

## MLE FOR PARTIALLY OBSERVED DIFFUSIONS: DIRECT MAXIMIZATION VS. THE EM ALGORITHM

Fabien CAMPILLO and François LE GLAND

*INRIA Sophia-Antipolis, F-06565 Valbonne, France*

Received 26 July 1988

Revised 22 February 1989

Two algorithms are compared for maximizing the likelihood function associated with parameter estimation in partially observed diffusion processes:

- the EM algorithm, investigated by Dembo and Zeitouni (1986), an iterative algorithm where, at each iteration, an auxiliary function is computed and maximized;
- the direct approach where the likelihood function itself is computed and maximized.

This yields to a comparison of nonlinear smoothing and nonlinear filtering for computing a class of conditional expectations related to the problem of estimation. In particular, it is shown that smoothing is indeed necessary for the EM algorithm approach to be efficient.

Time discretization schemes for the stochastic PDE's involved in the algorithms are given, and the link with the discrete time case (hidden Markov model) is explored.

Numerical results are presented with the conclusion that direct maximization should be preferred whenever some noise covariances associated with the parameters to be estimated are small.

parameter estimation \* maximum likelihood \* EM algorithm \* diffusion processes \* nonlinear filtering \* nonlinear smoothing \* Skorokhod integral \* time discretization

### 1. Introduction

The EM algorithm is an iterative algorithm for maximizing the likelihood function, in a context of partial information (Dempster, Laird and Rubin, 1977). Indeed, let  $\{P_\theta, \theta \in \Theta\}$  be a family of mutually absolutely continuous probability measures on a measurable space  $(\Omega, \mathcal{F})$ , with  $P_\theta \sim P'$  and let  $\mathcal{Y} \subset \mathcal{F}$  be the  $\sigma$ -algebra containing all the available information. Then, the log-likelihood function for the estimation of the parameter  $\theta$  can be defined as

$$l(\theta) \triangleq \log E' \left( \frac{dP_\theta}{dP'} \middle| \mathcal{Y} \right), \tag{1.1}$$

and the MLE (maximum likelihood estimate) as

$$\hat{\theta} \in \operatorname{argmax}_{\theta \in \Theta} l(\theta).$$

Research partially supported by USACCE under Contract DAJA45-87-M-0296, and by CNRS-GRECO "Traitement du Signal et Image".

The EM algorithm is based on the following straightforward application of Jensen's inequality

$$l(\theta) - l(\theta') = \log \mathbf{E}_{\theta'} \left( \frac{dP_{\theta}}{dP_{\theta'}} \middle| \mathcal{Y} \right) \geq \mathbf{E}_{\theta'} \left( \log \frac{dP_{\theta}}{dP_{\theta'}} \middle| \mathcal{Y} \right) \triangleq Q(\theta, \theta'), \quad (1.2)$$

which says that, for each value  $\theta'$  of the parameter, the log-likelihood function  $l(\theta)$  is globally bounded from below by the function  $l(\theta') + Q(\theta, \theta')$ , with equality at  $\theta = \theta'$ . The algorithm iterations are described by the following steps:

*Step 1.*  $p = 0$ , initial guess  $\hat{\theta}_0$ .

*Step 2.* Set  $\theta' = \hat{\theta}_p$ .

*Step 3 (E-step).* Compute  $Q(\cdot, \theta')$ .

*Step 4 (M-step).* Find  $\hat{\theta}_{p+1}$  such that  $Q(\hat{\theta}_{p+1}, \theta') \geq Q(\theta, \theta')$  for all  $\theta \in \Theta$ .

*Step 5.* Repeat from Step 2 with  $p = p + 1$ , unless a stopping test is satisfied, in which case set  $\theta^* = \hat{\theta}_{p+1}$ .

An interesting feature of the algorithm is that it generates a maximizing sequence  $\{\hat{\theta}_p, p = 0, 1, \dots\}$  in the sense that  $l(\hat{\theta}_{p+1}) > l(\hat{\theta}_p)$  unless  $\hat{\theta}_{p+1} = \hat{\theta}_p$ . Some general convergence results about the sequences  $\{l(\hat{\theta}_p), p = 0, 1, \dots\}$  and  $\{\hat{\theta}_p, p = 0, 1, \dots\}$  are proved in Wu (1983), under mild regularity assumptions on  $l(\theta)$  and  $Q(\theta, \theta')$  (see also Dembo and Zeitouni, 1986, Theorem 2).

To decide whether this algorithm is interesting from a computational point of view, the following three questions should be answered.

(E) How expensive is the computation of the auxiliary function  $Q(\theta, \theta')$ ?

(M) How easy is the maximization with respect to  $\theta$  of the auxiliary function  $Q(\theta, \theta')$ ?

(EM) How fast is the convergence of this sub-optimal iterative algorithm towards the MLE?

In Dembo and Zeitouni (1986), the EM algorithm has been applied in the context of continuous time partially observed stochastic processes. In the particular case of diffusion processes, the general expression of  $Q(\theta, \theta')$  has been derived and said to involve a nonlinear smoothing problem. The purpose of this work is to address the following three points:

- discuss the expression in Dembo and Zeitouni (1986) giving  $Q(\theta, \theta')$  in terms of nonlinear smoothing problem—this will involve generalized stochastic calculus (Skorokhod integral);

- get an equivalent expression, in terms of a nonlinear filtering problem, for  $Q(\theta, \theta')$  and its gradient  $\nabla^{1,0} Q(\theta, \theta')$  with respect to  $\theta$ —it will turn out that smoothing is indeed necessary for the point (M) introduced above to be satisfied, although filtering is enough to compute  $Q(\theta, \theta')$  for a given pair  $(\theta, \theta')$ ;

- get similar expressions for the original log-likelihood function  $l(\theta)$  and its gradient  $\nabla l(\theta)$ .

This will allow to compare, from a computational point of view, the two possible methods for maximum likelihood estimation:

- direct maximization of the likelihood function, as described in Le Gland (1981);

- the EM algorithm.

In particular, the point (M) will receive a positive answer, which is indeed the main motivation for the EM algorithm. On the other hand, it will be proved that computing the auxiliary function  $Q(\theta, \theta')$  is a more heavy task than computing the original log-likelihood function  $l(\theta)$ . As for the point (EM), numerical examples will show that the convergence of the EM algorithm may be very slow. This typically occurs in those cases where, for each  $\theta' \in \Theta$  the function  $l(\theta') + Q(\theta, \theta')$  is very sharp below the log-likelihood function  $l(\theta)$  (see Fig. 4 below). In such cases indeed, maximizing the auxiliary function does not allow to update significantly enough the current estimate at each M-step.

The statistical model is presented in Section 2, where expressions are given for  $l(\theta)$ ,  $\nabla l(\theta)$ ,  $Q(\theta, \theta')$  and  $\nabla^{1,0}Q(\theta, \theta')$  in terms of conditional expectations. It turns out that the last three expressions all belong to a certain class of conditional expectations. Two methods are then proposed in Section 3 for computing conditional expectations in this class—one based on nonlinear filtering, the other on nonlinear smoothing and involving generalized stochastic calculus (Skorokhod integral). These results are applied in Section 4 to the computation of  $l(\theta)$ ,  $\nabla l(\theta)$ ,  $Q(\theta, \theta')$  and  $\nabla^{1,0}Q(\theta, \theta')$  in terms of filtering and smoothing densities. Section 5 is devoted to the time discretization of the stochastic PDE's introduced in Section 4, and the link with MLE of parameters in partially observed Markov chains (hidden Markov models) is explored. A numerical example is presented in Section 6, and the influence of noise covariances is investigated.

This paper only deals with different ways to compute the MLE of parameters in partially observed diffusion processes. Whether the MLE is a good estimate of the unknown parameter  $\theta$  is not investigated here. See James and Le Gland (1989) where a consistency result is proved in the “small noise” asymptotics.

## 2. Statistical model

In this section, expressions for the log-likelihood function  $l(\theta)$  and the auxiliary function  $Q(\theta, \theta')$  will be derived in the following context (Dembo and Zeitouni, 1986, Section 3).

Suppose that on a measurable space  $(\Omega, \mathcal{F})$  are given:

- (a) a family  $\mathcal{M} = \{P_\theta, \theta \in \Theta\}$  of probability measures,
- (b) a pair of stochastic processes  $\{X_t, t \geq 0\}$  and  $\{Y_t, t \geq 0\}$  taking values in  $\mathbb{R}^m$  and  $\mathbb{R}^d$  respectively,

such that under  $P_\theta$ ,

$$\begin{aligned} dX_t &= b_\theta(X_t) dt + \sigma(X_t) dW_t^\theta, & X_0 &\sim p_\theta^\theta(x) dx, \\ dY_t &= h_\theta(X_t) dt + dV_t^\theta, \end{aligned} \tag{2.1}$$

where  $\{W_t^\theta, t \geq 0\}$  and  $\{V_t^\theta, t \geq 0\}$  are independent Wiener processes, with covariance

matrix  $I$  (identity) and  $r$  respectively, and the random variable  $X_0$  is independent of the Wiener processes.

