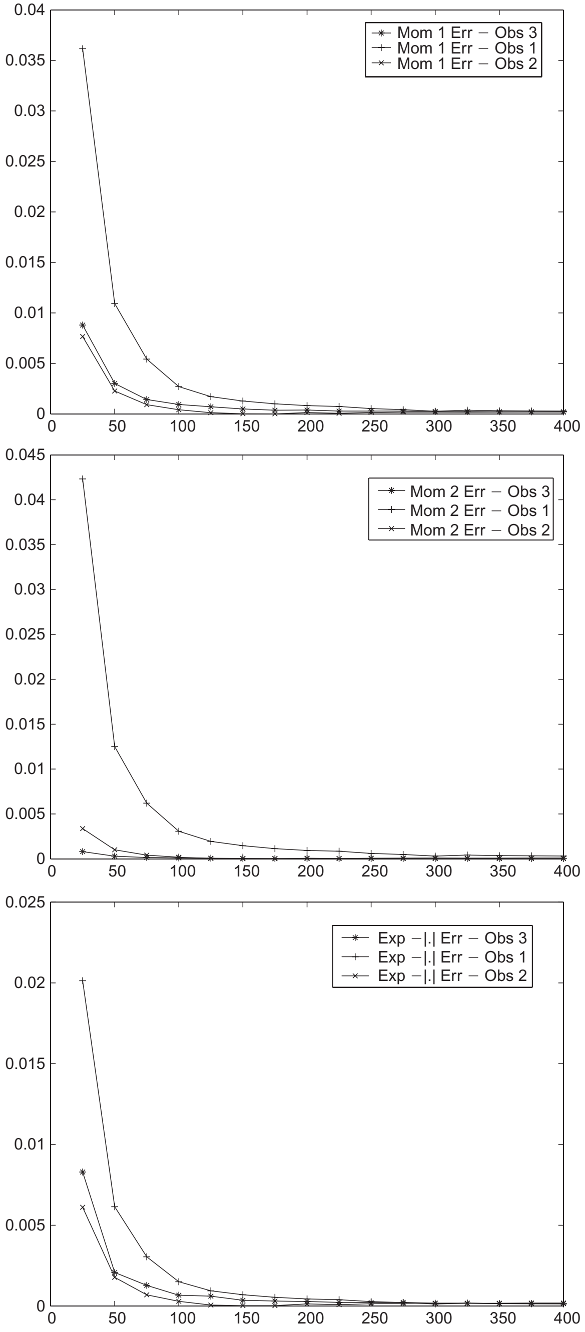

Figure 2. $d = 1$, $n = 15$. Errors and convergence rate as $\overline{N}$ grows for three different observation vectors.

