Optimal quantization methods for nonlinear filtering with discrete-time observations

Gilles PAGÈS  
Laboratoire de Probabilités et  
Modèles Aléatoires  
CNRS, UMR 7599  
Université Paris 6  
e-mail: gpa@ccr.jussieu.fr

Huyệ́n PHAM  
Laboratoire de Probabilités et  
Modèles Aléatoires  
CNRS, UMR 7599  
Université Paris 7  
e-mail: pham@math.jussieu.fr

and CREST, Laboratoire de Finance-Assurance

Abstract

We develop an optimal quantization approach for numerically solving nonlinear filtering problems associated with discrete-time or continuous-time state process and discrete-time observations. Two quantization methods are proposed: 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.

Key words: Nonlinear filtering, Markov chain, Euler scheme, vector quantization, stochastic gradient descent, stationary signal, numerical approximation.

1 Introduction

We address the following nonlinear discrete time filtering problem. In the sequel, 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 probability transition \(P_k(x, dx'), k \geq 1\) (i.e. the transition from time \(k - 1\) to time \(k\)). The initial law of \(X_0\) is known and denoted by \(\mu\). We have noisy observations \(\{Y_k, k \in \mathbb{N}^*\}\) valued in \(\mathbb{R}^q\) and our aim is to compute at some time \(n \geq 1\) 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

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

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

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

We consider an observation process (or design) where the pair \((X_k, Y_k)\) is a Markov chain and such that for all \(k \geq 1\):

\((H)\) The law of \(Y_k\) conditional on \((X_{k-1}, Y_{k-1}, X_k)\) admits a density \(y' \mapsto g_k(x, y, x', y')\).

Notice that the probability transition of the Markov chain \((X_k, Y_k)\) is then given by \(P_k(x, dx')g_k(x, y, x', y')dy'\).

An example of the observation scheme considered above is the model:

\[
\begin{align*}
X_k &= F_k(X_{k-1}, \varepsilon_k), \quad k = 1, \ldots, n, \\
Y_k &= G_k(X_{k-1}, Y_{k-1}, X_k, \eta_k), \quad k = 1, \ldots, n,
\end{align*}
\]

where \((\varepsilon_k)\) and \((\eta_k)\) are two i.i.d. independent sequences of random variables, and \(F_k, G_k\) are measurable functions. The pair \((X_k, Y_k)\) is then Markovian with respect to the filtration generated by \((\varepsilon_k, \eta_k)\). The basic assumption concerning \(G_k\) and \((\eta_k)\) 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] or [Di Masi et al. 1985] for approaches related to the one followed here. In these papers, for an observation scheme in the particular form

\[
Y_k = G(X_k) + \eta_k, \quad k = 1, \ldots, n,
\]

the method consists basically in approximating the signal Markov chain \((X_k)\) (or in the continuous-time problem, the signal diffusion \((X_t)\)) 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 \((X_k)_{1 \leq k \leq n}\) optimally “fitted” to its probabilistic features. Let us be more specific by considering the case of a single random vector \(X\). If one wishes to approximate \(X\) by a random vector taking its values in a finite grid \(\Gamma := \{x^1, \ldots, x^N\}\), we consider its projection \(\text{Proj}_\Gamma(X)\) on the grid following the closest neighbour rule. Then, the resulting mean \(L^p\)-error \((p \geq 1)\) is \(\|X - \text{Proj}_\Gamma(X)\|_p = \min_{1 \leq i \leq N} |X - x^i|_p\). It only depends on the distribution \(\mathbb{P}_X\) of \(X\) and the grid \(\Gamma\). For historical reasons, \(\text{Proj}_\Gamma(X)\) is often called the quantization of the r.v. \(X\) by the grid \(\Gamma\) 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 50’s. Thus, the \(L^p\)-mean quantization error is continuous as a function of the grid \(\Gamma\) and reaches a minimum over all the grids \(\Gamma\) with size at most \(N\). Furthermore, following Zador’s Theorem (see [Graf and Luschgy 2000]),

\[
\min_{\Gamma, |\Gamma| = N} \|X - \text{Proj}_\Gamma(X)\|_p = c(\mathbb{P}_X, p) \times c(d) N^{-\frac{1}{d}} + o(N^{-\frac{1}{d}})
\]

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 like 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 \(\|X - \text{Proj}_\Gamma(X)\|_p^p\) one derives a stochastic gradient descent that converges towards some (at least locally) optimal grids. The distribution of \(\text{Proj}_\Gamma(X)\) and the resulting quantization error can be obtained as a by-products, especially when in the quadratic case \(p = 2\) (see, e.g. [Pagès 1997]). Simulations like those carried out in [Pagès and Printems 2003] for the 2-dim Gaussian distribution confirm what could be a priori expected: the heavier 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 \(C^1\) with Lipschitz continuous differential \(Df\) then,

\[
\mathbb{E}f(\text{Proj}_{\Gamma^*}(X)) = \sum_{1 \leq i \leq N} f(x^{*,i}) p_i \quad \text{with} \quad p_i = \mathbb{P}(\text{Proj}_{\Gamma^*}(X) = x^i)
\]

\[
|\mathbb{E}f(\text{Proj}_{\Gamma^*}(X)) - \mathbb{E}f(X)| \leq [Df]_{Lip} \|X - \text{Proj}_{\Gamma^*}(X)\|_2^2 = O\left(\frac{1}{N^{2/d}}\right).
\]
this shows that weak approximation by quantization can be strictly more performing than strong approximation, so that quantization method may overperform Monte Carlo simulation at least up to 4 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 $N$.

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

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

where the grids $\Gamma_k$ minimize the $L^p$-quantization error $\|X_k - \text{Proj}_{\Gamma_k}(X_k)\|_p$ among the grids with size $N_k$ for every $k = 0, \ldots, n$. The sequence $(\hat{X}_k)_{0 \leq k \leq n}$ no longer has the Markov property. Then, one defines the approximate quantized filter by simply replacing in the forward explicit formula of 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 in the form (1.2)-(1.3), one sets

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

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

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

This Markovian quantization approach was introduced in [Pagès et al. 2004a] to approximate numerically some stochastic control problems for multi-dimensional 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 $\|Z_k - \text{Proj}_{\Gamma_k}(Z_k)\|_p$, 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 quantization, 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 have been originally introduced.

In Section 6, we analyze 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 \((X_k)_{0 \leq k \leq n}\) has been processed and kept off-line, it *instantaneously produces* for any set of observations some reproducible deterministic results. This underlines 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 multi-dimensional signal processes, at least up to 4-dimension 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 like 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 storing.

In the recent past 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] and [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 field of applications are quite different.

It may also be interesting to keep the same filter for close observation samples in order to spare 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] and [Newton 2000b]: some functional weak convergence results are stated toward the original filter. In a framework where both \(X\) and the observation process are quantized, some a priori error bounds are derived in [Sellami 2004].

The paper is organized as follows. Some preliminaries on nonlinear filtering are provided in Section 2. In particular, we recall the known forward inductive formula for the filter, and also the (less 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 get 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 process. We discuss in Section 7 how our results may be applied to the case of discretely observed diffusions. Finally, we provide in Section 8 several numerical experiments. Firstly, we compare our approximate filter with the explicit Kalman-Bucy filter: one 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 converging behaviour is emphasized as the quantization accuracy increases, although no reference value is available.

**Notations:**
- For every Borel function \(f : \mathbb{R}^d \to \mathbb{R}\) set
  \[
  \|f\|_\infty = \sup_{x \in \mathbb{R}^d} |f(x)| \quad \text{and} \quad [f]_{\text{Lip}} = \sup_{x \neq x'} \frac{|f(x) - f(x')|}{|x - x'|}.
  \]
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 $Pf(x) := \int f(x')P(x, dx').$

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 derives the Kallianpur-Striebel formula for the filter:

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

(2.1)

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

$$\pi_{y,n}f = \int f(x_n)\mu(dx_0)\prod_{k=1}^{n} g_k(x_{k-1}, y_{k-1}, x_k, y_k)P_k(x_{k-1}, dx_k),$$

(2.2)

$$\pi_{y,n}1 = E[f(X_n)L_{y,n}],$$

(2.3)

with

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

(2.4)

(convention $y_0 = 0$). Notice that

$$\pi_{y,n}1 = E[L_{y,n}] = \phi_n(y).$$

(2.5)

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

In the sequel, we shall denote for notational convenience

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

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

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

For convenience we also define

$$H_{y,0}f(x) := \pi_{y,0}f = E[f(X_0)] = \int f(x_0)\mu(dx_0), \quad x \in \mathbb{R}^d.$$  

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

$$\pi_{y,k}f = \pi_{y,k-1}H_{y,k}f, \quad k = 1, \ldots, n$$

(2.6)

so that

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

(2.7)
Equation (2.6) is called the \textit{forward expression} of the filter. One also derives from the \textquote{symmetric} expression (2.7) a \textit{backward expression} for the filter which turns out to be a useful tool for 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

\[
u_{y,n}(f)(x) = f(x), \quad u_{y,k-1}(f) = H_{y,k} u_{y,k}(f), \quad k = 0, \ldots, n.
\] (2.8)

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

We shall replace the true filter by a computable approximate filter constructed. 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) \( \hat{\pi}_{y,k} f = \pi_{y,-1} \hat{H}_{y,k} f \).