The set of parameters  $\Theta \subset \mathbb{R}^p$  is compact, and the coefficients satisfy the following hypotheses.

(i)  $\sigma$  is a continuous and bounded function on  $\mathbb{R}^m$  such that  $a \triangleq \sigma\sigma^*$  is a uniformly elliptic  $m \times m$  matrix, i.e.  $a(x) \geq \alpha I$ .

(ii) For all  $\theta \in \Theta$ ,  $p_\theta^\theta$  is a density on  $\mathbb{R}^m$ .

(iii) For all  $\theta \in \Theta$ ,  $b_\theta$  and  $h_\theta$  are measurable and bounded functions from  $\mathbb{R}^m$  to  $\mathbb{R}^m$  and  $\mathbb{R}^d$  respectively.

In addition:

(iv) The probability measures on  $\mathbb{R}^m$  with densities  $\{p_\theta^\theta, \theta \in \Theta\}$  are mutually absolutely continuous, and for all  $\theta, \theta' \in \Theta$  the function  $\log(p_\theta^\theta/p_{\theta'}^{\theta'})$  is integrable with respect to  $p_{\theta'}^{\theta'}$ .

Moreover, it is assumed that  $p_\theta^\theta$ ,  $b_\theta$  and  $h_\theta$  are continuously differentiable with respect to the parameter  $\theta$ , and that:

(v) For all  $\theta \in \Theta$ ,  $\nabla b_\theta$  and  $\nabla h_\theta$  are measurable and bounded functions from  $\mathbb{R}^m$  to  $\mathbb{R}^{m \times p}$  and  $\mathbb{R}^{d \times p}$  respectively.

(Throughout this paper,  $\nabla$  will denote the derivation with respect to the parameter  $\theta$ .) In addition:

(vi) For all  $\theta, \theta' \in \Theta$  the function  $\nabla \log p_\theta^\theta$  is integrable with respect to  $p_{\theta'}^{\theta'}$ .

The hypotheses ensure the existence and uniqueness of a weak solution to the stochastic differential system (2.1). There is no loss in generality in assuming that  $\Omega$  is the canonical space  $C([0, T]; \mathbb{R}^{m+d})$ , in which case  $X$  and  $Y$  are the canonical processes on  $C([0, T]; \mathbb{R}^m)$  and  $C([0, T]; \mathbb{R}^d)$  respectively, and  $P_\theta$  is the probability law of  $(X, Y)$ .

The probability measures in  $\mathcal{M}$  are mutually absolutely continuous:

- Define first a probability measure  $P_\theta^\dagger$  equivalent to  $P_\theta$ , with

$$Z^\theta \triangleq \frac{dP_\theta}{dP_\theta^\dagger} = \exp\left\{ \int_0^T h_\theta^*(X_s) r^{-1} dY_s - \frac{1}{2} \int_0^T h_\theta^*(X_s) r^{-1} h_\theta(X_s) ds \right\},$$

so that under  $P_\theta^\dagger$ ,

$$dX_t = b_\theta(X_t) dt + \sigma(X_t) dW_t^\theta, \quad X_0 \sim p_\theta^\theta(x) dx,$$

where  $\{W_t^\theta, t \geq 0\}$  and  $\{Y_t, t \geq 0\}$  are independent Wiener processes, with covariance matrix  $I$  and  $r$  respectively, and the random variable  $X_0$  is independent of the Wiener processes.

• Then, the probability measures  $\{P_\theta^\dagger, \theta \in \Theta\}$  are mutually absolutely continuous with Radon–Nikodym derivative

$$\begin{aligned} \Lambda_{\theta, \theta'}^\dagger &\triangleq \frac{dP_\theta^\dagger}{dP_{\theta'}^\dagger} \\ &= \frac{p_\theta^\theta}{p_{\theta'}^{\theta'}}(X_0) \cdot \exp\left\{ \int_0^T [b_\theta(X_s) - b_{\theta'}(X_s)]^* a^{-1}(X_s) \sigma(X_s) dW_s^{\theta'} \right\} \end{aligned}$$

$$-\frac{1}{2} \int_0^T [b_\theta(X_s) - b_{\theta'}(X_s)]^* a^{-1}(X_s) [b_\theta(X_s) - b_{\theta'}(X_s)] ds \Big\}.$$

Therefore

$$A_{\theta, \theta'} \triangleq \frac{dP_\theta}{dP_{\theta'}} = A_{\theta, \theta'}^\dagger \frac{Z^\theta}{Z^{\theta'}}.$$

It is assumed that only  $\{Y_t, 0 \leq t \leq T\}$  is observed. Let  $\{\mathcal{Y}_t, 0 \leq t \leq T\}$  denote the associated filtration. The likelihood function for estimating the parameter  $\theta$  in the statistical model  $\mathcal{M}$  can be expressed as

$$E_\alpha^\dagger \left( \frac{dP_\theta}{dP_\alpha^\dagger} \Big| \mathcal{Y}_T \right) = E_\alpha^\dagger (Z^\theta A_{\theta, \alpha}^\dagger | \mathcal{Y}_T),$$

with the particular choice  $P' = P_\alpha^\dagger$  ( $\alpha$  fixed in  $\Theta$ ) in (1.1). By the Bayes formula,

$$E_\alpha^\dagger (Z^\theta A_{\theta, \alpha}^\dagger | \mathcal{Y}_T) = E_\theta^\dagger (Z^\theta | \mathcal{Y}_T) \cdot E_\alpha^\dagger (A_{\theta, \alpha}^\dagger | \mathcal{Y}_T) = E_\theta^\dagger (Z^\theta | \mathcal{Y}_T),$$

since  $A_{\theta, \alpha}^\dagger$  is independent of  $\mathcal{Y}_T$  under  $P_\alpha^\dagger$ . This gives the following expression for the log-likelihood function

$$l(\theta) = \log E_\theta^\dagger (Z^\theta | \mathcal{Y}_T), \tag{2.2}$$

which is independent of the arbitrary choice of  $\alpha$ .

For the auxiliary function defined by (1.2), one has immediately

$$Q(\theta, \theta') = E_\theta (\lambda^{\theta, \theta'} | \mathcal{Y}_T) = E_\theta^\dagger (\lambda^{\theta, \theta'} Z^{\theta'} | \mathcal{Y}_T) / E_{\theta'}^\dagger (Z^{\theta'} | \mathcal{Y}_T), \tag{2.3}$$

where

$$\begin{aligned} \lambda^{\theta, \theta'} \triangleq \log A_{\theta, \theta'} &= \log \frac{P_\theta^\theta}{P_{\theta'}^\theta} (X_0) + \int_0^T [b_\theta(X_s) - b_{\theta'}(X_s)]^* a^{-1}(X_s) \sigma(X_s) dW_s^{\theta'} \\ &\quad - \frac{1}{2} \int_0^T [b_\theta(X_s) - b_{\theta'}(X_s)]^* a^{-1}(X_s) [b_\theta(X_s) - b_{\theta'}(X_s)] ds \\ &\quad + \int_0^T [h_\theta(X_s) - h_{\theta'}(X_s)]^* r^{-1} dV_s^{\theta'} \\ &\quad - \frac{1}{2} \int_0^T [h_\theta(X_s) - h_{\theta'}(X_s)]^* r^{-1} [h_\theta(X_s) - h_{\theta'}(X_s)] ds. \end{aligned} \tag{2.4}$$

Under additional regularity assumptions on the coefficients  $p_0^\theta$ ,  $b_\theta$  and  $h_\theta$ , it is easy to prove, using results in Sznitman (1982), that both  $l(\theta)$  and  $Q(\theta, \theta')$  have a.s. differentiable versions with respect to  $\theta$ , with gradients given by

$$\nabla l(\theta) = E_\theta (\lambda^\theta | \mathcal{Y}_T) = E_\theta^\dagger (\lambda^\theta Z^\theta | \mathcal{Y}_T) / E_\theta^\dagger (Z^\theta | \mathcal{Y}_T), \tag{2.5}$$

$$\nabla^{1,0} Q(\theta, \theta') = E_{\theta'} (\lambda^\theta | \mathcal{Y}_T) = E_{\theta'}^\dagger (\lambda^\theta Z^\theta | \mathcal{Y}_T) / E_{\theta'}^\dagger (Z^\theta | \mathcal{Y}_T), \tag{2.6}$$

respectively, where

$$\begin{aligned} \lambda^\theta &\triangleq \nabla^{1,0} \log \Lambda_{\theta,\theta'} = \nabla^{1,0} \lambda^{\theta,\theta'} \\ &= \frac{\nabla p_0^\theta}{p_0^\theta}(X_0) + \int_0^T [\nabla b_\theta(X_s)]^* a^{-1}(X_s) \sigma(X_s) dW_s^\theta \\ &\quad + \int_0^T [\nabla h_\theta(X_s)]^* r^{-1} dV_s^\theta \end{aligned}$$

is independent of  $\theta'$ .