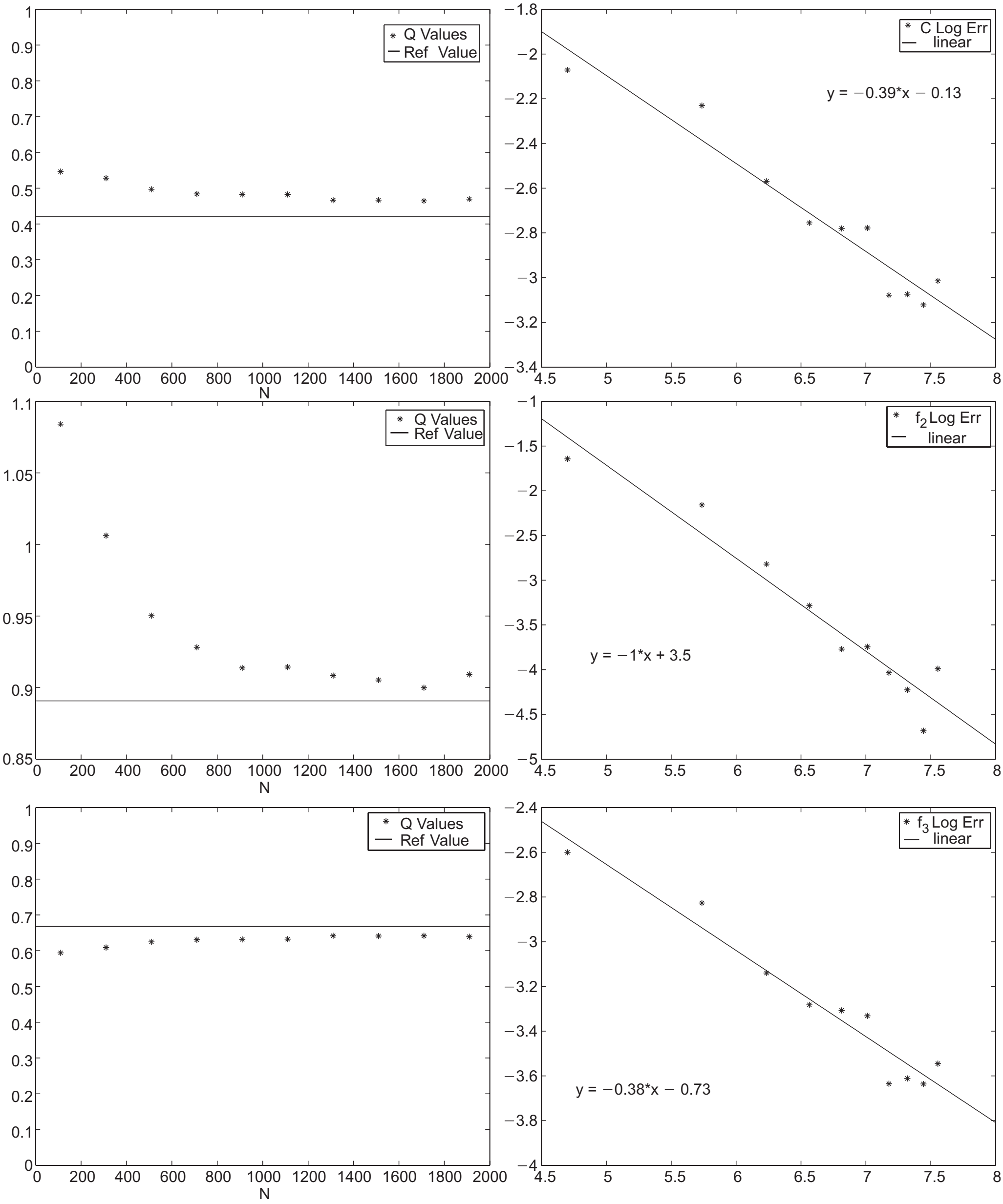

Figure 3. $d = 3$, $n = 15$. Convergence and convergence rate (on a log scale) as a function of $\overline{N}$.

$$\sum_0 = \begin{bmatrix} 1.01976 & 0.10041 \\ 0.10041 & 0.969493 \end{bmatrix}.$$

We set $\overline{N} = 600$ with n running from 1 to 100. On the left in Figure 4 are depicted

$$n \mapsto \Pi_n(x^i),\ n \mapsto \hat{\Pi}_n(x^i), \quad i = 1, 2 \qquad \text{and} \qquad n \mapsto \frac{\hat{\Pi}_n(|x|) - \Pi_n(|x|)}{\Pi_n(|x|)}, \quad n \in [1, 100],$$

and the linear regression line of these relative errors. These results are much more satisfactory than those induced by the a posteriori error bounds obtained in Theorem 3.1 or Theorem 4.1, although the process (X_k) is not 'rapidly mixing' since $|||A|||$ is close to 1 (this explains why the regression line is not completely flat). When $|||A|||$ is less than 0.8 as on the right in Figure 4, the true value and the quantized one become indistinguishable. This means that, as for interacting particle methods, the mixing property of the state variable X induces the stability of the filter as n increases. However, we do not yet have theoretical results to support this fact.

8.2. A stochastic volatility model

We consider a state model with multiplicative Gaussian noise process:

$$Y_k = \sigma(X_k)\eta_k \in \mathbb{R}, \qquad \text{with } X_k = \rho X_{k-1} + \varepsilon_k \in \mathbb{R}, \tag{8.10}$$

where ρ is a real constant, $\sigma(\cdot)$ is a positive Borel function on $\mathbb{R}$ and $(\varepsilon_k)_{k\geqslant 1}$, $(\eta_k)_{k\geqslant 1}$ are independent Gaussian processes. In terms of financial modelling, $(Y_k)_{k\geqslant 0}$ represents a (martingale) asset price model with stochastic volatility $\sigma(X_k)$. We still consider (8.10) as an Euler scheme, with step size Δt, of a continuous-time Ornstein–Uhlenbeck stochastic volatility model

$$\mathrm{d}X(t) = -\alpha X(t)\mathrm{d}t + \tau\,\mathrm{d}W(t), \qquad 0 \leqslant t \leqslant 1,$$

with positive parameters λ and τ. We then suppose that

$$\rho = 1 - \alpha\Delta t, \qquad \varepsilon_k \rightsquigarrow \mathcal{N}(0, \tau^2\Delta t), \qquad \eta_k \rightsquigarrow \mathcal{N}(0, \Delta t).$$

The filtering problem consists of estimating the volatility $\sigma(X_n)$ at step n given the observations of the prices $(Y_0, \ldots, Y_n)$. Here,

$$g_k(x, y) = g(x, y) = \frac{1}{\sqrt{2\pi}\Delta t\,\sigma(x)} \exp\left(-\frac{y^2}{2\,\sigma^2(x)\Delta t}\right).$$

The values of the parameters in our simulation are for $(\alpha, \tau, \Delta t) = (1, 0.5, 1/250)$. The Gaussian distribution of X_0 is such that the sequence $(X_k)_{k\geqslant 1}$ is stationary, that is, $X_0 \sim \mathcal{N}(0, \sum_0^2)$ with $\sum_0 = \tau\sqrt{\Delta t/(1-\rho^2)} = \tau/\sqrt{\alpha(2 - \alpha\Delta t)} \approx 0.354$.