3 Approximate filter by marginal quantization

3.1 Method

In this section, one considers a marginal quantization of the Markov chain \( (X_k)_k \), time by time \( i.e \)

\[
\hat{X}_k = \text{Proj}_{\Gamma_k}(X_k), \quad 0 \leq k \leq n,
\] (3.1)

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

\[
\hat{P}^i_j = \mathbb{P}[\hat{X}_k = x^i_k | \hat{X}_{k-1} = x^{i-1}_k], \quad i = 1, \ldots, N_{k-1}, \quad j = 1, \ldots, N_k.
\] (3.2)

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

\[
\hat{H}_{y,k} := \sum_{j=1}^{N_k} \hat{H}_{y,k}^i \delta_{x^i_{k-1}}, \quad k = 1, \ldots, n,
\] (3.3)

with

\[
\hat{H}_{y,k}^i = g_{y,k}(x^i_k, x^{i+1}_k) \hat{P}^i_j, \quad i = 1, \ldots, N_{k-1}, \quad j = 1, \ldots, N_k,
\] (3.4)

for \( k = 1, \ldots, n \), so that, for every function \( f : \Gamma_k \rightarrow \mathbb{R} \),

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

Finally, we set

\[
\hat{H}_{y,0} = \sum_{i=1}^{N_0} \hat{P}^i_0 \delta_{x^i_0} \quad \text{with} \quad \hat{P}^i_0 := \mathbb{P}[\hat{X}_0 = x^i_0], \quad i = 1, \ldots, N_0.
\]
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
\[
\hat{\pi}_{y,n} = \hat{H}_{y,0} \circ \cdots \circ \hat{H}_{y,n}.
\]
It is easily computed by the following forward induction:
\[
\begin{align*}
\hat{\pi}_{y,0} &= \hat{H}_{y,0}, \\
\hat{\pi}_{y,k} &= \pi_{y,k-1} \hat{H}_{y,k} := \left[ \sum_{i=1}^{N_{k-1}} \hat{H}_{y,j} \hat{\pi}_{y,k-1} \right]_{j=1, \ldots, N_k}, \quad k = 1, \ldots, n. \tag{3.5}
\end{align*}
\]
The approximate filter \( \hat{\Pi}_{y,n} \) is then given by:
\[
\hat{\Pi}_{y,n} = \sum_{i=1}^{N_n} \hat{\Pi}^i_{y,n} \delta_{x_n^i} \quad \text{with} \quad \hat{\Pi}^i_{y,n} = \frac{\hat{\pi}^i_{y,n}}{\sum_{i=1}^{N_n} \hat{\pi}^i_{y,n}}, \quad i = 1, \ldots, N_n.
\]

### 3.2 Error analysis

In this paragraph, we will estimate the accuracy of the approximate filter \( \hat{\Pi}_{y,n} \) in terms of the marginal quantization errors on the signal \( \| \Delta_k \|_1 \), \( k = 0, \ldots, n \), defined by:
\[
\Delta_k = X_k - \text{Proj}_{\Gamma_k}(X_k). \tag{3.6}
\]

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 \( X_{k-1}, Y_{k-1}, X_k \).

To be more precise, we first recall some definitions. We say that a probability transition \( P \) on \( \mathbb{R}^d \) is \( C \)-Lipschitz for some positive real constant \( C \) if for every Lipschitz function \( \varphi \) on \( \mathbb{R}^d \) with ratio \( [\varphi]_{Lip} \), \( P \varphi \) is Lipschitz and \( [P \varphi]_{Lip} \leq C[\varphi]_{Lip} \). Then, we may define the Lipschitz ratio
\[
[P]_{Lip} = \sup \left\{ \frac{[P \varphi]_{Lip}}{[\varphi]_{Lip}}, \varphi \text{ nonzero Lipschitz continuous function} \right\} < +\infty.
\]

\((A1)\) The Markov transition operators \( P_k(x, dx') \), \( k = 1, \ldots, n \), are Lipschitz, so that
\[
[P]_{Lip} := \max_{k=0, \ldots, n} [P_k]_{Lip} < +\infty.
\]

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

\[
(ii) \quad \text{For every } k = 1, \ldots, n, \text{ there exist two Borel functions}
\]
\[
[g_k^1]_{Lip} : \mathbb{R}^d \times \mathbb{R}^q \to \mathbb{R}^+ \text{ such that, for all } x, x', \tilde{x}, \tilde{x}' \in \mathbb{R}^d \text{ and } y, y' \in \mathbb{R}^q,
\]
\[
|g_k(x, y, x', y') - g_k(\tilde{x}, y, \tilde{x}', y')| \leq [g_k^1]_{Lip}(y, y')|x - \tilde{x}| + [g_k^2]_{Lip}(y, y')|x' - \tilde{x}'|.
\]

An essential tool for the proof of theorem 3.1 below is to introduce the sequence functions \((\hat{u}_{y,k}(f))_{-1 \leq k \leq n}\) which is the quantized counterpart of the sequence \((u_k(f))_{-1 \leq k \leq n}\) defined in (2.8) as the backward expression of the filter: mimicking this backward dynamic formula, one recursively define the \( \hat{u}_{y,k}(f) \) on \( \Gamma_k \), \( k = 0, \ldots, n \), by
\[
\hat{u}_{y,n}(f) = f, \quad \text{on the grid } \Gamma_n,
\]
\[
\hat{u}_{y,k}(f) = \hat{H}_{y,k+1} \hat{u}_{y,k+1}(f) \text{ on the grid } \Gamma_{k-1}, \quad k = -1, \ldots, n-1.
\]
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 functions \( f \) on \( \mathbb{R}^d \) and each \( n \)-tuple of observations \( y = (y_1, \ldots, y_n) \), we have, for every \( p \geq 1 \),

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

with \( \phi_n(y) := \hat{\pi}_{y,n}1 \),

\[
B_k^n(f,y,p) := (2 - \delta_{2,p})[P]^{n-k}_{-Lip}[f]_{Lip}^2 2 \left( \frac{\|f\|_\infty}{K_g} \left( |g_{k+1}^1|_{Lip}(y_k, y_{k+1}) + |g_{k}^2|_{Lip}(y_{k-1}, y_k) \right) \right) \tag{3.8}
\]

\[
+ (2 - \delta_{2,p}) \frac{\|f\|_\infty}{K_g} \sum_{j=k+1}^{n} [P]^{j-1}_{Lip} \left( |g_{j}^1|_{Lip}(y_{j-1}, y_j) + |P|_{Lip} |g_{j}^2|_{Lip}(y_{j-1}, y_j) \right).
\]  \tag{3.9}

(Convention: \( g_0 = g_{n+1} \equiv 0 \) and \( \delta_{r,p} \) is for the usual Kronecker symbol)

**Remark 3.1** Note that

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

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 \( D_k^n(f,y,p) \) are smaller than in the \( L^1 \)-case (other bounds are trivial since the \( L^p \)-norm is nondecreasing as a function of \( p \)).

**Remark 3.3** If one introduces \( |g|_{Lip} := \max_{k=1,\ldots,n} \sup_{y,y' \in \mathbb{R}^q} \left(|g_{k+1}^1|_{Lip}(y,y') + |g_{k}^2|_{Lip}(y,y')\right)\), the \( B_k^n(f,y,p) \) is upper-bounded by the simpler coefficient:

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

with the usual convention \( \frac{1}{u-1} \times (u^m - 1) = m \) when \( u = 1 \) and \( m \in \mathbb{N} \).

**Remark 3.4** If 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 \( |g_k|_{Liploc}^1, |g_k|_{Liploc}^2 : \mathbb{R}^q \times \mathbb{R}^q \rightarrow \mathbb{R}_+ \) such that, for all \( x, x', \hat{x}, \hat{x}' \in \mathbb{R}^d \) and \( y, y' \in \mathbb{R}^q \),

\[
|g_k(x, y, x', y') - g_k(\hat{x}, y, \hat{x}', y')| \leq |g_k|_{Liploc}^1(y, y')(1 + |x| + |x'| + |\hat{x}| + |\hat{x}'|)|x - \hat{x}|
\]

\[
+ |g_k|_{Liploc}^2(y, y')(1 + |x| + |x'| + |\hat{x}| + |\hat{x}'|)|x' - \hat{x}'|,
\]
then we may state a similar estimate for the approximate filter as in Theorem 3.1: for every 
$p \geq 1$ and every $p', q' \in (1, \infty)$, $1/p' + 1/q' = 1$:

$$
|\Pi_{y,n} f - \tilde{\Pi}_{y,n} f| \leq \frac{K^n}{\phi_n(y)} \sum_{k=0}^{n} B^n_k(p, pq', f) \|\Delta_k\|_{p''},
$$

with

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

and

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

Here we have set $[g]_{Lip loc} := \max_{k=1,\ldots,n} \sup_{y,y' \in \mathbb{R}^d} ([g^1_k]_{Lip loc}(y, y') + [g^2_k]_{Lip loc}(y, y'))$, $\|X\|_q = \max_{k=0,\ldots,n} \|X_k\|_q$
and $\|\hat{X}\|_q = \max_{k=0,\ldots,n} \|\hat{X}_k\|_q$.