**Remark 2.1.** One can check from (2.5) and (2.6) that

$$\nabla^{1,0} Q(\theta, \theta')|_{\theta=\theta'} = \nabla l(\theta').$$

In the next section, two different methods will be given—by means of stochastic PDE’s—to compute the various quantities introduced so far:  $l(\theta)$ ,  $\nabla l(\theta)$ ,  $Q(\theta, \theta')$  and  $\nabla^{1,0} Q(\theta, \theta')$ . This will make possible numerical implementation of algorithms for maximizing the likelihood function.

### 3. Smoothing vs. filtering for computing a class of conditional expectations

For the sake of simplicity, any reference to the parameter  $\theta$  will be dropped throughout this section. In particular,  $P$  will denote the probability measure under which

$$\begin{aligned} dX_t &= b(X_t) dt + \sigma(X_t) dW_t, \quad X_0 \sim p_0(x) dx, \\ dY_t &= h(X_t) dt + dV_t, \end{aligned}$$

where  $\{W_t, 0 \leq t \leq T\}$  and  $\{V_t, 0 \leq t \leq T\}$  are independent Wiener processes, with covariance matrix  $I$  and  $r$  respectively, and the random variable  $X_0$  is independent of the Wiener processes, whereas under  $P^\dagger$ ,

$$dX_t = b(X_t) dt + \sigma(X_t) dW_t, \quad X_0 \sim p_0(x) dx,$$

where  $\{W_t, 0 \leq t \leq T\}$  and  $\{Y_t, 0 \leq t \leq T\}$  are independent Wiener processes, with covariance matrix  $I$  and  $r$  respectively, and the random variable  $X_0$  is independent of the Wiener processes. Therefore  $P = Z_T \cdot P^\dagger$ , where the process  $\{Z_t, 0 \leq t \leq T\}$  is defined by

$$Z_t \triangleq \exp \left\{ \int_0^t h^*(X_s) r^{-1} dY_s - \frac{1}{2} \int_0^t h^*(X_s) r^{-1} h(X_s) ds \right\}.$$

The purpose of this section is to provide two different methods—one based on nonlinear filtering, the other on nonlinear smoothing—for computing the following

class of conditional expectations

$$\begin{aligned}
 A \triangleq & \mathbf{E} \left( \beta(X_0) + \int_0^T \xi(X_s) \, ds + \int_0^T \eta^*(X_s) \, dV_s \right. \\
 & \left. + \int_0^T \chi^*(X_s) \sigma(X_s) \, dW_s \mid \mathcal{Y}_T \right), \tag{3.1}
 \end{aligned}$$

where  $\beta$ ,  $\xi$ ,  $\eta$  and  $\chi$  are measurable and bounded functions from  $\mathbb{R}^m$  to  $\mathbb{R}$ ,  $\mathbb{R}$ ,  $\mathbb{R}^d$  and  $\mathbb{R}^m$  respectively. It is readily seen from (2.3-2.7) that the computation of either  $\nabla l(\theta)$ ,  $Q(\theta, \theta')$  or  $\nabla^{1,0}Q(\theta, \theta')$  involves such conditional expectations.

It is clear from the definition that  $A$  depends linearly on  $(\beta, \xi, \eta, \chi)$ . It will turn out that nonlinear smoothing is the only way to make this dependence explicit, although nonlinear filtering—which is simpler—is enough to just compute  $A$ .

Rewriting  $A$  as

$$\begin{aligned}
 A = & \mathbf{E}(\beta(X_0) \mid \mathcal{Y}_T) + \int_0^T \mathbf{E}(\xi(X_s) - \eta^*(X_s)h(X_s) \mid \mathcal{Y}_T) \, ds \\
 & + \mathbf{E} \left( \int_0^T \eta^*(X_s) \, dY_s \mid \mathcal{Y}_T \right) + \mathbf{E} \left( \int_0^T \chi^*(X_s) \sigma(X_s) \, dW_s \mid \mathcal{Y}_T \right), \tag{3.2}
 \end{aligned}$$

one would like to interchange conditional expectation and stochastic integral in the third term of (3.2). However, the resulting expression

$$\left\langle \int_0^T \mathbf{E}(\eta^*(X_s) \mid \mathcal{Y}_T) \, dY_s \right\rangle \tag{3.3}$$

is not an Itô integral, since the integrand is obviously not adapted to the filtration  $\{\mathcal{Y}_t, 0 \leq t \leq T\}$ , and needs to be given a rigorous meaning. It will be proved in Proposition 3.5 below that the correct statement is

$$\begin{aligned}
 \mathbf{E} \left( \int_0^T \eta^*(X_s) \, dY_s \mid \mathcal{Y}_T \right) &= \mathbf{E} \left( \int_0^T \eta^*(X_s) \circ dY_s \mid \mathcal{Y}_T \right) \\
 &= \int_0^T \mathbf{E}(\eta^*(X_s) \mid \mathcal{Y}_T) \circ dY_s \\
 &\neq \int_0^T \mathbf{E}(\eta^*(X_s) \mid \mathcal{Y}_T) \, dY_s,
 \end{aligned}$$

where the non-adapted stochastic integrals are respectively a generalized Stratonovich integral and a Skorokhod integral, as defined in Nualart and Pardoux (1988).

In the particular case where  $\chi$  is a gradient vector field, one has the following.

**Proposition 3.1.** Assume there exists a scalar function  $U$  defined in  $\mathbb{R}^m$ , twice continuously differentiable with bounded derivatives, such that  $\chi = U'$ . Then

$$\begin{aligned} & \mathbf{E} \left( \int_0^T \chi^*(X_s) \sigma(X_s) dW_s \mid \mathcal{Y}_T \right) \\ &= \mathbf{E}(U(X_T) \mid \mathcal{Y}_T) - \mathbf{E}(U(X_0) \mid \mathcal{Y}_T) - \int_0^T \mathbf{E}(LU(X_s) \mid \mathcal{Y}_T) ds, \end{aligned} \tag{3.4}$$

where  $L$  is the infinitesimal generator of the diffusion process  $\{X_t, 0 \leq t \leq T\}$ , see (3.7) below.  $\square$

This proposition follows immediately from Itô's lemma.

At this point, it is necessary to introduce some notations and definitions related to nonlinear filtering and smoothing.

**Notations and definitions.** *Filtering:* Let  $\pi_t$  (resp.  $p_t$ ) denote the normalized (resp. unnormalized) conditional density of the random variable  $X_t$  given  $\mathcal{Y}_t$ , i.e.

$$(\pi_t, \phi) \triangleq \mathbf{E}(\phi(X_t) \mid \mathcal{Y}_t), \quad (p_t, \phi) \triangleq \mathbf{E}^\dagger(\phi(X_t) Z_t \mid \mathcal{Y}_t), \tag{3.5}$$

for any test-function  $\phi$ . By the Bayes formula:

$$(\pi_t, \phi) = (p_t, \phi) / (p_t, 1).$$

The equation for  $\{p_t, 0 \leq t \leq T\}$  is the Zakai equation (Pardoux, 1979),

$$dp_t = L^* p_t dt + h^* p_t r^{-1} dY_t, \tag{3.6}$$

where  $L^*$  is the adjoint operator of the infinitesimal generator  $L$  of the diffusion process  $\{X_t, 0 \leq t \leq T\}$  defined by

$$L \triangleq \frac{1}{2} \sum_{i,j=1}^m a^{i,j} \frac{\partial^2}{\partial x_i \partial x_j} + \sum_{i=1}^m b^i \frac{\partial}{\partial x_i}. \tag{3.7}$$

*Smoothing (fixed-interval):* Let  $T > 0$  denote the fixed end-time, and  $\rho_t$  (resp.  $q_t$ ) denote the normalized (resp. unnormalized) conditional density of the random variable  $X_t$  given  $\mathcal{Y}_T$ , i.e.

$$(\rho_t, \phi) \triangleq \mathbf{E}(\phi(X_t) \mid \mathcal{Y}_T), \quad (q_t, \phi) \triangleq \mathbf{E}^\dagger(\phi(X_t) Z_T \mid \mathcal{Y}_T).$$

Again

$$(\rho_t, \phi) = (q_t, \phi) / (q_t, 1).$$

Introducing the backward Zakai equation

$$dv_t + Lv_t dt + h^* v_t r^{-1} dY_t = 0, \quad v_T \equiv 1, \tag{3.8}$$

it is proved by Pardoux (1979, 1984) that  $(p_t, v_t)$  is independent of  $t$ ,  $q_t = p_t v_t$ , and satisfies

$$\dot{q}_t + p_t Lv_t = v_t L^* p_t. \tag{3.9}$$

Note that

$$(q_t, 1) = (p_T, 1), \quad 0 \leq t \leq T. \tag{3.10}$$

Existence, uniqueness and regularity results for stochastic PDE's can be found in Pardoux (1979), Krylov and Rozovskii (1977). The stochastic calculus of variations for stochastic PDE's is presented in Ocone (1988).