The selected model here is

$$(ABS) \equiv \sigma(X_k) = \gamma + |X_k|, \qquad \text{with } \gamma = 0.05.$$

Figure 5 shows the graph $\overline{N} \mapsto |\Pi_n(f_i) - \hat{\Pi}_n(f_i)|$ which strongly suggests convergence for the three functions (although no reference value is available).

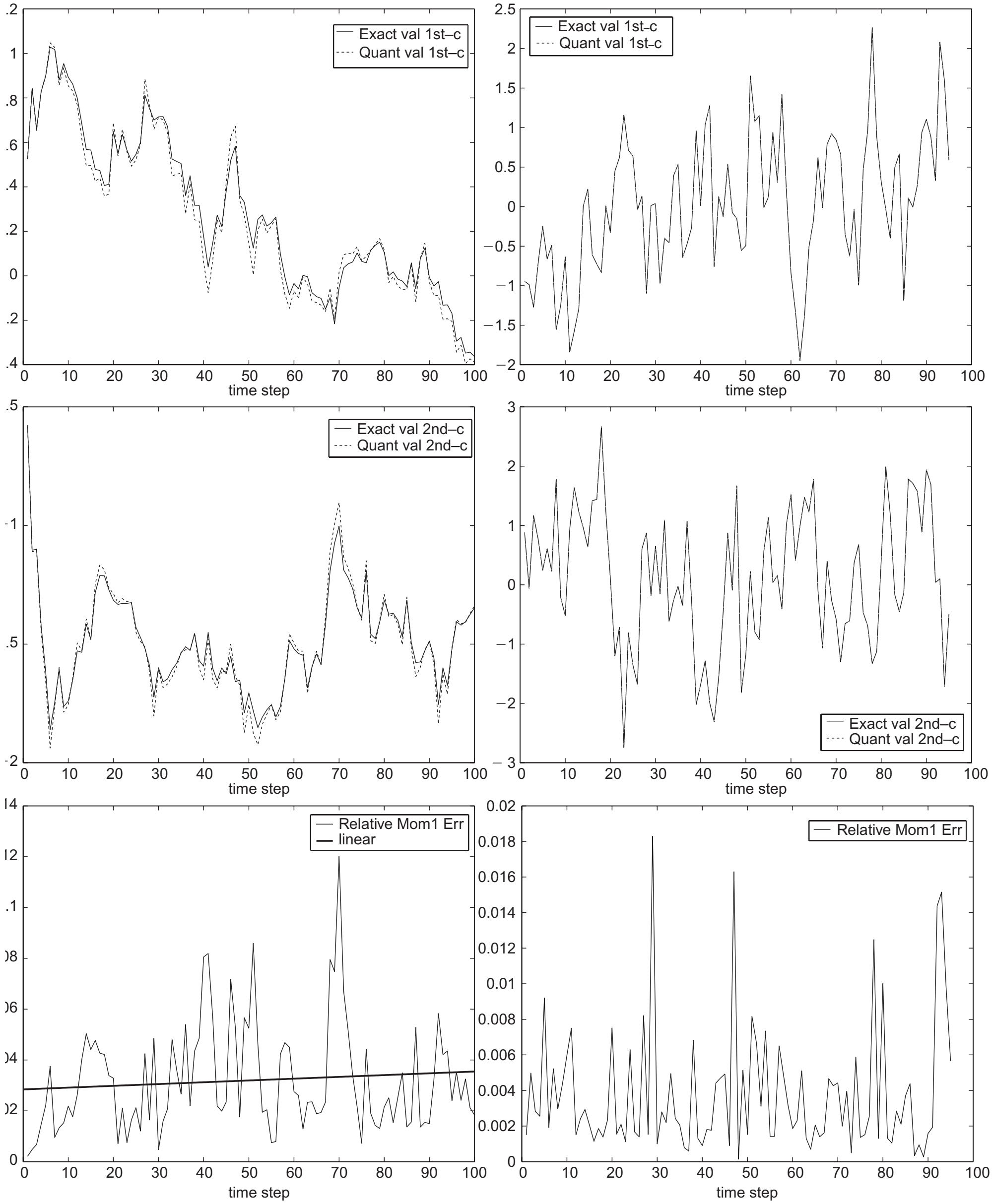

Figure 4. First moment error and relative error $\overline{N} = 600$, $1 \leqslant n \leqslant 100$, with $|||A|||$ close to 1 (left) and less than 0.8 (right).