Furthermore, when $\hat{X}$ is an optimal quadratic quantization one shows (see [Graf and Luschgy 2000]
or (B.7) in Appendix B) that $\hat{X}_k = \mathbb{E}(X_k | \hat{X}_k)$ so that $\|\hat{X}_k\|_r \leq \|X_k\|_r$ for every $r \geq 1$.
Hence, one may take

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

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 get the announced error bound, we need three steps. The first one in lemma 3.1 below
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 some bound
for the Lipschitz coefficient of the functions $\nu_k, f$ defined in (2.8). In the third step –
which is the proof of the theorem itself –, we will bound $\|\nu_k f(X_k) - \tilde{\nu}_k f(\hat{X}_k)\|_p$ by a
backward induction, having in mind that $|\pi_{y,n} f - \tilde{\pi}_{y,n} f| = |u_y f(X_k) - \tilde{u}_y f(\hat{X}_k)|$.

**Lemma 3.1** Let $(\mu_y)$ and $(\nu_y)$ 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 $f$

$$
\left| \int d\mu_y - \int f d\nu_y \right| \leq \|f\|_{Lip} R(\mu_y, \nu_y) + [f]_{Lip} S(\mu_y, \nu_y)
$$

(3.10)

then

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

**Proof:**

$$
\left| \frac{\int d\mu_y}{\mu_y(E)} - \frac{\int f d\nu_y}{\nu_y(E)} \right| \leq \frac{\left| \int d\mu_y - \int f d\nu_y \right|}{\mu_y(E)} + \frac{\|f\|_{Lip} S(\mu_y, \nu_y)}{\mu_y(E)} \leq \frac{\left( \|f\|_{Lip} R(\mu_y, \nu_y) + [f]_{Lip} S(\mu_y, \nu_y) \right)}{\mu_y(E)} + \frac{\|f\|_{Lip} S(\mu_y, \nu_y)}{\mu_y(E)} \leq \frac{\left( \|f\|_{Lip} R(\mu_y, \nu_y) + [f]_{Lip} S(\mu_y, \nu_y) + \|f\|_{Lip} S(\mu_y, \nu_y) \right)}{\mu_y(E)}
$$
In particular, for every \( f \) assume and from the Lipschitz property of the transitions \( \text{Proof of Theorem 3.1.} \)

One completes the proof by a symmetry argument. \( \square \)

**Lemma 3.2** Assume (A1) and (A2) hold. Let \((y_k)_{k=1,\ldots,n}\) 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 \( [u_{y,k}(f)]_{Lip} \) satisfying

\[
[u_{y,k}(f)]_{Lip} \leq [P_{k+1}]_{Lip} (K_g[u_{y,k+1}(f)]_{Lip} + \|u_{y,k+1}(f)\|_\infty [g_y^2]_{Lip} + \|u_{y,k+1}(f)\|_\infty [g_y^1]_{Lip})
\]

and

\[
\|u_{y,k}(f)\|_\infty \leq K_g^{n-k}\|f\|_\infty, \quad k = 0, \ldots, n.
\]

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

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

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 derives the first two formulae from the recursive definition (2.8) of the \( u_{k}^{\prime} \)’s:

\[
u_{k}^{\prime}(x) = \mathbb{E}[g_{k+1}(x, X_{k+1})u_{k+1}(X_{k+1}) | X_{k} = x] = \int g_{k+1}(x, x')u_{k+1}(x')P_{k+1}(x, dx')
\]

and from the Lipschitz property of the transitions \( P_{k}(x, dx') \):

\[
[u_{k}]_{Lip} \leq [P_{k+1}]_{Lip} \sup_{x \in \mathbb{R}^d} |x| \rightarrow g_{k+1}(x, x')u_{k+1}(x')|_{Lip} + \|u_{k+1}\|_\infty [g_k^1]_{Lip}.
\]

Now \( \|u_n\|_\infty = \|f\|_\infty \) and \( \|u_k\|_\infty \leq K_g\|u_{k+1}\|_\infty \) so that \( \|u_k\|_\infty \leq K_g^{-(k+1)}\|f\|_\infty \). Hence

\[
[u_k]_{Lip} \leq A[u_{k+1}]_{Lip} + B K_g^{-(k+1)}[g_k^2]_{Lip} + CK_g^{-(k+1)}[g_k^1]_{Lip}
\]

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

**Proof of Theorem 3.1.** To obtain an upper-bound for \( \|u_k(X_k) - \hat{u}_k(\hat{X}_k)\|_p \), one proceeds by induction. Set temporarily for every \( k \leq n-1 \) 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,

\[
\varphi(x_k, x_{k+1}, x_{k+1}') = \mathbb{E}(\varphi(X_k, X_{k+1}, X_{k+1}) | X_k) - \mathbb{E}(g_{k+1}(\hat{X}_k, \hat{X}_{k+1})\hat{u}_{k+1}(\hat{X}_{k+1}) | \hat{X}_k)\]

Using that \( \mathbb{E}(. | \hat{X}_k) \) is a \( L^p \)-contraction, one gets for every \( k \in \{0, \ldots, n-1\} \),

\[
\|u_k(X_k) - \hat{u}_k(\hat{X}_k)\|_p \leq \|\mathbb{E}(\varphi(X_k, X_{k+1}, X_{k+1}) | X_k) - \mathbb{E}(\varphi(\hat{X}_k, \hat{X}_{k+1}, X_{k+1}) | \hat{X}_k)\|_p
\]

\[
+\|\mathbb{E}(\varphi(\hat{X}_k, \hat{X}_{k+1}, X_{k+1}) - g_{k+1}(\hat{X}_k, \hat{X}_{k+1})\hat{u}_{k+1}(\hat{X}_{k+1}) | \hat{X}_k)\|_p
\]

\[
\leq \|\mathbb{E}(\varphi(X_k, X_{k+1}, X_{k+1}) | X_k) - \mathbb{E}(\varphi(\hat{X}_k, \hat{X}_{k+1}, X_{k+1}) | \hat{X}_k)\|_p + K_g\|u_{k+1}(X_{k+1}) - \hat{u}_{k+1}(\hat{X}_{k+1})\|_p. \quad (3.11)
\]

11
Now
\[
\|\mathbb{E}(\varphi(X_k, X_{k+1}, X_{k+1}) | X_k) - \mathbb{E}(\varphi(\tilde{X}_k, \tilde{X}_{k+1}, X_{k+1}) | \tilde{X}_k)\|_p \leq \|u_k(X_k) - \mathbb{E}(u_k(X_k) | \tilde{X}_k)\|_p + \|\mathbb{E}(u_k(X_k) | \tilde{X}_k) - \mathbb{E}(\varphi(\tilde{X}_k, \tilde{X}_{k+1}, X_{k+1}) | \tilde{X}_k)\|_p,
\]
(3.12)
Then
\[
\|u_k(X_k) - \mathbb{E}(u_k(X_k) | \tilde{X}_k)\|_p \leq \|u_k(X_k) - u_k(\tilde{X}_k)\|_p + \|\mathbb{E}(u_k(\tilde{X}_k) - u_k(X_k) | \tilde{X}_k)\|_p \leq 2\|u_k(X_k) - u_k(\tilde{X}_k)\|_p
\]

since conditional expectation is a $L^p$-contraction. When $p = 2$,
\[
\|u_k(X_k) - \mathbb{E}(u_k(X_k) | \tilde{X}_k)\|_2 = \min \left\{ \|u_k(X_k) - \psi(\tilde{X}_k)\|_2, \psi(\tilde{X}_k) \in L^2 \right\} \leq \|u_k(X_k) - u_k(\tilde{X}_k)\|_2.
\]

Now $\tilde{X}_k$ being $\sigma(X_k)$-measurable, $\mathbb{E}(u_k(X_k) | \tilde{X}_k) = \mathbb{E}(\varphi(X_k, X_{k+1}, X_{k+1}) | \tilde{X}_k)$, so that
\[
\|\mathbb{E}(u_k(X_k) | \tilde{X}_k) - \mathbb{E}(\varphi(\tilde{X}_k, \tilde{X}_{k+1}, X_{k+1}) | \tilde{X}_k)\|_p \leq \|\varphi(X_k, X_{k+1}, X_{k+1}) - \varphi(\tilde{X}_k, \tilde{X}_{k+1}, X_{k+1})\|_p \leq \|u_{k+1}\|_\infty \|g_{k+1}(X_k, X_{k+1}) - g_{k+1}(\tilde{X}_k, \tilde{X}_{k+1})\|_p.
\]

Consequently, plugging into Equation (3.12) yields
\[
\|\mathbb{E}(\varphi(X_k, X_{k+1}, X_{k+1}) | X_k) - \mathbb{E}(\varphi(\tilde{X}_k, \tilde{X}_{k+1}, X_{k+1}) | \tilde{X}_k)\|_p \leq (2 - \delta_{p,2})\|u_k\|_{L^{p}}\|X_k - \tilde{X}_k\|_p + \|u_{k+1}\|_\infty \left( \|g_{k+1}\|_{L^{p}}\|X_k - \tilde{X}_k\|_p + \|g_{k+1}\|_{L^{p}}\|X_{k+1} - \tilde{X}_{k+1}\|_p \right)
\]

Since we are dealing with marginal quantization $\Delta_k = X_k - \tilde{X}_k$, plugging into Equation (3.11) yields the following induction:

for every $k \in \{0, \ldots, n - 1\}$,
\[
\|u_k(X_k) - \tilde{u}_k(\tilde{X}_k)\|_p \leq K_\|u_{k+1}(X_{k+1}) - \tilde{u}_{k+1}(\tilde{X}_{k+1})\|_p + \alpha_k\|\Delta_k\|_p + \beta_k\|\Delta_{k+1}\|_p
\]
with
\[
\alpha_k := (2 - \delta_{p,2})\|u_k\|_{L^{p}} + \|u_{k+1}\|_\infty \|g_{k+1}\|_{L^{p}}, 0 \leq k \leq n - 1,
\]
and
\[
\beta_k := \|g_{k}\|_{L^{p}}\|u_k\|_\infty, 1 \leq k \leq n.
\]