### 3.1. Filtering approach

Define

$$\lambda_t \triangleq \beta(X_0) + \int_0^t \xi(X_s) ds + \int_0^t \eta^*(X_s) dV_s + \int_0^t \chi^*(X_s) \sigma(X_s) dW_s,$$

so that, by the Bayes formula

$$A = \mathbf{E}(\lambda_T | \mathcal{Y}_T) = \mathbf{E}^\dagger(\lambda_T Z_T | \mathcal{Y}_T) / \mathbf{E}^\dagger(Z_T | \mathcal{Y}_T).$$

A first method would be to compute the joint conditional density of  $(X_T, \lambda_T)$  given  $\mathcal{Y}_T$ , and then integrate over the first variable to get the marginal conditional density of  $\lambda_T$  given  $\mathcal{Y}_T$ . An alternative method is to find an equation for  $\{w_t, 0 \leq t \leq T\}$  defined by

$$(w_t, \phi) \triangleq \mathbf{E}^\dagger(\phi(X_t) \lambda_t Z_t | \mathcal{Y}_t).$$

Indeed, by Itô's lemma

$$\begin{aligned} d[\phi(X_t) \lambda_t Z_t] &= \lambda_t Z_t L\phi(X_t) dt + \lambda_t Z_t [\phi'(X_t)]^* \sigma(X_t) dW_t \\ &\quad + \phi(X_t) Z_t \xi(X_t) dt + \phi(X_t) Z_t \eta^*(X_t) dV_t \\ &\quad + \phi(X_t) Z_t \chi^*(X_t) \sigma(X_t) dW_t + \phi(X_t) \lambda_t Z_t h^*(X_t) r^{-1} dY_t \\ &\quad + \phi(X_t) \eta^*(X_t) h(X_t) Z_t dt + Z_t \chi(X_t)^* a(X_t) \phi'(X_t) dt. \end{aligned}$$

Using properties of conditional expectation given the observation  $\sigma$ -algebra under the reference probability measure  $P^\dagger$ , and the definition (3.5), gives

$$\begin{aligned} (w_t, \phi) &= (p_0, \beta\phi) + \int_0^t (w_s, L\phi) ds + \int_0^t (w_s, h^*\phi) r^{-1} dY_s \\ &\quad + \int_0^t (p_s, \xi\phi) ds + \int_0^t (p_s, \eta^*\phi) dY_s + \int_0^t (p_s, J(\chi)\phi) ds, \end{aligned}$$

where

$$J(\chi)\phi \triangleq \chi^* a \phi' = \sum_{i,j=1}^m a^{i,j} \chi_j \frac{\partial \phi}{\partial x_i}, \tag{3.11}$$

so that  $\{w_t, 0 \leq t \leq T\}$  solves

$$\begin{aligned} dw_t &= L^* w_t dt + h^* w_t r^{-1} dY_t + \xi p_t dt + \eta^* p_t dY_t + J^*(\chi) p_t dt, \\ w_0 &= \beta p_0. \end{aligned} \tag{3.12}$$

The following theorem has been proved.

**Theorem 3.2.** Let  $\{p_t, 0 \leq t \leq T\}$  and  $\{w_t, 0 \leq t \leq T\}$  be the unique solutions of (3.6) and (3.12) respectively. Then, the following expression holds for  $A$  defined in (3.1),

$$A = (w_T, 1) / (p_T, 1). \quad \square \tag{3.13}$$

To get an expression for  $A$  in terms of normalized conditional densities, define  $\alpha_t$  by

$$(\alpha_t, \phi) \triangleq (w_t, \phi) / (p_t, 1) = \mathbf{E}(\phi(X_t) \lambda_t | \mathcal{Y}_t).$$

By Itô's lemma,

$$\begin{aligned} d\alpha_t &= L^* \alpha_t dt + [h^* - (\pi_t, h^*)] \alpha_t r^{-1} [dY_t - (\pi_t, h) dt] + \xi \pi_t dt \\ &\quad + \eta^* \pi_t [dY_t - (\pi_t, h) dt] + J^*(\chi) \pi_t dt. \end{aligned}$$

Therefore

$$A = (\pi_0, \beta) + \int_0^T (\pi_s, \xi) ds + \int_0^T \{(\pi_s, \eta^*) + [(\alpha_s, h^*) - (\pi_s, h^*)(\alpha_s, 1)] r^{-1}\} [dY_s - (\pi_s, h) ds], \tag{3.14}$$

which is the expression given in Lipster and Shiriyayev (1977, Theorem 8.1).

**Remark 3.3.** From the computational point of view, it is enough in (3.13) to integrate the unnormalized conditional densities at final time  $T$ , whereas in (3.14) one has (i) at each time  $t$ , to integrate some functions involving in particular  $(\xi, \eta)$  against the normalized conditional densities, and (ii) to integrate the resulting processes over the interval  $[0, T]$ .

The expression (3.13) is actually computable. Unfortunately, the linear dependence of  $(w_T, 1)$  on  $(\beta, \xi, \eta, \chi)$  is not made explicit, which should be the case for the point (M) introduced in the Introduction to be satisfied. Therefore, the next step will be to make this dependence more explicit. This will involve nonlinear smoothing and generalized stochastic calculus (Skorokhod integral). Actually, the stochastic integral in (3.3) will be given a rigorous meaning, and the last term in (3.2) will also be given a computable expression, whether or not  $\chi$  is a gradient vector field.

### 3.2. Smoothing approach

The idea here is to compute the stochastic differential of the scalar product  $(w_t, v_t)$ , where  $\{v_t, 0 \leq t \leq T\}$  is the solution of the backward Zakai equation (3.8). Since (3.12) is a forward stochastic PDE and (3.8) is a backward stochastic PDE, one has to use the two-sided stochastic calculus introduced in Pardoux and Protter (1987) and Pardoux (1987). This gives

$$\begin{aligned} d(w_t, v_t) &= (L^* w_t, v_t) dt + (h^* w_t, v_t) r^{-1} dY_t + (\xi p_t, v_t) dt + (\eta^* p_t, v_t) dY_t \\ &\quad + (J^*(\chi) p_t, v_t) dt - (w_t, Lv_t) dt - (w_t, h^* v_t) r^{-1} dY_t \\ &= (q_t, \xi) dt + (q_t, \eta^*) dY_t + (p_t, J(\chi) v_t) dt. \end{aligned}$$

Integrating from 0 to  $T$  gives

$$(w_T, 1) = (q_0, \beta) + \int_0^T (q_s, \xi) ds + \int_0^T (q_s, \eta^*) dY_s + \int_0^T (p_s, J(\chi)v_s) ds,$$

where the stochastic integral is a two-sided stochastic integral.

**Remark 3.4.** Since  $s \mapsto q_s$  is differentiable (see (3.9)) the stochastic integral could indeed be defined using an integration by parts.

To get an expression for  $A$  in terms of normalized conditional densities, the first step is to use (3.10), which gives

$$A = (\rho_0, \beta) + \int_0^T (\rho_s, \xi) ds + A' + A''.$$

*Study of  $A'$ :*

$$A' \triangleq \left( \int_0^T (q_s, \eta^*) dY_s \right) / (p_T, 1). \tag{3.15}$$

Note that

$$E(A') = E^\dagger(Z_T A') = E^\dagger(E^\dagger(Z_T | \mathcal{Y}_T) A') = E^\dagger\left(\int_0^T (q_s, \eta^*) dY_s\right) = 0,$$

where the last equality follows from results on two-sided stochastic integrals. This was expected, since

$$A' = E\left(\int_0^T \eta^*(X_s) dV_s | \mathcal{Y}_T\right).$$

Expressions in terms of normalized conditional densities are given by the following.