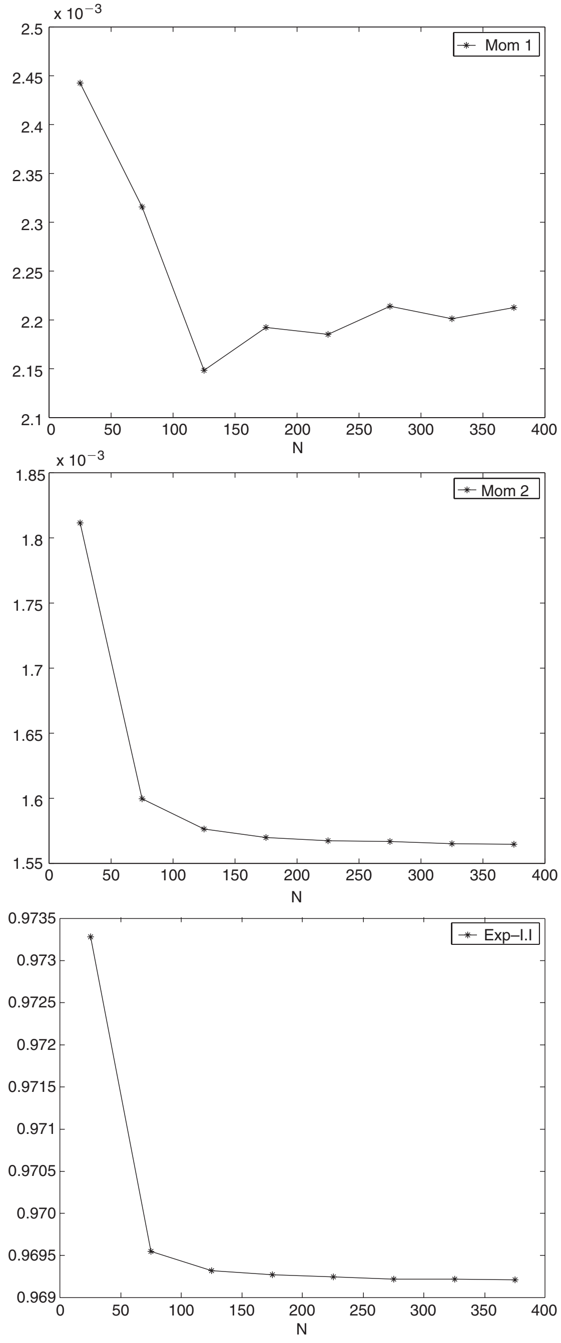

Figure 5. Stochastic volatility model: filter values for f_1, f_2 and f_3 functions and $(\rho, \theta, \gamma) = (0.996, 0.0316, 0.05)$ as $\overline{N}$ grows.

Appendix A: Lipschitz condition on the conditional density of the Euler scheme

We consider the particular case where the coefficients b, σ and γ of the diffusion (X, Y) in (7.1)–(7.2) are real constants. We then assume without loss of generality that $\sigma = I_d$ and $\gamma = I_q$. We also assume that the function β is bounded and differentiable with bounded derivatives. We show that then the conditional density $g^{(m)}$ of the Euler scheme is bounded and satisfies the locally Lipschitz continuous condition (7.7). First, we recall that

$$g^{(m)}(x_0, y_0, x_m, y_m) = \frac{r^{(m)}(x_0, y_0, x_m, y_m)}{p^{(m)}(x_0, x_m)} \tag{A.1}$$

with

$$r^{(m)}(x_0, y_0, x_m, y_m) = \int \prod_{i=0}^{m-1} \phi(x_i, x_{i+1}) \psi(x_i, y_i, y_{i+1})\, \mathrm{d}x_1 \ldots \mathrm{d}x_{m-1} \mathrm{d}y_1 \ldots \mathrm{d}y_{m-1},$$

$$p^{(m)}(x_0, x_m) = \int \prod_{i=0}^{m-1} \phi(x_i, x_{i+1}) \mathrm{d}x_1 \ldots \mathrm{d}x_{m-1},$$

where

$$\phi(x, x') = \frac{m^{d/2}}{(2\pi)^{d/2}} \exp\left[-\frac{m}{2}\left|x' - x\frac{b}{m}\right|^2\right], \qquad \psi(x, y, y') = \frac{m^{q/2}}{(2\pi)^{q/2}} \exp\left[-\frac{m}{2}\left|y' - y - \frac{\beta(x, y)}{m}\right|^2\right].$$

First, by noting that $\psi(x_{m-1}, y_{m-1}, y_m)$ is bounded by $(m/2\pi)^{q/2}$ and using the fact that, for every $x_0, \ldots, x_{m-1} \in \mathbb{R}$, $y_0 \in \mathbb{R}$, $(y_1, \ldots, y_{m-1}) \mapsto \prod_{i=0}^{m-1} \psi(x_i, y_i, y_{i+1})$ is a (Gaussian) density function, we see by Fubini's theorem that

$$g^{(m)} \leqslant \left(\frac{m}{2\pi}\right)^{q/2}. \tag{A.2}$$

Both functions $p^{(m)}(x_0, x_m)$ and $r^{(m)}(x_0, y_0, x_m, y_m)$ are clearly differentiable with respect to x_0 and x_m with derivatives given by