For notational convenience, one sets $\alpha_n := (2 - \delta_{p,2})[f]_{L^{p}}$ (in fact $\alpha_n = [f]_{L^{p}}$ would always be suitable) and $\beta_0 := 0$. Then, standard computations using Lemma 3.2 yield
\[
\|\bar{\pi}_{y,n}f - \tilde{\pi}_{y,n}f\| = \|u_0(f)(X_0) - \tilde{u}_0(f)(\tilde{X}_0)\|_p \leq \|u_0(f)(X_0) - \tilde{u}_0(f)(\tilde{X}_0)\|_p \leq \sum_{k=0}^{n} C^n_k(f, y, p)\|\Delta_k\|_p
\]
where, for every $0 \leq k \leq n$,
\[
C^n_k(f, y, p) := K_{g}^{k+1}(\alpha_k K_g + \beta_k),
\]
\[
= (2 - \delta_{2,p})K_{g}^{k}[u_k]_{L^{p}} + K_{g}^{n-1}\|f\|_\infty ([g_{k+1}]_{L^{p}} + [g_{k}]_{L^{p}})
\]
\[
\leq K_{g}^{n}(2 - \delta_{2,p})[P]_{L^{p}}^{-k}[f]_{L^{p}}
\]
\[
+ \|f\|_\infty \left( [g_{k+1}]_{L^{p}} + [g_{k}]_{L^{p}} + (2 - \delta_{2,p})\sum_{m=1}^{n-k} [P]_{L^{p}}^{m-1} ([g_{k+m}]_{L^{p}} + [P]_{L^{p}}[g_{k+m}]_{L^{p}}) \right)
\]
One concludes by Lemma 3.1. \qed
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 $\Gamma_k$ 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 $(\hat{X}_k, \hat{Y}_k)_k$ defined by:

$$\hat{X}_k = \text{Proj}_{\Gamma_k} \left( F_k(\hat{X}_{k-1}, \varepsilon_k) \right), \quad k = 1, \ldots, n$$

(4.1)

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

(4.2)

with $\hat{X}_0 = \text{Proj}_{\Gamma_0}(X_0)$ and $\hat{Y}_0 = Y_0 = 0$. Here $\text{Proj}_{\Gamma_k}$ still denotes the closest neighbour projection on $\Gamma_k$. 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 $\hat{Y} = (\hat{Y}_1, \ldots, \hat{Y}_n)$ are fixed to $y = (y_1, \ldots, y_n)$.

Since $\hat{X}_n$ is valued in the finite grid $\Gamma_n$ made 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 $\hat{\Pi}_{y,n} = \mathbb{P}[\hat{X}_n = x_n^i | \hat{Y} = y]$, $i = 1, \ldots, N_n$; 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 $\delta_x$ is the Dirac mass at $x$. By same arguments as in Section 2, using Bayes rule and Markov property of $(\hat{X}_k, \hat{Y}_k)_k$, we have:

$$\hat{\Pi}_{y,n}^i = \frac{\tilde{\pi}_{y,n}^i}{\sum_{i=1}^{N_n} \tilde{\pi}_{y,n}^i}, \quad i = 1, \ldots, N_n,$$

(4.3)

where

$$\tilde{\pi}_{y,n}^i = \mathbb{E} \left[ 1_{\hat{X}_n = x_n^i} \hat{L}_{y,n} \right], \quad i = 1, \ldots, N_n,$$

(4.4)

and

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

(4.5)

From an algorithmic viewpoint, the unnormalized approximate filter $\tilde{\pi}_{y,n}$ may be computed either in a forward or backward induction in view of 1) or 2) of Section 2. We describe here the forward procedure which is less costly in terms of complexity. We denote by $\hat{P}_0 = (\hat{P}_0^i)_{i=1, \ldots, N_0}$ the probability law of $\hat{X}_0$, i.e. $\hat{P}_0^i = \mathbb{P}[\hat{X}_0 = x_0^i]$, $i = 1, \ldots, N_0$, and for $k = 1, \ldots, n$, by $(\hat{P}_k)$ the probability transition matrix of the finite state space Markovian process $(\hat{X}_k)_k$, i.e.

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

(4.6)

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

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

(4.7)

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

$$\tilde{\pi}_{y,0} = \hat{P}_0, \quad \tilde{\pi}_{y,k}^j = \sum_{i=1}^{N_{k-1}} \hat{H}_{y,k}^{ij} \tilde{\pi}_{y,k-1}^i, \quad j = 1, \ldots, N_k, \quad k = 1, \ldots, n.$$  

(4.8)
4.2 Error analysis

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

\[
\Delta_k = F_k(\tilde{X}_{k-1}, \varepsilon_k) - \text{Proj}_{\Gamma_k} \left( F_k(\tilde{X}_{k-1}, \varepsilon_k) \right), \ k \geq 1 \quad (4.9)
\]

\[
\Delta_0 = X_0 - \text{Proj}_{\Gamma_0}(X_0), \ k = 0. \quad (4.10)
\]

We make the following Lipschitz assumptions on the model (1.2)-(1.3).

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

\[
\forall x, \tilde{x} \in \mathbb{R}^d, \quad \| F_k(x, \varepsilon_k) - F_k(\tilde{x}, \varepsilon_k) \|_1 \leq [F_k]_{\text{Lip}} |x - \tilde{x}|.
\]

We then set \([F]_{\text{Lip}} = \max_{k=1,\ldots,n} [F_k]_{\text{Lip}}\).

Comments on assumptions \((A1')\) and \((A1)\): Note that \([P_k]_{\text{Lip}} \leq [F_k]_{\text{Lip}}\). In fact it may happen for some models that \([P_k]_{\text{Lip}} < \infty = [F_k]_{\text{Lip}}\) which means that the field of application of the marginal quantization is wider than that of Markovian quantization, at least theoretically. For example, set

\[
X_{k+1} = \text{sign}(X_k - \varepsilon_{k+1}) G(X_k, \varepsilon_{k+1}),
\]

where \((\varepsilon_k)\) is an i.i.d. sequence, \( P_{\varepsilon_1}(du) = g(u) \lambda_q(du) \) (\( \lambda_q \) lebesgue measure on \( \mathbb{R}^d \)) and \((x, u) \mapsto G(x, u)\) is Lipschitz continuous in \( x \) uniformly in \( u \) with ratio \([G]_{\text{Lip}}\). Then, one easily shows that

\[
[P]_{\text{Lip}} \leq [G]_{\text{Lip}} < +\infty
\]

whereas \( x \mapsto F(x, \varepsilon_1) = \text{sign}(x - \varepsilon_1) G(x, \varepsilon_1)\) is not even continuous so that \((A1')\) is not satisfied.

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 usual discretization scheme of Gaussian diffusions like 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 nondegenerate diffusions.

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

\[
\left| \Pi_{y,n} f - \hat{\Pi}_{y,n} f \right| \leq \frac{K_{\phi_n}(y)}{\phi_n(y)} \sum_{k=0}^n A_{\phi_n}^n(f, y) \| \Delta_k \|_1. \quad (4.11)
\]
with \( \hat{\theta}_n(y) := \tilde{\pi}_{y,n} \mathbf{1} \),

\[
A^n_k(f, y) = [F]_{\text{Lip}}^{n-k}[f]_{\text{Lip}} + 2 \frac{\|f\|_{\infty}}{K_g} [g_k^2]_{\text{Lip}}(y_{k-1}, y_k) + 2 \frac{\|f\|_{\infty}}{K_g} \sum_{j=k+1}^{n} [F]_{\text{Lip}}^{j-k-1} ( [g_j^1]_{\text{Lip}}(y_{j-1}, y_j) + [F]_{\text{Lip}} [g_j^2]_{\text{Lip}}(y_{j-1}, y_j) ) ,
\]

(\text{with the convention that the sum in (4.13) is zero for } k = n).\]

\textbf{Remark 4.2} By introducing \([g]_{\text{Lip}} := \max_{k=1,\ldots,n} \sup_{y,y' \in \mathbb{R}^q} ( [g_k^1]_{\text{Lip}}(y, y') \vee [g_k^2]_{\text{Lip}}(y, y') )\), the \(A^n_k(f, y)\) is bounded by the simpler quantity

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

with the usual convention that \( \frac{1}{u^{-1}}(u^m - 1) = m \) if \( u = 1 \) and \( m \in \mathbb{N} \).