**Proposition 3.5.** *Let  $\{\rho_t, 0 \leq t \leq T\}$  denote the normalized smoothing density. Then*

$$\begin{aligned} A' &= \int_0^T (\rho_s, \eta^*) dY_s - \int_0^T (\rho_s, \eta^*)(\rho_s, h) ds \\ &= \int_0^T (\rho_s, \eta^*) \circ dY_s - \int_0^T (\rho_s, \eta^* h) ds, \end{aligned}$$

where the non-adapted stochastic integrals are respectively a Skorokhod integral and a generalized Stratonovich integral (Nualart and Pardoux, 1988).

**Proof.** The idea is to get the factor  $F \triangleq 1/(p_T, 1)$  inside the stochastic integral in (3.15), using the generalized stochastic calculus developed in Nualart and Pardoux (1988).

On the probability space  $(\Omega, \mathcal{F}, P^*)$ , let  $D$  denote the derivative with respect to the  $d$ -dimensional Wiener process  $\{Y_t, 0 \leq t \leq T\}$  in the direction of the vector space  $H^1(0, T; \mathbb{R}^d)$ . Since the two-sided integral is a particular case of the Skorokhod integral, it follows from Nualart and Pardoux (1988, Proposition 3.2) that

$$\begin{aligned} A' &= F \int_0^T (q_s, \eta^*) dY_s = \int_0^T F(q_s, \eta^*) dY_s + \int_0^T (q_s, \eta^*) r D_s F ds \\ &= \int_0^T \frac{(q_s, \eta^*)}{(q_s, 1)} dY_s - \int_0^T \frac{(q_s, \eta^*)}{(q_s, 1)^2} r(D_s p_T, 1) ds, \end{aligned}$$

where the stochastic integral is a Skorokhod integral, and  $r$  is the covariance matrix of the Wiener process  $\{Y_t, 0 \leq t \leq T\}$ .

For  $s$  fixed in  $[0, T]$ , consider the  $d$ -dimensional random process  $\{z_t, 0 \leq t \leq T\}$  defined by  $z_t \triangleq D_s p_t$ . Clearly  $z_t \equiv 0$  for  $0 \leq t < s$ . On the other hand, it follows from Ocone (1988) that the process  $\{z_t, s \leq t \leq T\}$  is the unique solution of the forward stochastic PDE:

$$dz_t = L^* z_t dt + h^* z_t r^{-1} dY_t, \quad z_s = r^{-1} h p_s.$$

Introducing the solution  $\{v_t, 0 \leq t \leq T\}$  of the backward Zakai equation (3.8) and using again the two-sided stochastic calculus gives  $d(z_t, v_t) = 0$  for  $s \leq t \leq T$ . Therefore

$$(D_s p_T, 1) = (z_T, 1) = (z_s, v_s) = r^{-1}(q_s, h),$$

so that

$$\begin{aligned} A' &= \int_0^T \frac{(q_s, \eta^*)}{(q_s, 1)} dY_s - \int_0^T \frac{(q_s, \eta^*)(q_s, h)}{(q_s, 1)^2} ds \\ &= \int_0^T (\rho_s, \eta^*) dY_s - \int_0^T (\rho_s, \eta^*)(\rho_s, h) ds. \end{aligned}$$

To get the second expression, consider the  $d$ -dimensional random process  $\{u_t, 0 \leq t \leq T\}$  defined by  $u_t \triangleq (\rho_t, \eta)$ . The Skorokhod-Stratonovich transformation for generalized stochastic integrals gives (Nualart and Pardoux, 1988, Theorem 7.3),

$$\int_0^T u_s^* dY_s = \int_0^T u_s^* \circ dY_s - \frac{1}{2} \int_0^T (D_s^+ u_s + D_s^- u_s) ds,$$

where

$$D_s^+ u_s \triangleq \lim_{t \downarrow s} \sum_{i,j=1}^d r_{i,j} D_s^i u_t^j, \quad D_s^- u_s \triangleq \lim_{t \uparrow s} \sum_{i,j=1}^d r_{i,j} D_s^i u_t^j.$$

It turns out that

$$D_s^i u_t^j = D_s^i(\rho_t, \eta^j) = D_s^i \frac{(q_t, \eta^j)}{(q_t, 1)} = \frac{(D_s^i q_t, \eta^j)}{(q_t, 1)} - \frac{(q_t, \eta^j)(D_s^i q_t, 1)}{(q_t, 1)^2}.$$

Therefore

$$\sum_{i,j=1}^d r_{i,j} D_s^i u_t^j = \frac{((r\eta)^*, D_s q_t)}{(q_t, 1)} - \frac{(q_t, (r\eta)^*)(D_s q_t, 1)}{(q_t, 1)^2}.$$

Next  $D_s q_t = (D_s p_t) v_t + p_t (D_s v_t)$ . In particular  $D_s p_t$  has already been studied, and a similar argument for  $D_s v_t$  shows that

$$\lim_{t \downarrow s} D_s q_t = \lim_{t \uparrow s} D_s q_t = r^{-1} h q_s.$$

Therefore

$$D_s^+ u_s = D_s^- u_s = \frac{(q_s, \eta^* h)}{(q_s, 1)} - \frac{(q_s, \eta^*)(q_s, h)}{(q_s, 1)^2} = (\rho_s, \eta^* h) - (\rho_s, \eta^*)(\rho_s, h).$$

This finally gives

$$A' = \int_0^T (\rho_s, \eta^*) \circ dY_s - \int_0^T (\rho_s, \eta^* h) ds,$$

where the stochastic integral is now a generalized Stratonovich integral.  $\square$

**Remark 3.6.** In terms of conditional expectations,

$$\begin{aligned} A' &= \int_0^T \mathbf{E}(\eta^*(X_s) | \mathcal{Y}_T) dY_s - \int_0^T \mathbf{E}(\eta^*(X_s) | \mathcal{Y}_T) \mathbf{E}(h(X_s) | \mathcal{Y}_T) ds \\ &= \int_0^T \mathbf{E}(\eta^*(X_s) | \mathcal{Y}_T) \circ dY_s - \int_0^T \mathbf{E}(\eta^*(X_s) h(X_s) | \mathcal{Y}_T) ds. \end{aligned}$$

*Study of  $A''$ :*

$$A'' \triangleq \left( \int_0^T (p_s, J(\chi) v_s) ds \right) / (p_T, 1), \tag{3.16}$$

where the first-order partial differential operator  $J(\chi)$  is defined in (3.11). Note that

$$\begin{aligned} \mathbf{E}(A'') &= \mathbf{E}^\dagger(Z_T A'') = \mathbf{E}^\dagger(\mathbf{E}^\dagger(Z_T | \mathcal{Y}_T) A'') \\ &= \mathbf{E}^\dagger\left(\int_0^T (p_s, J(\chi) v_s) ds\right) = \int_0^T (\mathbf{E}^\dagger(p_s), J(\chi) \mathbf{E}^\dagger(v_s)) ds, \end{aligned}$$

where the last equality follows from the independence of  $p_s$  and  $v_s$  under the probability measure  $P^\dagger$ . Now  $J(\chi) \mathbf{E}^\dagger(v_s) \equiv 0$  since  $\mathbf{E}^\dagger(v_s) \equiv 1$ . Therefore  $\mathbf{E}(A'') = 0$ , which was expected since

$$A'' = \mathbf{E}\left(\int_0^T \chi^*(X_s) \sigma(X_s) dW_s | \mathcal{Y}_T\right).$$

The following two other expressions for  $A''$ , in terms of normalized conditional densities, are easily obtained:

$$A'' = \int_0^T \left( \pi_s, J(\chi) \left( \frac{\rho_s}{\pi_s} \right) \right) ds = \int_0^T \left( \rho_s, J(\chi) \left( \log \frac{\rho_s}{\pi_s} \right) \right) ds.$$

In the particular case where  $\chi$  is a gradient vector field, it can be checked that (3.16) reduces to the expression (3.4) given in Proposition 3.1. Indeed:

**Proposition 3.7.** *Assume there exists a scalar function  $U$  defined in  $\mathbb{R}^m$ , twice continuously differentiable with bounded derivatives, such that  $\chi = U'$ . Then*

$$A'' = E(U(X_T) | \mathcal{Y}_T) - E(U(X_0) | \mathcal{Y}_T) - \int_0^T E(LU(X_s) | \mathcal{Y}_T) ds,$$

which is exactly (3.4).