$$\frac{\partial r^{(m)}}{\partial x_0} = \int \left[m I_d \left(x_1 - x_0 - \frac{b}{m}\right) + \frac{\partial \beta}{\partial x_0}\left(y_1 - y_0 - \frac{\beta(x_0, y_0)}{m}\right)\right]$$

$$\times \prod_{i=0}^{m-1} \phi(x_i, x_{i+1}) \psi(x_i, y_i, y_{i+1}) \mathrm{d}x_1 \ldots \mathrm{d}x_{m-1} \mathrm{d}y_1 \ldots \mathrm{d}y_{m-1},$$

$$\frac{\partial r^{(m)}}{\partial x_m} = \int \left[-m I_d \left(x_m - x_{m-1} - \frac{b}{m}\right)\right]$$

$$\times \prod_{i=0}^{m-1} \phi(x_i, x_{i+1}) \psi(x_i, y_i, y_{i+1}) \mathrm{d}x_1 \ldots \mathrm{d}x_{m-1} \mathrm{d}y_1 \ldots \mathrm{d}y_{m-1},$$

$$\frac{\partial p^{(m)}}{\partial x_0} = \int\left[mI_d\left(x_1 - x_0 - \frac{b}{m}\right)\right]\prod_{i=0}^{m-1}\phi(x_i, x_{i+1})\mathrm{d}x_1 \ldots \mathrm{d}x_{m-1},$$

$$\frac{\partial p^{(m)}}{\partial x_m} = \int\left[-mI_d\left(x_m - x_{m-1} - \frac{b}{m}\right)\right]\prod_{i=0}^{m-1}\phi(x_i, x_{i+1})\mathrm{d}x_1 \ldots \mathrm{d}x_{m-1}.$$

Now using the same arguments as for (A.2), one obtains

$$\left|\frac{\partial r^{(m)}}{\partial x_0}\right| \leqslant Cm^{q/2}\int\left(m\left|x_1 - x_0 - \frac{b}{m}\right| + 1\right)\prod_{i=0}^{m-1}\phi(x_i, x_{i+1})\mathrm{d}x_1 \ldots \mathrm{d}x_{m-1},$$

for some positive constant C. Hence,

$$\left|\frac{\partial r^{(m)}/\partial x_0}{p^{(m)}}\right| \leqslant Cm^{q/2}(1 + mB(x_0, x_m)), \tag{A.3}$$

where

$$B(x_0, x_m) = \frac{\displaystyle\int |x_1 - x_0 - (b/m)|\Pi_{i=0}^{m-1}\phi(x_i, x_{i+1})\mathrm{d}x_1 \ldots \mathrm{d}x_{m-1}}{\displaystyle\int \Pi_{i=0}^{m-1}\phi(x_i, x_{i+1})\mathrm{d}x_1 \ldots \mathrm{d}x_{m-1}}.$$

By making the change of variables $x_i \to x_i - x_{i-1} - b/m$, $i = 1, \ldots, m-1$, we have $B(x_0, x_m) = \bar{B}(x_m - x_0 - b)$ with

$$\bar{B}(x) = \frac{\displaystyle\int |x_1|\exp\left[-(m/2)\left(\textstyle\sum_{i=1}^{m-1}|x_i|^2 + \left|\sum_{i=1}^{m-1}x_i - x\right|^2\right)\right]\mathrm{d}x_1 \ldots \mathrm{d}x_{m-1}}{\displaystyle\int \exp\left[-(m/2)\left(\textstyle\sum_{i=1}^{m-1}|x_i|^2 + \left|\sum_{i=1}^{m-1}x_i - x\right|^2\right)\right]\mathrm{d}x_1 \ldots \mathrm{d}x_{m-1}}. \tag{A.4}$$

By writing the sum in parentheses in the previous relation as a canonical square sum in x_{m-1}, that is,

$$\sum_{i=1}^{m-1}|x_i|^2 + \left|\sum_{i=1}^{m-1}x_i - x\right|^2 = 2\left|x_{m-1} + \frac{\sum_{i=1}^{m-2}x_i - x}{2}\right|^2 + \sum_{i=1}^{m-2}|x_i|^2 + \tfrac{1}{2}\left|\sum_{i=1}^{m-2}x_i - x\right|^2,$$

we obtain by integrating in (A.4) first with respect to x_{m-1} (by Fubini's theorem):

$$\bar{B}(x) = \frac{\displaystyle\int |x_1|\exp\left[-(m/2)\left(\textstyle\sum_{i=1}^{m-2}|x_i|^2 + \frac{1}{2}\left|\sum_{i=1}^{m-2}x_i - x\right|^2\right)\right]\mathrm{d}x_1 \ldots \mathrm{d}x_{m-2}}{\displaystyle\int \exp\left[-(m/2)\left(\textstyle\sum_{i=1}^{m-2}|x_i|^2 + \frac{1}{2}\left|\sum_{i=1}^{m-2}x_i - x\right|^2\right)\right]\mathrm{d}x_1 \ldots \mathrm{d}x_{m-2}}.$$