\textbf{Remark 4.3} If the Lipschitz condition (A2)(ii) is weakened into (A2)(ii') like in Section 3 and if (A1) is strengthened into the slightly more stringent condition than \((A1')\),

\textbf{(A1')p} \quad \text{For each } k = 1, \ldots, n, \text{ there exists a positive constant } [F_k]_{\text{Lip}} \text{ such that}

\[
\|F_k(x, \varepsilon_k) - F_k(\hat{x}, \varepsilon_k)\|_p \leq [F_k]_{\text{Lip}} |x - \hat{x}|,
\]

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 \( 1/p + 1/q = 1 \)):

\[
\left| \Pi_{y,n} f - \tilde{\Pi}_{y,n} f \right| \leq \frac{K_g}{\phi_n(y) \vee \phi_n(y)} \sum_{k=0}^{n} \tilde{A}^n_k(q, f) \|\Delta_k\|_p,
\]

with \( \tilde{A}^n_k(q, f) = [F]_{\text{Lip}}^{n-k}[f]_{\text{Lip}} + 2 \frac{\|f\|_{\infty}}{K_g} [g]_{\text{Lip}} \left( \frac{[F]_{\text{Lip}} + 1}{[F]_{\text{Lip}} - 1} ( [F]_{\text{Lip}}^{n-k} - 1 ) + 1 \right) N_n(q) \),

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

Here we have set \([g]_{\text{Lip}} := \max_{k=0,\ldots,n} \sup_{y,y' \in \mathbb{R}^q} ( [g_k^1]_{\text{Lip}} \vee [g_k^2]_{\text{Lip}}(y, y') )\) and \( \|X\|_q = \max_{k=0,\ldots,n} \|X_k\|_q \).

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

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

\[
|L_{y,n} - \tilde{L}_{y,n}| \leq K_g^{n-1} \sum_{k=1}^{n} [g_k^1]_{\text{Lip}}(y_{k-1}, y_k) |X_{k-1} - \hat{X}_{k-1}| + [g_k^2]_{\text{Lip}}(y_{k-1}, y_k) |X_k - \hat{X}_k|.
\]
**Proof.** In order to alleviate notations, 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 \( k = 1, \ldots, n \),

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

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

\[
\left| L_k - \hat{L}_k \right| \leq K_g^{k-1} \left( \|g_k\|_{L_{lip}}(y_{k-1}, y_k) |X_{k-1}-\hat{X}_{k-1}| + \|g_k^2\|_{L_{lip}}(y_{k-1}, y_k) |X_k-\hat{X}_k| \right) + K_g \left| L_{k-1} - \hat{L}_{k-1} \right|.
\]

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

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

\[
\|X_k - \hat{X}_k\|_1 \leq \sum_{j=0}^{k} [F]_{L_{lip}}^{k-j} \|\Delta_j\|_1.
\]

**Proof.** From the definitions (1.2) and (4.1) of \( X_k \) and \( \hat{X}_k \), and (4.9) of \( \Delta_k \), we obviously get for each \( k \geq 1 \):

\[
\|X_k - \hat{X}_k\|_1 \leq \left\| F_k(X_{k-1}, \varepsilon_k) - F_k(\hat{X}_{k-1}, \varepsilon_k) \right\|_1 + \|\Delta_k\|_1.
\]

By Assumption (A1’) and since \( \varepsilon_k \) is independent of \( X_{k-1} \) and \( \hat{X}_{k-1} \), we then obtain:

\[
\|X_k - \hat{X}_k\|_1 \leq [F_k]_{L_{lip}}\|X_{k-1} - \hat{X}_{k-1}\|_1 + \|\Delta_k\|_1.
\]

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

**Proof of Theorem 4.1:** From expressions (2.3) and (4.4), one derives that

\[
|\pi_{y,n} f - \hat{\pi}_{y,n} f| = \left| \mathbb{E}[f(X_n)\hat{L}_{y,n}] - \mathbb{E}[f(\hat{X}_n)\hat{L}_{y,n}] \right|
\leq \|f\|_{\infty} \|L_{y,n} - \hat{L}_{y,n}\| + [f]_{L_{lip}} \mathbb{E}[|X_n - \hat{X}_n|\hat{L}_{y,n}]
\leq \|f\|_{\infty} \|L_{y,n} - \hat{L}_{y,n}\| + [f]_{L_{lip}} K_g^n \|X_n - \hat{X}_n\|_1.
\]

Lemmata 3.1, 4.1 and 4.2 complete the proof. \( \square \)

## 5 Convergence of the quantized filters. Optimization

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

\[
|\Pi_{y,n} f - \hat{\Pi}_{y,n} f| \leq \|f\|_{\infty} \vee [f]_{L_{lip}} K_g^n \sum_{k=0}^{n} D^p_k(y, p) \|\Delta_k\|_p,
\]

for all \( p \in [1, \infty) \), where \( \Delta_k = Z_k - \text{Proj}_{\Gamma_k}(Z_k) \) is the difference between a simulated random variable, say \( Z_k \) and its projection \( \text{Proj}_{\Gamma_k}(Z_k) \) following the closest neighbour rule.
onto the grid $\Gamma_k$ 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

$$D^n_k(y, p) = \begin{cases} B^n_k(f_0, y, p) & \text{for the marginal quantization,} \\ A^n_k(f_0, y) & \text{for the Markovian quantization} \end{cases}$$

where $f_0(x) = \frac{|x|}{1 + |x|}$ (so that $\|f_0\|_\infty = [f_0]_{Lip} = 1$). If one assumes that, at every time $k = 0, 1, \ldots, n$, the grid $\Gamma_k$ is $L^p$-optimal i.e. minimizes the mean $L^p$-quantization error among all grids with size $N_k$, then the asymptotic of optimal quantization stated in Zador’s Theorem 8.1 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

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

so that

$$|\Pi_{y,n}f - \hat{\Pi}_{y,n}f| \leq \|f\|_\infty \vee \begin{cases} [f]_{Lip} K^n_k \sum_{k=0}^n \theta_k D^n_k(y, p)N_k^{-1/d} \\ \end{cases}$$

(5.1)

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

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

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

(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, like 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 replicate it at every time produces better results.

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

$$dX_t = b(X_t)dt + \sigma(X_t)dW_t,$$

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 $\sigma$ are Lipschitz, condition (A1’) is satisfied with:

$$[F]_{Lip} = 1 + \frac{c}{n},$$
for some positive constant independent of $n$. We 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 $N_k = \tilde{N} := N/(n + 1)$ points at each grid $\Gamma_k, k = 0, \ldots, n,$ (5.1) provides a rate of convergence for the approximate filters of order:

$$\frac{K_g^n}{\phi_n(y)} \frac{n + 1}{\tilde{N}^{\frac{1}{2}}}.$$ 

This has to be compared with the rate of convergence obtained by the particle Monte-Carlo methods using $\tilde{N}$ interacting particles, see [Del Moral et al. 2001]:

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

### 6 Some facts about practical implementation

#### 6.1 General features, complexity

At this stage, it is important to mention when and how a quantization method can be implemented. First, one must keep 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 to implement the quantization approach is to assume that the probabilistic features of the state process $(X_k)$ do not change too fast and/or too often. This is not a real restriction when one thinks about some usual applications of filtering problems (surge of the sea, satellite pursuit, 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 details the three phases of the filter approximation.

**Phase I (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 number of $N$ points. This means

- specifying the sizes $N_k$ of the grids $\Gamma_k, 1 \leq k \leq n$ (*a priori* dispatching).

- optimizing the every grid $\Gamma_k := \{x^i_k, 1 \leq i \leq N_k\}, 1 \leq k \leq n$ that is solving the optimization problem

$$\min_{\|\Gamma_k\| \leq N_k} \|X_k - \text{Proj}_{\Gamma_k}(X_k)\|_2$$

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

The dispatching is made *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 are carried out either simultaneously or by a new Monte Carlo simulation (see [Bally and Pagès 2003]). They have been extensively investigated in former papers and we refer to them for a precise description of the procedure (see e.g. [Bally and Pagès 2003], [Pagès et al. 2004b]). This optimization phase is computationally the most demanding. This means less than 10 minutes of CPU time to compute the whole quantization tree (“height” \( n = 20 \), size \( N = 20000 \)) of a(n asymptotically) non-stationary 4-dimensional process using a 1 GHz micro-processor.

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 below), which drastically cuts down 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 \( N(0; I_d) \) (see [Pagès and Printems 2003], the files are available at the URL – www.proba.jussieu.fr/pageperso/pages.html or – www.univ-paris12.fr/www/labos/cmup/homepages/printems).

The crucial fact is that 

\[
\text{the optimization phase does not depend on the observations}
\]

which explains why its results can be kept off-line.

**Phase II (Computation of the quantized filter distribution):** One computes the weight vector \((\hat{\pi}_j^{y,n})_{1 \leq j \leq N_n}\) by plugging the observation vector \( y = (y_1, \ldots, y_n) \) and the transition weights \( \hat{P}_{ij}^{k} \) into the forward representations of the approximate filter i.e. (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 \( \frac{nN^2}{(n+1)^2} \approx \frac{N^2}{n} \) (when \( N_k = \frac{N}{n+1}, k = 0, \ldots, n \)). In practice, many transitions are 0 (when \( x_{k-1}^i \) and \( x_k^j \) are remote) and every knot \( x_k^i \) of the tree has approximately the same number \( \nu \) of “active connections” with knots of time \( k + 1 \). So, the resulting complexity, after an appropriate pruning of the quantization tree, is approximately \( \nu \times n \times \bar{N} \) where \( \bar{N} = N/(n + 1) \) is the average number of points per time step. In all our numerical experiments, this phase is almost instantaneous (less than \( \frac{1}{10} \) second with \( N = 20000 \) points in dimension \( d = 4 \) with the same 1 GHz micro-processor).

This distribution approximation phase does not depend on the function \( f \).