**Proof.** It follows from the identity  $L(Uv_s) = ULv_s + v_sLU + J(\chi)v_s$ , and from (3.9) that

$$\begin{aligned} (p_s, J(\chi)v_s) &= (p_s, L(Uv_s)) - (p_s, ULv_s) - (p_s, v_sLU) \\ &= (v_sL^*p_s - p_sLv_s, U) - (p_sv_s, LU) = (\dot{q}_s, U) - (q_s, LU). \end{aligned}$$

Integrating from 0 to  $T$  gives

$$\int_0^T (p_s, J(\chi)v_s) ds = (q_T, U) - (q_0, U) - \int_0^T (q_s, LU) ds.$$

Dividing by  $(p_T, 1)$  and using (3.10) gives

$$A'' = (\rho_T, U) - (\rho_0, U) - \int_0^T (\rho_s, LU) ds,$$

which finishes the proof.  $\square$

The following theorem has been proved.

**Theorem 3.8.** *Let  $\{\pi_t, 0 \leq t \leq T\}$  and  $\{\rho_t, 0 \leq t \leq T\}$  be the normalized filtering and smoothing densities, computed from the unique solutions  $\{p_t, 0 \leq t \leq T\}$  and  $\{v_t, 0 \leq t \leq T\}$  of (3.6) and (3.8) respectively. Then, the following expression holds for  $A$  defined in (3.1):*

$$A = (\rho_0, \beta) + \int_0^T (\rho_s, \xi) ds + A' + A'',$$

with

$$\begin{aligned} A' &= \int_0^T (\rho_s, \eta^*) dY_s - \int_0^T (\rho_s, \eta^*)(\rho_s, h) ds \\ &= \int_0^T (\rho_s, \eta^*) \circ dY_s - \int_0^T (\rho_s, \eta^*h) ds, \\ A'' &= \int_0^T \left( \pi_s, J(\chi) \left( \frac{\rho_s}{\pi_s} \right) \right) ds = \int_0^T \left( \rho_s, J(\chi) \left( \log \frac{\rho_s}{\pi_s} \right) \right) ds, \end{aligned}$$

where the non-adapted stochastic integrals are respectively a Skorokhod integral and a generalized Stratonovich integral (Nualart and Pardoux, 1988), and where the first-order partial differential operator  $J(\chi)$  is defined in (3.11).  $\square$

The advantage of smoothing over filtering is that the linear dependence on  $(\beta, \xi, \eta, \chi)$  is made explicit: provided the underlying probability measure does not change, evaluating  $A$  for a different set of data  $(\beta, \xi, \eta, \chi)$  will not require the computation of a new infinite-dimensional conditional density. In the filtering approach, one would have to solve a new stochastic PDE, with the same dynamics and a different “right-hand side”.

On the other hand, from the computational point of view, solving the equation for the smoothing density requires not only the computation but also the storage of the filtering density, and is therefore more expensive. Moreover, in the filtering approach it is enough to integrate the unnormalized filtering density at final time  $T$ , whereas in the smoothing approach one has (i) at each time  $t$ , to integrate some functions involving  $(\xi, \eta, \chi)$  against the normalized smoothing density, and (ii) to integrate the resulting processes over the interval  $[0, T]$ .

#### 4. Application to the MLE problem

In this section, computable expressions will be given for quantities related to the direct maximization of the likelihood function and to the EM algorithm. Results of Section 2 are used to check that the quantity to be computed belongs to the class of conditional expectations considered in Section 3. Then, Theorem 3.2 or Theorem 3.8 is applied to get a computable expression in terms of filtering or smoothing densities respectively. Expression involving normalized conditional densities will also be provided.

##### 4.1. Direct maximization of the likelihood function

It follows from (2.2) that the log-likelihood function  $l(\theta)$  can be expressed as

$$l(\theta) = \log(p_T^\theta, 1),$$

with  $\{p_t^\theta, 0 \leq t \leq T\}$  given by, see (3.6),

$$dp_t^\theta = L_\theta^* p_t^\theta dt + h_\theta^* p_t^\theta r^{-1} dY_t, \tag{4.1}$$

and

$$L_\theta \triangleq \frac{1}{2} \sum_{i,j=1}^m a^{i,j} \frac{\partial^2}{\partial x_i \partial x_j} + \sum_{i=1}^m b_\theta^i \frac{\partial}{\partial x_i}.$$

In terms of normalized conditional densities

$$l(\theta) = \int_0^T (\pi_s^\theta, h_\theta^*) r^{-1} dY_s - \frac{1}{2} \int_0^T (\pi_s^\theta, h_\theta^*) r^{-1} (\pi_s^\theta, h_\theta) ds,$$

where  $\{\pi_t^\theta, 0 \leq t \leq T\}$  is the normalized filtering density, computed from the unique solution  $\{p_t^\theta, 0 \leq t \leq T\}$  of (4.1).

It follows from (2.5) and (2.7) that  $\nabla l(\theta)$  belongs to the class of conditional expectations considered in Section 3. The approach based on filtering (Theorem 3.2) gives

$$\nabla l(\theta) = (w_T^\theta, 1) / (p_T^\theta, 1),$$

with  $\{p_t^\theta, 0 \leq t \leq T\}$  and  $\{w_t^\theta, 0 \leq t \leq T\}$  given respectively by (4.1) and, see (3.12),

$$\begin{aligned} dw_t^\theta &= L_\theta^* w_t^\theta dt + h_\theta^* w_t^\theta r^{-1} dY_t + J_\theta^* p_t^\theta dt + [\nabla h_\theta]^* p_t^\theta r^{-1} dY_t, \\ w_0^\theta &= \nabla p_0^\theta, \end{aligned} \tag{4.2}$$

where

$$J_\theta \phi \triangleq [\nabla b_\theta]^* \phi' = \sum_{i=1}^m \nabla b_\theta^i \frac{\partial \phi}{\partial x_i}, \tag{4.3}$$

i.e.  $J_\theta = \nabla L_\theta$ .

**Remark 4.1.** Note that  $w_t^\theta = \nabla p_t^\theta$  and that equation (4.2) could be obtained by differentiating formally equation (4.1) with respect to  $\theta$ . This was done indeed in Le Gland (1981), relying on the existence of a “robust” (i.e. continuous with respect to observation sample paths) version of the Zakai equation.

If  $\theta$  is a  $p$ -dimensional parameter, then the gradient  $\{w_t^\theta, 0 \leq t \leq T\}$  is a  $p$ -dimensional vector: each component of this vector actually solves a stochastic PDE which is coupled only with  $\{p_t^\theta, 0 \leq t \leq T\}$  and with no other component; moreover each of the  $(p + 1)$  stochastic PDE’s has the same dynamics. In other words, one has to solve the same stochastic PDE with  $(p + 1)$  different “right-hand side”. Note that smoothing could provide a more efficient method to deal with such a problem.

#### 4.2. The EM algorithm

It follows from (2.3) and (2.4) that the auxiliary function  $Q(\theta, \theta')$  belongs to the class of conditional expectations considered in Section 3. The approach based on filtering (Theorem 3.2) gives

$$Q(\theta, \theta') = (w_T^{\theta, \theta'}, 1) / (p_T^{\theta, \theta'}, 1),$$

with  $\{p_t^{\theta, \theta'}, 0 \leq t \leq T\}$  and  $\{w_t^{\theta, \theta'}, 0 \leq t \leq T\}$  given respectively by (4.1) and, see (3.12),

$$\begin{aligned} dw_t^{\theta, \theta'} &= L_\theta^* w_t^{\theta, \theta'} dt + h_\theta^* w_t^{\theta, \theta'} r^{-1} dY_t + J_{\theta, \theta'}^* p_t^{\theta'} dt \\ &\quad - \frac{1}{2} [b_\theta - b_{\theta'}]^* a^{-1} [b_\theta - b_{\theta'}] p_t^{\theta'} dt \\ &\quad + [h_\theta - h_{\theta'}]^* p_t^{\theta'} r^{-1} dY_t - \frac{1}{2} [h_\theta - h_{\theta'}]^* r^{-1} [h_\theta - h_{\theta'}] p_t^{\theta'} dt, \\ w_0^{\theta, \theta'} &= p_0^{\theta'} \log(p_0^\theta / p_0^{\theta'}), \end{aligned}$$

where

$$J_{\theta, \theta'} \phi \triangleq [b_\theta - b_{\theta'}]^* \phi' = \sum_{i=1}^m [b_\theta^i - b_{\theta'}^i] \frac{\partial \phi}{\partial x_i},$$

i.e.  $J_{\theta, \theta'} = L_\theta - L_{\theta'}$ .