By induction, this yields

$$\bar{B}(x) = \frac{\displaystyle\int |x_1| \exp[-(m/2)(|x_1|^2 + \tfrac{1}{m-1}|x_1 - x|^2)]\mathrm{d}x_1}{\displaystyle\int \exp\,[-(m/2)(|x_1|^2 + \tfrac{1}{m-1}|x_1 - x|^2)]\mathrm{d}x_1}.$$

Using again the canonical square sum in x_1, we then obtain

$$\bar{B}(x) = \frac{\displaystyle\int |x_1| \exp(-(m^2/2)(m-1)|x_1 - x/m|^2)\mathrm{d}x_1}{\displaystyle\int \exp(-m^2/2(m-1)|x_1 - x/m|^2)\mathrm{d}x_1}.$$

With the change of variable $x_1 \mapsto m(x_1 - x/m)/\sqrt{m-1}$, it is then clear that

$$\bar{B}(x) \leqslant \frac{C}{m}(\sqrt{m} + |x|), \qquad \forall x \in \mathbb{R}^d,$$

for some positive constant C. From (A.3), we deduce that

$$\left| \frac{\partial r^{(m)}/\partial x_0}{p^{(m)}} \right| \leqslant Cm^{q/2}(\sqrt{m} + |x_0| + |x_m|).$$

By the same arguments as above, we have

$$\left| \frac{\partial p^{(m)}/\partial x_0}{p^{(m)}} \right| \leqslant C(\sqrt{m} + |x_0| + |x_m|).$$

Therefore,

$$\left| \frac{\partial g^{(m)}}{\partial x_0} \right| \leqslant Cm^{g/2}(\sqrt{m} + |x_0| + |x_m|). \tag{A.5}$$

By the same arguments as above, we also show that

$$\left| \frac{\partial g^{(m)}}{\partial x_m} \right| \leqslant Cm^{q/2+1}(\sqrt{m} + |x_0| + |x_m|). \tag{A.6}$$

The local Lipschitz assumption (7.7) straightforwardly follows from (A.5)–(A.6). □

Appendix B: Optimal quantization: numerical aspects

As mentioned in the introduction, quantization consists of replacing an $\mathbb{R}^d$-valued random vector X by its projection according to a nearest-neighbour rule onto a grid $\Gamma \subset \mathbb{R}^d$, $\hat{X}^\Gamma := \mathrm{Proj}_\Gamma(X)$. For a grid $\Gamma := \{x^1, \ldots, x^N\}$, such a projection is defined by a Borel partition $C_1(\Gamma), \ldots, C_N(\Gamma)$ of $\mathbb{R}^d$ (called the Voronoi tessellation of Γ) satisfying $C_i(\Gamma) \subset \left\{\xi \in \mathbb{R}^d \colon |\xi - x^i| = \min_{x^j \in \Gamma} |\xi - x^j|\right\}$, $i = 1, \ldots, N$, where $|\cdot|$ denotes the usual canonical Euclidean norm. We then set

$$\hat{X}^{\Gamma} = \sum_{i=1}^{N} x^i \mathbf{1}_{C_i(\Gamma)}(X). \tag{B.1}$$

If $X \in L^p$, the L^p error induced by this projection – called the L^p quantization error – is given by $\|X - \hat{X}\|_p$. It is obvious that this quantization error depends on the grid Γ. In fact, one can easily derives from the nearest-neighbour rule that if $\Gamma = \{x^1, \ldots, x^N\}$, then

$$\|X - \hat{X}^{\Gamma}\|_p^p = \mathbb{E}\left(\min_{1 \leqslant i \leqslant N} |X - x^i|^p\right). \tag{B.2}$$

So if one identifies a grid Γ of size N with the N-tuple $(x^1, \ldots, x^N)$ or any permutation of it, the pth power of the L^p quantization error – called the L^p *distortion* – appears as a symmetric function

$$Q_N^p(x^1, \ldots, x^N) := \int \min_i |\xi - x^i|^p \mathbb{P}_X(\mathrm{d}\xi)$$

which can obviously be defined on the whole $(\mathbb{R}^d)^N$. The function $\sqrt[p]{Q_N^p}$ is Lipschitz continuous and does reach a minimum. If $|X(\Omega)|$ is infinite, then any N-tuple that achieves the minimum has pairwise distinct components and this minimum decreases toward 0 as N goes to infinity. Its rate of convergence is governed by Zador's theorem (see Graf and Luschgy 2000):

Theorem B.1. *Assume that* $\mathbb{E}|X|^{p+\varepsilon} < +\infty$ *for some* $\varepsilon > 0$. *Then*

$$\lim_N \left(N^{1/d} \min_{|\Gamma| \leqslant N} \|X - \hat{X}^{\Gamma}\|_p \right) = \tilde{J}_{p,d} \left(\int_{\mathbb{R}^d} \varphi(\xi)^{d/(d+p)} \mathrm{d}\xi \right)^{1/p+1/d} \tag{B.3}$$

where $\mathbb{P}_X(\mathrm{d}\xi) = \varphi(\xi)\lambda_d(\mathrm{d}\xi) + \nu(\mathrm{d}\xi)$, $\nu \perp \lambda_d$ *(*λ_d *being Lebesgue measure on* $\mathbb{R}^d$*). The constant* $\tilde{J}_{p,d}$ *corresponds to the case of the uniform distribution on* $[0, 1]^d$.