**Phase III (Computation of \( \hat{\Pi}_{y,n}f \)):** one computes for every (needed) function \( f \)

\[
\int f(x)\hat{\Pi}_{y,n}(dx) = \sum_{i=1}^{N_n} f(x_i^n)\hat{\Pi}_{n}^i
\]

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

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

It may happen for some observation vectors that the optimal filter and the prior distribution of the process $X$ assign some masses at 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 of interest: marginal quantization of a stationary signal

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

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

Then $\hat{\Gamma}_k := \hat{\Gamma}$, $0 \leq k \leq n$ make up the optimal quantization of the chain. The companion parameters are

- the quantization of the distribution $\nu$ induced by $\hat{\Gamma}$, that is $\hat{P}_0 = \text{Proj}_{\hat{\Gamma}}(X_0)$
- and a single transition matrix

$$\hat{P}^{ij}_k = \hat{P}^{ij}_1 := \mathbb{P}(\hat{X}_1 = \hat{x}^j \mid \hat{X}_0 = \hat{x}^i), \quad 0 \leq i, j \leq \hat{N}.$$ 

The size of the parameters to be stored is obviously divided by a factor $n$ (or the possible quantization size for the distribution $\nu$ 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 in the form:

$$dX_t = b(X_t)dt + \sigma(X_t)dW_t, \quad X_0 \sim \mu \quad (7.1)$$
$$dY_t = \beta(X_t, Y_t)dt + \gamma(X_t, Y_t)dB_t, \quad Y_0 = 0, \quad (7.2)$$

where $W$ is a $d$-dimensional Brownian motion independent of the $q$-dimensional Brownian motion $B$, $\mu$ is a known distribution, and $b$, $\beta$, $\sigma$, $\gamma$ are known functions. We set $\Sigma(x) = \sigma(x)\sigma(x)^*$ and $\Lambda(x, y) = \gamma(x, y)\gamma(x, y)^*$. We assume that $x \mapsto \Sigma(x)$ and $(x, y) \mapsto \Lambda(x, y)$ are uniformly nondegenerate functions and we denote by $x \mapsto \Sigma^1(x)$ and $(x, y) \mapsto \Lambda^1(x, y)$ their square roots functions\(^{(1)}\) which are clearly uniformly nondegenerate. The process $(X, Y)$ is Markov with a transition semi-group denoted by $(R_t)_t$.

\(^{(1)}\)Every nonnegative symmetric $S$ matrix admits a unique square root $S^{1/2}$ satisfying $S^{1/2}$ is nonnegative, symmetric, $S^{1/2} S^{1/2} = S$ and $S^{1/2}$ commutes with $S$.\n
20
We consider 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 \((Y_1, \ldots, Y_n)\) settled 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 chain 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 \(b, \sigma, \gamma\) 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 \((H)\) of the introduction: the law of \(Y_k\) conditional on \((X_{k-1}, Y_{k-1}, X_k) = (x, y, x')\) admits 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 closely follow here the arguments of [Del Moral et al. 2001]. More precisely, 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

\[
X(x)^{(m)}_0 = x,
X(x)^{(m)}_{i+1} = X(x)^{(m)}_i + b(X(x)^{(m)}_i) \frac{1}{m} + \sigma(X(x)^{(m)}_i) \frac{\varepsilon_{i+1}}{\sqrt{m}},
\]

\[
Y(x, y)^{(m)}_0 = y,
Y(x, y)^{(m)}_{i+1} = Y(x, y)^{(m)}_i + \beta(Y(x, y)^{(m)}_i, Y(x, y)^{(m)}_i) \frac{1}{m} + \gamma(Y(x, y)^{(m)}_i, Y(x, y)^{(m)}_i) \frac{\eta_{i+1}}{\sqrt{m}},
\]

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

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

\[
\psi(x, y, y') = \frac{m^{\frac{d}{2}}}{(2\pi)^{\frac{d}{2}} \text{det}(\Sigma^y(x, y))} \exp \left[ -\frac{m}{2} \left( \Sigma^y(x, y) \right)^{-1} \left( y' - y - \frac{\beta(x, y)}{m} \right)^2 \right].
\]

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 \( b, \beta, \sigma, \gamma \) such that:

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

\[
    |x - x'| + |y - y'| > \frac{2}{m} \quad \Rightarrow \quad |r(x, y, x', y') - r^{(m)}(x, y, x', y')| \leq \frac{C}{m} \exp\left(-C(|x - x'|^2 + |y - y'|^2)\right). \tag{7.4}
\]

The law \( P^{(m)}(x, dx') \) of \( X(x)^m \) has a density \( x' \to p^{(m)}(x, x') = \int r^{(m)}(x, y, x')dy' \). We have then 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 \( \tilde{\Pi}_{y,n}^{(m)} \) defined by the marginal quantization algorithm in Section 3 where we replace the unknown function \( g \) by \( g^{(m)} \). The estimation error is measured through

\[
    \left| \Pi_{y,n}f - \tilde{\Pi}_{y,n}^{(m)} f \right| \leq \left| \Pi_{y,n}f - \Pi_{y,n}^{(m)} f \right| + \left| \Pi_{y,n}^{(m)} f - \tilde{\Pi}_{y,n}^{(m)} f \right| , \tag{7.6}
\]

where \( \Pi_{y,n}^{(m)} \) is the filter given by the formulas (2.1)- (2.2), with the transition probability distribution \( P(x, dx') = p(x, x')dx' \) replaced by \( P^{(m)}(x, dx') = p^{(m)}(x, x')dx' \) 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:

\[
    \Pi_{y,0} = \mu, \quad \Pi_{y,k} = \frac{\Pi_{y,k-1}H_{y,k}f}{\Pi_{y,k-1}H_{y,k}1}, \quad k = 1, \ldots, n,
\]

with

\[
    H_{y,k}f(x) = \int f(x')g(x, y_{k-1}, x', y_k)P(x, dx') = \int f(x')r(x, y_{k-1}, x', y_k)dx'.
\]

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

\[
    \Pi_{y,0}^{(m)} = \mu, \quad \Pi_{y,k}^{(m)} = \frac{\Pi_{y,k-1}^{(m)}H_{y,k}^{(m)}f}{\Pi_{y,k-1}^{(m)}H_{y,k}^{(m)}1}, \quad k = 1, \ldots, n,
\]

with

\[
    H_{y,k}^{(m)}f(x) = \int f(x')g^{(m)}(x, y_{k-1}, x', y_k)p^{(m)}(x, dx') = \int f(x')r^{(m)}(x, y_{k-1}, x', y_k)dx'.
\]

By using estimations (7.3)-(7.4), it is easily checked that for any bounded function \( f \),

\[
    \left\| H_{y,k}f - H_{y,k}^{(m)} f \right\|_{\infty} \leq \frac{C}{m} \left\| f \right\|_{\infty},
\]

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:

\[
    \left| \Pi_{y,n}f - \Pi_{y,n}^{(m)} f \right| \leq \frac{C}{m} \left\| f \right\|_{\infty} \rho_n(y)^{n+1} \frac{\rho_n(y)\phi_n(y) - \rho_n(y)}{K_g\rho_n(y) - 1},
\]

with \( \rho_n(y) = 2K^n_g/\phi_n(y) \). The second term in (7.6) is given by estimation result as in Theorem 3.1 provided one can check some Lipschitz condition for \( g^{(m)} \). Actually, it is
proved in the Appendix that when the functions $b$, $\sigma$ and $\gamma$ 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 $y, y' \in \mathbb{R}^q$,

\[ g^{(m)}(x, y, x', y') - g^{(m)}(\hat{x}, y, \hat{x}', y') \leq Cm^{2+1} \left( 1 + |x| + |x'| + |\hat{x}| + |\hat{x}'| \right) |x - \hat{x}| \]

\[ + Cm^{2+3} \left( 1 + |x| + |x'| + |\hat{x}| + |\hat{x}'| \right) |x' - \hat{x}'|. \]  

8.1 The Kalman-Bucy model

\[ X_k = AX_{k-1} + \tau \varepsilon_k \quad \in \mathbb{R}^d \]  

\[ Y_k = BX_k + \theta \eta_k \quad \in \mathbb{R}^q \]  

for $k \in \mathbb{N}$, and $X_0$ is normally distributed with mean $m_0 = 0$ and covariance matrix $\Sigma_0^2$. Here $A$, $B$, $\tau$ and $\theta$ are matrices of appropriate dimensions, and $(\varepsilon_k)_{k \geq 1}$, $(\eta_k)_{k \geq 1}$ are independent centered Gaussian processes, $\varepsilon_k \sim N(0, I_d)$, $\eta_k \sim N(0, I_d)$. In this case, we have

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

If $\|A\| < 1$, then $(X_k)_{k \geq 0}$ is stationary if $\Sigma_0^2 = \sum_{n \geq 0} (A^n \tau')(A^n \tau')'$ (unique solution of $\Sigma_0^2 = A \Sigma_0^2 A' + \tau' \tau$ where $M'$ denotes the transpose of $M$). Of course, the filter $\Pi_{y,n}$ is explicitly known, see e.g. [Elliott et al. 1995]: it is a Gaussian distribution of mean $m_n$ and covariance matrix $C_n$ given by the inductive equations:

\[ C_{k+1} = (I_d - K_{k+1}B)(\tau' + AC_k A'), \quad C_0 := 0, \]

\[ m_{k+1} = Am_k + K_{k+1}(y_{k+1} - BA m_k), \quad m_0 := 0, \]

where

\[ K_{k+1} = (\tau \tau' + AC_k A' B')(B(\tau' + AC_k A' B' + \theta \theta')^{-1}. \]

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 $\Delta t$ of the following (linear) Gaussian diffusion system

\[ dX(t) = -\alpha X(t)dt + \sigma_X dW^X(t) \]

\[ dY(t) = B dX(t) + \sigma_Y dW^Y(t) \quad \text{with} \quad < W^Y, W^Y > = 0 \]

if one sets $A = I_d - \Delta t \alpha$, $\tau = \sqrt{\Delta t} \sigma_X$, $\theta = \sqrt{\Delta t} \sigma_Y$.