On the other hand, smoothing (Theorem 3.8) gives

$$\begin{aligned} Q(\theta, \theta') &= (\rho_0^{\theta'}, \log(p_0^\theta/p_0^{\theta'})) + \int_0^T \left( \pi_s^{\theta'}, J_{\theta, \theta'} \left( \frac{\rho_s^{\theta'}}{\pi_s^{\theta'}} \right) \right) ds \\ &\quad - \frac{1}{2} \int_0^T (\rho_s^{\theta'}, [b_\theta - b_{\theta'}]^* a^{-1} [b_\theta - b_{\theta'}]) ds \\ &\quad + \int_0^T (\rho_s^{\theta'}, [h_\theta - h_{\theta'}]^* r^{-1} \circ dY_s - \int_0^T (\rho_s^{\theta'}, [h_\theta - h_{\theta'}]^* r^{-1} h_{\theta'}) ds \\ &\quad - \frac{1}{2} \int_0^T (\rho_s^{\theta'}, [h_\theta - h_{\theta'}]^* r^{-1} [h_\theta - h_{\theta'}]) ds, \end{aligned} \tag{4.4}$$

where  $\{\pi_t^{\theta'}, 0 \leq t \leq T\}$  and  $\{\rho_t^{\theta'}, 0 \leq t \leq T\}$  are the normalized filtering and smoothing densities, computed from the unique solutions  $\{p_t^{\theta'}, 0 \leq t \leq T\}$  and  $\{v_t^{\theta'}, 0 \leq t \leq T\}$  of (4.1) and, see (3.8),

$$dv_t^{\theta'} + L_{\theta'} v_t^{\theta'} dt + h_{\theta'}^* v_t^{\theta'} r^{-1} dY_t = 0, \quad v_T^{\theta'} = 1, \tag{4.5}$$

respectively. Moreover, the non-adapted stochastic integral in (4.4) is a generalized Stratonovich integral (Nualart and Pardoux, 1988).

**Remark 4.2.** It is now possible to give a new formulation of the E-step and M-step of the algorithm. Indeed,  $\theta'$  being fixed:

*Step 3 (E-step).* Compute the normalized smoothing density  $\{\rho_t^{\theta'}, 0 \leq t \leq T\}$ : this requires in particular to compute the normalized filtering density  $\{\pi_t^{\theta'}, 0 \leq t \leq T\}$ .

*Step 4 (M-step).* Maximize  $Q(\theta, \theta')$  with respect to  $\theta$ , where for each  $\theta \in \Theta$  the computation of  $Q(\theta, \theta')$  requires according to (4.4) (i) at each time  $t$ , to integrate some functions depending on  $(\theta, \theta')$  against the normalized smoothing density  $\rho_t^{\theta'}$ , and (ii) to integrate the resulting processes over the interval  $[0, T]$ .

**Remark 4.3.** A partial answer can be given to the question (M) raised in the introduction. Indeed:

- the differentiability of  $Q(\theta, \theta')$  with respect to  $\theta$  relies in an obvious way on the existence of derivatives for  $p_0^\theta$ ,  $b_\theta$  and  $h_\theta$ ;
- computing the corresponding derivatives, and maximizing  $Q(\theta, \theta')$  with respect to  $\theta$  will not involve the computation of any other infinite-dimensional conditional density.

Moreover, as was pointed out in Dembo and Zeitouni (1986), there are particular cases in which the M-step can be done explicitly. This includes the case where  $\log p_0^\theta$  depends quadratically on  $\theta$ , and  $b_\theta$  and  $h_\theta$  depend linearly on  $\theta$ .

It follows from (2.6) and (2.7) that  $\nabla^{1,0}Q(\theta, \theta')$  belongs to the class of conditional expectations considered in Section 3. The approach based on filtering (Theorem 3.2) gives

$$\nabla^{1,0}Q(\theta, \theta') = (w_T^{\theta, \theta'}, 1) / (p_T^{\theta'}, 1),$$

with  $\{p_t^{\theta'}, 0 \leq t \leq T\}$  and  $\{w_t^{\theta, \theta'}, 0 \leq t \leq T\}$  given respectively by (4.1) and, see (3.12),

$$\begin{aligned} dw_t^{\theta, \theta'} &= L_{\theta'}^* w_t^{\theta, \theta'} dt + h_{\theta'}^* w_t^{\theta, \theta'} r^{-1} dY_t + J_{\theta'}^* p_t^{\theta'} dt - [\nabla b_{\theta'}]^* a^{-1} [b_{\theta} - b_{\theta'}] p_t^{\theta'} dt \\ &\quad + [\nabla h_{\theta'}]^* p_t^{\theta'} r^{-1} dY_t - [\nabla h_{\theta'}]^* r^{-1} [h_{\theta} - h_{\theta'}] p_t^{\theta'} dt, \\ w_0^{\theta, \theta'} &= (p_0^{\theta'} / p_0^{\theta}) \nabla p_0^{\theta}, \end{aligned}$$

where the first-order partial differential operator  $J_{\theta}$  is defined in (4.3).

**Remark 4.4.** Comparing with (4.2), one can check once again that

$$\nabla^{1,0}Q(\theta, \theta')|_{\theta=\theta'} = \nabla l(\theta').$$

As for the smoothing approach, one can use again the results of Section 3. Alternatively, one can directly differentiate with respect to  $\theta$  the expression (4.4) for  $Q(\theta, \theta')$ , thus illustrating the point (M). This gives

$$\begin{aligned} \nabla^{1,0}Q(\theta, \theta') &= \left( \rho_0^{\theta'}, \frac{\nabla p_0^{\theta}}{p_0^{\theta}} \right) + \int_0^T \left( \pi_s^{\theta'}, J_{\theta} \left( \frac{\rho_s^{\theta'}}{\pi_s^{\theta'}} \right) \right) ds \\ &\quad - \int_0^T (\rho_s^{\theta'}, [\nabla b_{\theta}]^* a^{-1} [b_{\theta} - b_{\theta'}]) ds \\ &\quad + \int_0^T (\rho_s^{\theta'}, [\nabla h_{\theta}]^*) r^{-1} \circ dY_s - \int_0^T (\rho_s^{\theta'}, [\nabla h_{\theta}]^* r^{-1} h_{\theta'}) ds \\ &\quad - \int_0^T (\rho_s^{\theta'}, [\nabla h_{\theta}]^* r^{-1} [h_{\theta} - h_{\theta'}]) ds, \end{aligned}$$

where the non-adapted stochastic integral is a generalized Stratonovich integral (Nualart and Pardoux, 1988).

### 5. Time discretization, and relation with hidden Markov models

In this section, an approximation procedure is described, which allows to actually compute the expressions obtained for  $l(\theta)$ ,  $\nabla l(\theta)$  and  $Q(\theta, \theta')$ . From the results of the previous section, this should reduce in some sense to discretizing the stochastic PDE's (4.1), (4.2) and (4.5).

However, instead of discretizing separately these stochastic PDE's a global approximation of the original continuous time problem by a discrete time problem will be presented. Expressions will be given for the log-likelihood function  $\bar{l}(\theta)$  and the auxiliary function  $\bar{Q}(\theta, \theta')$  associated with the discrete time model, in terms of filtering and smoothing densities. In other words:

- the approximation  $\bar{l}(\theta)$  to the log-likelihood function  $l(\theta)$  of the continuous time problem, will be interpreted as the log-likelihood function of the discrete time problem;

- the approximation  $\bar{Q}(\theta, \theta')$  to the auxiliary function  $Q(\theta, \theta')$  of the continuous time problem will be such that the fundamental relation (1.2) will hold for the discrete time problem, i.e.  $\bar{l}(\theta) - \bar{l}(\theta') \geq \bar{Q}(\theta, \theta')$ .

In addition, the expressions for  $\bar{l}(\theta)$  and  $\bar{Q}(\theta, \theta')$  will be compared with the corresponding expressions for  $l(\theta)$  and  $Q(\theta, \theta')$  in the continuous time model.

5.1. Discrete time statistical model

Let  $\{t_n, 0 \leq n \leq N\}$  be a uniform partition of the interval  $[0, T]$  with time step  $\Delta t$ . Suppose that on a measurable space  $(\Omega, \mathcal{F})$  are given

(a) a family  $\bar{\mathcal{M}} = \{\bar{P}_\theta, \theta \in \Theta\}$  of probability measures,

(b) a pair of stochastic processes  $\{\bar{X}_t, t \geq 0\}$  and  $\{Y_t, t \geq 0\}$  taking values in  $\mathbb{R}^m$  and  $\mathbb{R}^d$  respectively, with  $\{\bar{X}_t, t \geq 0\}$  constant on each time interval  $[t_n, t_{n+1})$ , such that under  $\bar{P}_\theta$ , the discrete time process  $\{x_n, 0 \leq n \leq N\}$  defined by  $x_n \triangleq \bar{X}_{t_n}$  is a Markov chain with transition probabilities kernel

$$\Pi_\theta \triangleq (I - \Delta t L_\theta)^{-1} \tag{5.1}$$

and initial density  $p_\theta^0$ , and

$$dY_t = h_\theta(\bar{X}_t) dt + dV_t,$$

where  $\{V_t, 0 \leq t \leq T\}$  is a Wiener process with covariance matrix  $r$ , independent of the Markov chain  $\{x_n, 0 \leq n \leq N\}$ .