Except in one dimension ($\tilde{J}_{p,1} = 1/2(p+1)^{1/p}$, $\tilde{J}_{2,2} = \sqrt{(5/18\sqrt{3})}, \ldots$) the true value of $\tilde{J}_{p,d}$ is unknown. However, $\tilde{J}_{p,d} \sim (d/2\pi \mathrm{e})^{1/2}$ as d goes to infinity (see Graf and Luschgy 2000). This theorem says that $\min_{|\Gamma| \leqslant N} \|X - \hat{X}^{\Gamma}\|_p \sim \theta_{X,p,d} N^{-1/d}$. This is in accordance with the rates $O(N^{-1/d})$ obtained in numerical integration with uniform grid methods (when $N = M^d$) although optimal quantizers are never uniform grids (except for the uniform distribution in one dimension): optimal quantization provides the 'best-fitting' grid of size N for a given distribution $\mu = \mathbb{P}_X$. Such grids correspond to the real constant $\theta_{X,p,d}$.

B.1. Stochastic gradient descent

When the dimension d is greater than 1 and independent copies of X can easily be simulated on a computer, an efficient approach consists of *differentiating* the integral representation of the quantization error function to implement a stochastic gradient algorithm. That is, set for every $x = (x^1, \ldots, x^N) \in (\mathbb{R}^d)^N$ and every $\xi \in \mathbb{R}^d$,

$$q^p_{_N}(x,\xi) := \min_{1\leqslant i\leqslant N} |x^i - \xi|^p.$$

For notational convenience, we will temporarily denote by $C_i(x)$ the Voronoi cell of x^i in the grid $\Gamma := \{x^1, \ldots, x^N\}$ (instead of $C_i(\Gamma)$). One can show (see Pagès 1997) that, if $p > 1$, $Q^p_{_N}$ is continuously differentiable at every N-tuple $x \in (\mathbb{R}^d)^N$ having pairwise distinct components and a $\mathbb{P}_X$-negligible Voronoi boundary $\cup_{i=1}^N \partial C_i(x)$. Its gradient $\nabla Q^p_{_N}$ is obtained by formal differentiation:

$$\nabla Q^p_{_N}(x) = \mathbb{E}\Big[\nabla_x\, q^p_{_N}(x, X)\Big], \tag{B.4}$$

where

$$\nabla_x\, q^p_{_N}(x,\xi) = \left(\frac{\partial q^p_{_N}}{\partial x^i}(x,\xi)\right)_{1\leqslant i\leqslant N} := p\left(\frac{x^i-\xi}{|x^i-\xi|}|x^i-\xi|^{p-1}\mathbf{1}_{C_i(x)}(\xi)\right)_{1\leqslant i\leqslant N}$$

with the convention that $0/|0| = 0$. Note that then, $\nabla_x q^p_{_N}(x,\xi)$ has exactly one non-zero component $i(x,\xi)$ defined by $\xi \in C_{i(x,\xi)}(x)$. The above differentiability result still holds for $p = 1$ if $\mathbb{P}_X$ is continuous (i.e. weights no point in $\mathbb{R}^d$).

Then, one may process a stochastic gradient descent algorithm (starting from an initial grid Γ^0 with N pairwise distinct components) defined by

$$\Gamma^{s+1} = \Gamma^s - \frac{\delta_{s+1}}{p}\nabla_x\, q^p_{_N}(\Gamma^s, \xi^{s+1}), \tag{B.5}$$

where $(\xi^s)_{s\geqslant 1}$ is an i.i.d. sequence of X-distributed random vectors and $(\delta_s)_{s\geqslant 1}$ a $(0, 1)$-valued sequence of step parameters satisfying the usual conditions,

$$\sum_s \delta_s = +\infty \quad \text{and} \quad \sum_s \delta_s^2 < +\infty. \tag{B.6}$$

Note that (B.5) almost surely implies by induction that Γ^s has pairwise distinct components for every s. Under some appropriate assumptions, such a stochastic descent procedure almost surely converges toward a local minimum of its potential function; here it would be $Q^p_{_N}$. Although these assumptions are not fulfilled by the function $Q^p_{_N}$, some theoretical problems may be overcome (see Pagès 1997). However, it provides satisfactory results a posteriori (this is a common situation when implementing gradient descents). The companion parameters ($\mathbb{P}_X$-weights of the cells and L^p quantization errors) can be obtained as by-products of the procedure. For more details, we refer to Pagès (1997), Bally and Pagès (2003) and Pagès *et al*. (2004a), where these questions are extensively investigated and discussed.