The numerical experiments have 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 $N(0; \Sigma_0^2)$. Instead, we selected some grids of the form

\[ \Gamma_0 = \Sigma_0 \Gamma^* := \{ \Sigma_0 \xi, \xi \in \Gamma^* \} \]
where $\Gamma^*$ 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 interest of keeping off-line some tabulations.

Two kinds of tests have been carried out with the Kalman-Bucy filter in order to track the behaviour of the quantized filter as a function of $N$ (or $\tilde{N}$) and $n$. The choice of a stationary setting is motivated by the possibility to detect 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 with (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 $\tilde{N} = N/(n+1)$ of the grid $\Gamma_0$. We set $\Delta t = 1/250$, $n = 15$ and considered three functions

$$ f_1(x) = x^d, \quad f_2(x) = |x|^2, \quad f_3(x) = \exp(-|x|) $$

that we implemented in dimensions $d = 1$ and $d = 3$ (closed forms exist for $\Pi_n f_i$, $i = 1, 2, 3$). In both dimensions the results are summed up in a figure showing for every function $f_i$, $i = 1, 2, 3$, the graphs $\tilde{N} \mapsto \Pi_n(f_i)$ (or $\tilde{N} \mapsto |\Pi_n(f_i) - \Pi_n(f_i)|$) and $\log \tilde{N} \mapsto \log|\Pi_n(f_i) - \Pi_n(f_i)|$ (i.e. a log-scale) with its least square regression line denoted by “$y = -ax + b$” ($a$ and $b$ appearing as numerical values). This means that

$$ |\Pi_n(f_i) - \Pi_n(f_i)| \approx \frac{e^b}{N^a}.$$ 

- $d = 1$: $\alpha = B = 1$, $\sigma^X = 0.5$ and $\sigma^Y = 1$ so that $A = 0.996$, $\tau = 0.0316$ (so that $\Sigma_0 = \frac{1}{\sqrt{1-A}} \approx 0.354$) and $\theta = 0.0663$. The grid size $\tilde{N}$ runs in the interval $[50, 400]$. Figure 1 contains the results. Furthermore, we added in Figure 2, for every function $f_i$, $i = 1, 2, 3$, a graph $\tilde{N} \mapsto |\Pi_n(f_i) - \Pi_n(f_i)|$ for three different observation vectors.

- $d = 3$: 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 $A = \begin{bmatrix} 0.9942 & -0.0022 & -0.0031 \\ -0.0022 & 0.9962 & -0.0009 \\ -0.00311 & -0.0009 & 0.9936 \end{bmatrix}$ ;

set $\tau = \begin{bmatrix} 0.0179 & 0.0317 & 0.0444 \\ 0.0317 & 0.0793 & 0.0127 \\ 0.0444 & 0.0127 & 0.1173 \end{bmatrix}$ so that $\Sigma_0 = \begin{bmatrix} 0.0900 & 0.0219 & 0.0449 \\ 0.0189 & 0.0449 & 1.0343 \end{bmatrix}$. Then, set $B = I_3$, $\theta = 0.5 I_3$. The grid size $\tilde{N}$ runs in the interval $[50, 600]$. Figure 3 contains the results.

**Test 2:** Stability of the filter for a fixed grid size $\tilde{N} = \tilde{N}_{\text{max}}$ as $n$ grows. We consider now a $d = 2$-dimensional model with $\alpha = \begin{bmatrix} 1.1625 & -0.8488 \\ -0.8488 & 1.5875 \end{bmatrix}$ so that $A := \begin{bmatrix} 0.99535 & 0.00340 \\ 0.003340 & 0.99365 \end{bmatrix}$, $\tau := \begin{bmatrix} 0.08830 & -0.02419 \\ -0.02419 & 0.10041 \end{bmatrix}$, $B = I_2$, $\theta := 0.5 I_2$ (so that $\Sigma_0 = \begin{bmatrix} 1.01976 & 0.10041 \\ 0.10041 & 0.969493 \end{bmatrix}$). We set $\tilde{N} = 600$ with $n$ running from 1 up to 100. In Figure 4(left) are depicted respectively

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

24
and the linear regression line of these relative errors. These results are much more satisfactory than those induced by the a posteriori errors 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 like in Figure 4(right), the true value and the quantized one become indistinguishable. This means that, like for interacting particle methods, the mixing property of the state variable \( X \) induces the stability of the filter as \( n \) increases. However, we have not yet theoretical results to support this fact.

8.2 A stochastic volatility model:

We consider a state model with multiplicative Gaussian noise process:

\[
Y_k = \sigma(X_k)\eta_k \in \mathbb{R} \quad \text{with} \quad X_k = \rho X_{k-1} + \varepsilon_k \in \mathbb{R} \quad (8.10)
\]

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

\[
dX(t) = -\alpha X(t) dt + \tau dW(t), \quad 0 \leq t \leq 1,
\]

with positive parameters \( \lambda \) and \( \tau \). We then suppose

\[
\rho = 1 - \alpha \Delta t, \quad \varepsilon_k \sim \mathcal{N}(0, \tau^2 \Delta t) \quad \text{and} \quad \eta_k \sim \mathcal{N}(0, \Delta t).
\]

The filtering problem consists in 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 \( (X_k)_{k \geq 1} \) is stationary i.e. \( X_0 \sim \mathcal{N}(0, \Sigma_0^2) \) with \( \Sigma_0 = \tau \sqrt{\Delta t/(1 - \rho^2)} = \tau / \sqrt{\alpha(2 - \alpha \Delta t)} \approx 0.354 \).

The selected model is here:

\[
(ABS) \equiv \sigma(X_k) = \gamma + |X_k| \quad \text{with} \quad \gamma = 0.05.
\]

In Figure 5 is depicted the graph \( \bar{N} \mapsto |\Pi_n(f_i) - \hat{\Pi}_n(f_i)| \) which strongly suggests convergence for the three functions (although we have no reference value at hand).

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

We consider the particular case where the coefficients \( b, \sigma \) and \( \gamma \) of the diffusion \( (X, Y) \) in (7.1)-(7.2) are real constants. We then assume w.l.o.g. \( \sigma = I_d \) and \( \gamma = I_q \). We also assume that the function \( \beta \) is bounded and differentiable with bounded derivatives. We show in this appendix 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

\[
r^{(m)}(x_0, y_0, x_m, y_m) = \int \prod_{i=0}^{m-1} \phi(x_i, x_{i+1}) \psi(x_i, y_i, y_{i+1}) \, dx_1 \ldots dx_{m-1} dy_1 \ldots dy_{m-1},
\]

\[
p^{(m)}(x_0, x_m) = \int \prod_{i=0}^{m-1} \phi(x_i, x_{i+1}) \, dx_1 \ldots dx_{m-1},
\]

and

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

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

\[
g^{(m)} \leq \left( \frac{m}{2\pi} \right)^\frac{d}{2}. \tag{A.2}
\]

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:

\[
\frac{\partial r^{(m)}}{\partial x_0} = \int \left[ m I_d \left( x_1 - x_0 - \frac{b}{m} \right) + \frac{\partial \beta}{\partial x_0} \left( y_1 - y_0 - \frac{\beta(x_0, y_0)}{m} \right) \right] \prod_{i=0}^{m-1} \phi(x_i, x_{i+1}) \psi(x_i, y_i, y_{i+1}) \, dx_1 \ldots dx_{m-1} dy_1 \ldots dy_{m-1},
\]

\[
\frac{\partial r^{(m)}}{\partial x_m} = \int \left[ -m I_d \left( x_m - x_{m-1} - \frac{b}{m} \right) \right] \prod_{i=0}^{m-1} \phi(x_i, x_{i+1}) \psi(x_i, y_i, y_{i+1}) \, dx_1 \ldots dx_{m-1} dy_1 \ldots dy_{m-1},
\]

\[
\frac{\partial p^{(m)}}{\partial x_0} = \int \left[ m I_d \left( x_1 - x_0 - \frac{b}{m} \right) \right] \prod_{i=0}^{m-1} \phi(x_i, x_{i+1}) \, dx_1 \ldots dx_{m-1},
\]

\[
\frac{\partial p^{(m)}}{\partial x_m} = \int \left[ -m I_d \left( x_m - x_{m-1} - \frac{b}{m} \right) \right] \prod_{i=0}^{m-1} \phi(x_i, x_{i+1}) \, dx_1 \ldots dx_{m-1}.
\]

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

\[
\left| \frac{\partial r^{(m)}}{\partial x_0} \right| \leq C m^\frac{d}{2} \int \left( m \left| x_1 - x_0 - \frac{b}{m} \right| + 1 \right) \prod_{i=0}^{m-1} \phi(x_i, x_{i+1}) \, dx_1 \ldots dx_{m-1},
\]

for some positive constant \( C \). Hence,

\[
\left| \frac{\partial r^{(m)}}{\partial x_0} \right| \leq C m^\frac{d}{2} \left( 1 + mB(x_0, x_m) \right), \tag{A.3}
\]

where

\[
B(x_0, x_m) = \frac{\int \left| x_1 - x_0 - \frac{b}{m} \right| \prod_{i=0}^{m-1} \phi(x_i, x_{i+1}) \, dx_1 \ldots dx_{m-1}}{\int \prod_{i=0}^{m-1} \phi(x_i, x_{i+1}) \, dx_1 \ldots dx_{m-1}}.
\]
By making the change of variables, $x_i \to x_i - x_{i-1} - b/m$, $i = 1, \ldots, m-1$, we have $B(x_0, x_m) = B(x_m - x_0 - b)$ with

$$B(x) = \frac{\int |x_1| \exp \left[ -\frac{m}{2} \left( \sum_{i=1}^{m-1} |x_i|^2 + \sum_{i=1}^{m-1} x_i - x \right)^2 \right] dx_1 \ldots dx_{m-1}}{\int \exp \left[ -\frac{m}{2} \left( \sum_{i=1}^{m-1} |x_i|^2 + \sum_{i=1}^{m-1} x_i - x \right)^2 \right] dx_1 \ldots dx_{m-1}}. \quad \text{(A.4)}$$

By writing the sum in parentheses in the previous relation as a canonical square sum in $x_{m-1}$, i.e.