The quadratic case $p = 2$ is the most commonly implemented for applications and is known as the competitive learning vector quantization algorithm.

B.2. Stationary quantizers

The differentiability of $Q^2_{_N}$ also has a noticeable theoretical consequence. Since $Q^2_{_N}$ is differentiable at any N-tuple x lying in $\mathrm{argmin} Q^2_{_N}$ – even in $\mathrm{argminloc} Q^2_{_N}$ if $\mathbb{P}_{_X}$ weights no hyperplane – any such N-tuple is a stationary quantizer, that is, $\nabla Q^2_{_N}(x) = 0$. Standard computations then show that this equation reads

$$\hat{X} = \mathbb{E}[X \mid \hat{X}]. \tag{B.7}$$

In particular this implies that, for every $p \in [1, +\infty]$, $\|\hat{X}\|_p \leqslant \|X\|_p$.

Acknowledgement

We are grateful to Afef Sellami and Jacques Printems who played an important part in the numerical experiments. We would like to thank François LeGland for useful remarks.

References

Bally, V. and Pagès G. (2003) A quantization algorithm for solving multidimensional discrete-time optimal stopping problems. *Bernoulli*, **9**, 1003–1049.

Bally, V. and Talay, D. (1996) The law of the Euler scheme for stochastic differential equations: II. Approximation of the density. *Monte Carlo Methods Appl.*, **2**, 93–128.

Bally, V., Pagès, G. and Printems, J. (2001) A stochastic quantization method for nonlinear problems. *Monte Carlo Methods Appl.*, **7**, 21–34.

Crisan, D. and Lyons, T. (1997) Nonlinear filtering and measure-valued processes. *Probab. Theory Related Fields*, **109**, 217–244.

Del Moral, P. (1998) Measure-valued processes and interacting particle systems, application to nonlinear filtering problem. *Ann. Appl. Probab.*, **8**, 438–495.

Del Moral, P., Jacod, J. and Protter, P. (2001) The Monte-Carlo method for filtering with discrete-time observations, *Probab. Theory Related Fields*, **120**, 346–368.

Devroye, L. (1986) *Non-uniform Random Variate Generation*. New York: Springer-Verlag.

Di Masi, G. and Runggaldier, W. (1982) Approximation and bounds for discrete-time nonlinear filtering, for the nonlinear filtering problem with error bounds. In A. Bensoussan and J.L. Lions (eds), *Analysis and Optimization of systems*, Lecture Notes in Control and Inform. Sci. 44. Berlin: Springer-Verlag.

Di Masi, G., Pratelli, M. and Runggaldier, W. (1985) An approximation for the nonlinear filtering problem with error bounds. *Stochastics*, **14**, 247–271.

Elliott, R., Aggoun, L. and Moore, J.B. (1995) *Hidden Markov Models, Estimation and Control*. Berlin: Springer-Verlag.

Florchinger, P. and LeGland, F. (1992) Particle Approximation for First-Order SPDE's. In I. Karatzas and D. Ocone (eds), *Applied Stochastic Analysis*, Lecture Notes in Control and Inform. Sci. 177, pp. 121–133, Berlin: Springer-Verlag.

Graf, S. and Luschgy H. (2000) *Foundations of Quantization for Probability Distributions*, Lecture Notes in Math. 1730. Berlin: Springer-Verlag.

Kallianpur, G. and Striebel, C. (1968) Arbitrary system process with additive white noise observation errors. *Ann. Math. Statist.*, **39**, 785–801.

Kushner, H.J. (1977) *Probability Methods for Approximations in Stochastic Control and for Elliptic Equations*. New York: Academic Press.

Newton, N. (2000a) Observations preprocessing and quantisation for nonlinear filters. *SIAM J. Control Optim.*, **38**, 482–502.

Newton, N. (2000b) Observation sampling and quantisation for continuous-time estimators. *Stochastic Process. Appl.*, **87**, 311–337.

Pagès, G. (1997) A space vector quantization method for numerical integration. *J. Appl. Comput. Math.*, **89**, 1–38.

Pagès, G. and Printems, J. (2003) Optimal quadratic quantization for numerics: the Gaussian case. *Monte Carlo Methods Appl.*, **9**, 135–166.

Pagès, G., Pham, H. and Printems, J. (2004a) A quantization algorithm for multidimensional stochastic control problems. *Stochastics and Dynamics*, **4**(4), 1–45.

Pagès, G., Pham, H. and Printems, J. (2004b) Optimal quantization methods and applications to numerical problems in finance. In S.T. Rachev (ed.), *Handbook of Computational and Numerical Methods in Finance*. Boston: Birkhäuser.

Sellami, A. (2004) Non linear filtering with quantization of the observations. Preprint, LPMA.

Received April 2003 and revised November 2004