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

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

$$B(x) = \frac{\int |x_1| \exp \left[ -\frac{m}{2} \left( \sum_{i=1}^{m-2} |x_i|^2 + \frac{1}{2} \sum_{i=1}^{m-2} x_i - x \right)^2 \right] dx_1 \ldots dx_{m-2}}{\int \exp \left[ -\frac{m}{2} \left( \sum_{i=1}^{m-2} |x_i|^2 + \frac{1}{2} \sum_{i=1}^{m-2} x_i - x \right)^2 \right] dx_1 \ldots dx_{m-2}}.$$

By induction, this yields

$$B(x) = \frac{\int |x_1| \exp \left[ -\frac{m}{2} \left( |x_1|^2 + \frac{1}{m-1} |x_1 - x|^2 \right) \right] dx_1}{\int \exp \left[ -\frac{m}{2} \left( |x_1|^2 + \frac{1}{m-1} |x_1 - x|^2 \right) \right] dx_1}.$$

Using again canonical square sum in $x_1$, we then get

$$B(x) = \frac{\int |x_1| \exp \left( -\frac{m^2}{2(m-1)} |x_1 - \frac{x}{m}|^2 \right) dx_1}{\int \exp \left( -\frac{m^2}{2(m-1)} |x_1 - \frac{x}{m}|^2 \right) dx_1}.$$

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

$$B(x) \leq \frac{C}{m} (\sqrt{m} + |x_0|, \forall x \in \mathbb{R}^d),$$

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

$$\left| \frac{\partial g^{(m)}}{\partial x_0} \right| \left| p^{(m)} \right| \leq C m^{\frac{d}{2}} \left( \sqrt{m} + |x_0| + |x_m| \right).$$

By same arguments as above, we have

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

Therefore,

$$\left| \frac{\partial g^{(m)}}{\partial x_0} \right| \leq C m^{\frac{d}{2}} \left( \sqrt{m} + |x_0| + |x_m| \right). \quad \text{(A.5)}$$

By same arguments as above, we also show that

$$\left| \frac{\partial g^{(m)}}{\partial x_m} \right| \leq C m^{\frac{d}{2}+1} \left( \sqrt{m} + |x_0| + |x_m| \right). \quad \text{(A.6)}$$

The local Lipschitz assumption (7.7) straightforwardly follows from (A.5)-(A.6) \hspace{1cm} \Box
Appendix B: Optimal quantization (numerical aspects)

As mentioned in the introduction, quantization consists in replacing a \( \mathbb{R}^d \)-valued random vector \( X \) by its projection following a closest neighbour rule onto a grid \( \Gamma \subset \mathbb{R}^d \), \( \hat{X}^\Gamma = \text{Proj}_\Gamma(X) \). 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 Voronoi tesselation of \( \Gamma \)) satisfying \( C_i(\Gamma) \subset \{ \xi \in \mathbb{R}^d : |\xi - x^i| = \min_{x \in C_i} |\xi - x| \} \), \( i = 1, \ldots, N \), where \( |\cdot| \) denotes the usual canonical Euclidean norm. We then set

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

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

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

So, if one identifies a grid \( \Gamma \) of size \( N \) with the \( N \)-tuple \( (x^1, \ldots, x^N) \) or any of its permutation, the \( p^{\text{the}} \) power of the \( L^p \)-quantization error – called \( L^p\)-distortion – appears as a symmetric function

\[
Q_N^p(x^1, \ldots, x^N) := \int_\mathbb{R} \min_{1 \leq i \leq N} |\xi - x^i|^p \mathbb{P}_X(d\xi)
\]

which van obviously be defined on the whole \( (\mathbb{R}^d)^N \). The function \( \sqrt{Q_N^p} \) is Lipschitz continuous and does reach a minimum. If \( |X(\Omega)| \) is infinite, then any \( N \)-tuple that achieves the minimum has pairwise distinct and this minimum decreases toward 0 as \( N \) goes to infinity. Its rate of convergence is ruled by the so-called Zador’s theorem.

**Theorem 8.1** (see [Graf and Luschgy 2000]) Assume that \( \mathbb{E}|X|^{p+\varepsilon} < +\infty \) for some \( \varepsilon > 0 \). Then

\[
\lim_{N \to \infty} \left( N^{\frac{1}{d}} \min_{|\Gamma| \leq N} \|X - \hat{X}^\Gamma\|_p \right) = \tilde{J}_{p,d} \left( \int_{\mathbb{R}^d} \varphi(\xi)\frac{d\nu(\xi)}{\nu(\xi)} d\xi \right)^{\frac{1}{d}} = \tilde{J}_{p,d} \tag{B.3}
\]

where \( \mathbb{P}_X(d\xi) = \varphi(\xi) \lambda_d(d\xi) + \nu(d\xi) \), \( \nu \perp \lambda_d \) (\( \lambda_d \) 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 1 dimension (\( \tilde{J}_{p,1} = \frac{1}{2(p+1)^\frac{1}{2}} \)), \( \tilde{J}_{2,2} = \sqrt{\frac{2}{18\sqrt{\pi}}} \ldots \) the true value of \( \tilde{J}_{p,d} \) is unknown. However, \( \tilde{J}_{p,d} \sim \left( \frac{d}{2\pi} \right)^\frac{1}{2} \) as \( d \) goes to infinity (see [Graf and Luschgy 2000]). This theorem says that \( \min_{|\Gamma| \leq N} \|X - \hat{X}^\Gamma\|_p \sim \theta_{X,p,d}N^{-\frac{1}{2}} \). This is in accordance with the rates \( O(N^{-1/d}) \) obtained in numerical integration with uniform grid methods although optimal quantizers are never uniform grids (except for the uniform distribution in 1-dimension): optimal quantization provides the “best fitting” grid for a given distribution \( \mu = \mathbb{P}_X \). Such grids correspond to the real constant \( \theta_{X,p,d} \).

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

\[
q_N^p(x, \xi) := \min_{1 \leq i \leq N} |x^i - \xi|^p.
\]

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 shows (see [Pagès 1997]) that, if \( p > 1 \), \( Q_N^p \) 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 $\nabla Q^p_N(x)$ is obtained by formal differentiation:

$$\nabla Q^p_N(x) = \mathbb{E} \left[ \nabla_x q^p_N(x, X) \right],$$

where

$$\nabla_x q^p_N(x, \xi) = \frac{\partial q^p_N}{\partial x}(x, \xi)_{1 \leq i \leq N} := p \left( \frac{x^i - \xi}{|x^i - \xi|} |x^i - \xi|^{p-1} 1_{C_i(x)}(\xi) \right)_{1 \leq i \leq N} \quad (B.4)$$

with the convention $\frac{0}{0} = 0$. Note that then, $\nabla_x q^p_N(x, \xi)$ has exactly one non-zero component 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 $\Gamma^0$ with $N$ pairwise distinct components) defined by

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

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

$$\sum_s \delta_s = +\infty \quad \text{and} \quad \sum_s \delta^2_s < +\infty. \quad (B.6)$$

Note that (B.5) a.s. grants by induction that $\Gamma^s$ has pairwise distinct components. Under some appropriate assumptions, such a stochastic descent procedure a.s. 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, 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 (CLVQ) algorithm.

**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 $\text{argmin} Q^2_N$ — even in $\text{argminloc} Q^2_N$ if $\mathbb{P}_X$ weights no hyperplane —, any such $N$-tuple is a stationary quantizer i.e. $\nabla Q^2_N(x) = 0$. Standard computations then show that this equation reads

$$\hat{X} = \mathbb{E}[X | \hat{X}] \quad (B.7)$$

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

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

**References**


Figure 1: $d = 1$, $n = 15$. Convergence and convergence rate (in a log-scale) as $\bar{N}$ grows
Figure 2: $d = 1$, $n = 15$. Errors and convergence rate as $\bar{N}$ grows for three different observation vectors.
Figure 3: $d = 3, n = 15$. Convergence and convergence rate (in a log-scale) as a function of $\bar{N}$.
Figure 4: First moment error and relative error $\bar{N} = 600$, $1 \leq n \leq 100$, 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($\rho, \theta, \gamma$) = (0.996, 0.0316, 0.05) as $\bar{N}$ grows