**Remark 5.1.** Equivalently, one can consider that the Markov chain  $\{x_n, 0 \leq n \leq N\}$  is observed through the measurements

$$z_n \triangleq \Delta Y_n / \Delta t = h_\theta(x_n) + v_n, \quad \Delta Y_n \triangleq Y_{t_{n+1}} - Y_{t_n},$$

where  $\{v_n, 0 \leq n \leq N\}$  is a Gaussian white noise sequence with covariance matrix  $r\Delta t^{-1}$ , independent of the Markov chain.

There is no loss in generality in assuming that  $\Omega$  is the *canonical space*  $D([0, T]; \mathbb{R}^m) \times C([0, T]; \mathbb{R}^d)$ , in which case  $X$  and  $Y$  are the *canonical processes* on  $D([0, T]; \mathbb{R}^m)$  and  $C([0, T]; \mathbb{R}^d)$  respectively, and  $P_\theta$  is the probability law of  $(X, Y)$ .

The probability measures in  $\bar{\mathcal{M}}$  are mutually absolutely continuous:

- Define first a probability measure  $\bar{P}_\theta^*$  equivalent to  $\bar{P}_\theta$ , with

$$\bar{Z}^\theta \triangleq \frac{d\bar{P}_\theta}{d\bar{P}_\theta^*} = \prod_{i=0}^{N-1} \Psi_i^\theta(x_i),$$

where

$$\Psi_i^\theta(x) \triangleq \exp\{h_\theta^*(x)r^{-1}\Delta Y_i - \frac{1}{2}h_\theta^*(x)r^{-1}h_\theta(x)\Delta t\}, \tag{5.2}$$

so that under  $\bar{P}_\theta^\dagger$ ,  $\{Y_t, 0 \leq t \leq T\}$  is a Wiener process with covariance matrix  $r$ , independent of the Markov chain  $\{x_n, 0 \leq n \leq N\}$ .

• Next, it follows from hypotheses of Section 2 that  $\forall x \in \mathbb{R}^m, \{\Pi_\theta(x, \cdot), \theta \in \Theta\}$  are mutually absolutely continuous probability measures on  $\mathbb{R}^m$ . Define

$$f_{\theta, \theta'}(x, y) \triangleq \Pi_\theta(x, dy) / \Pi_{\theta'}(x, dy),$$

as the corresponding Radon–Nikodym derivative. Then, the probability measures  $\{\bar{P}_\theta^\dagger, \theta \in \Theta\}$  are mutually absolutely continuous with Radon–Nikodym derivative

$$\bar{A}_{\theta, \theta'}^\dagger \triangleq \frac{d\bar{P}_\theta^\dagger}{d\bar{P}_{\theta'}^\dagger} = \frac{p_0^\theta}{p_0^{\theta'}}(x_0) \prod_{i=0}^{N-1} f_{\theta, \theta'}(x_i, x_{i+1}).$$

Therefore

$$\bar{A}_{\theta, \theta'} \triangleq \frac{d\bar{P}_\theta}{d\bar{P}_{\theta'}} = \bar{A}_{\theta, \theta'}^\dagger \frac{\bar{Z}^\theta}{\bar{Z}^{\theta'}}.$$

Let again  $\{\mathcal{Y}_t, 0 \leq t \leq T\}$  denote the observation filtration. The log-likelihood function for estimating the parameter  $\theta$  in the statistical model  $\bar{\mathcal{M}}$  is defined by

$$\bar{l}(\theta) = \log \bar{E}_\theta^\dagger(\bar{Z}^\theta | \mathcal{Y}_T), \tag{5.3}$$

whereas the auxiliary function is defined by

$$\bar{Q}(\theta, \theta') = \bar{E}_{\theta'}(\log \bar{A}_{\theta, \theta'} | \mathcal{Y}_T) = \bar{E}_{\theta'}^\dagger(\log \bar{A}_{\theta, \theta'} \bar{Z}^{\theta'} | \mathcal{Y}_T) / \bar{E}_{\theta'}^\dagger(\bar{Z}^{\theta'} | \mathcal{Y}_T). \tag{5.4}$$

### 5.2. Direct maximization of the likelihood function

The idea is to find an equation for  $\{\bar{p}_n^\theta, 0 \leq n \leq N\}$  defined by

$$(\bar{p}_n^\theta, \phi) \triangleq \bar{E}_\theta^\dagger(\phi(x_n) \bar{Z}_n^\theta | \mathcal{Y}_{t_n}),$$

where

$$\bar{Z}_n^\theta \triangleq \prod_{i=0}^{n-1} \Psi_i^\theta(x_i).$$

By definition

$$\begin{aligned} (\bar{p}_{n+1}^\theta, \phi) &= \bar{E}_\theta^\dagger(\phi(x_{n+1}) \bar{Z}_{n+1}^\theta | \mathcal{Y}_{t_{n+1}}) = \bar{E}_\theta^\dagger(\phi(x_{n+1}) \Psi_n^\theta(x_n) \bar{Z}_n^\theta | \mathcal{Y}_{t_{n+1}}) \\ &= \bar{E}_\theta^\dagger(\Psi_n^\theta(x_n) [\Pi_\theta \phi](x_n) \bar{Z}_n^\theta | \mathcal{Y}_{t_{n+1}}) = (\bar{p}_n^\theta, \Psi_n^\theta[\Pi_\theta \phi]), \end{aligned}$$

which results in the following equation

$$\bar{p}_{n+1}^\theta = \Pi_\theta^*(\Psi_n^\theta \bar{p}_n^\theta), \quad \bar{p}_0^\theta = p_0^\theta. \tag{5.5}$$

Using expression (5.1) for the transition probabilities kernel gives the following discretization scheme for the Zakai equation (4.1), which combines a Trotter-like product formula and a Euler implicit scheme

$$(I - \Delta t L_\theta^*) \bar{p}_{n+1}^\theta = \Psi_n^\theta \bar{p}_n^\theta, \quad \bar{p}_0^\theta = p_0^\theta. \tag{5.6}$$

It follows from (5.3) that the log-likelihood function  $l(\theta)$  is therefore approximated by

$$\bar{l}(\theta) = \log(\bar{p}_N^\theta, 1).$$

To approximate the gradient  $\nabla l(\theta)$ , one could either discretize directly equation (4.2), or derive the exact expression for the gradient of the approximated log-likelihood function  $\bar{l}(\theta)$ . The second method is preferred, and gives

$$\nabla \bar{l}(\theta) = (\bar{w}_N^\theta, 1) / (\bar{p}_N^\theta, 1),$$

where  $\{\bar{w}_n^\theta, 0 \leq n \leq N\}$  is defined by  $\bar{w}_n^\theta \triangleq \nabla \bar{p}_n^\theta$  and satisfies, deriving equation (5.6) with respect to the parameter  $\theta$ ,

$$(I - \Delta t L_\theta^*) \bar{w}_{n+1}^\theta = \Psi_n^\theta \bar{w}_n^\theta + J_\theta^* \bar{p}_{n+1}^\theta \Delta t + [\nabla \Psi_n^\theta] \bar{p}_n^\theta, \quad \bar{w}_0^\theta = \nabla p_0^\theta,$$

where the first-order partial differential operator  $J_\theta = \nabla L_\theta$  is defined in (4.3). Using the notation  $\bar{p}_{n+1/2}^\theta \triangleq \Psi_n^\theta \bar{p}_n^\theta$ , and the identity

$$\nabla \log \Psi_n^\theta = [\nabla h_\theta]^* r^{-1} (\Delta Y_n - h_\theta \Delta t),$$

gives

$$(I - \Delta t L_\theta^*) \bar{w}_{n+1}^\theta = \Psi_n^\theta \bar{w}_n^\theta + J_\theta^* \bar{p}_{n+1}^\theta \Delta t + [\nabla h_\theta]^* \bar{p}_{n+1/2}^\theta r^{-1} (\Delta Y_n - h_\theta \Delta t),$$

to be compared with (4.2).

**Remark 5.2** (normalization). To avoid numerical overflow, one should rather use normalized quantities, along the following steps:

- (i)  $j_{n+1}^\theta = (\bar{\pi}_n^\theta, \Psi_n^\theta)$ ,
- (ii)  $\bar{\pi}_{n+1/2}^\theta = \Psi_n^\theta \bar{\pi}_n^\theta / j_{n+1}^\theta$ ,
- (iii)  $(I - \Delta t L_\theta^*) \bar{\pi}_{n+1}^\theta = \bar{\pi}_{n+1/2}^\theta$ .

It is easily seen that  $\bar{p}_n^\theta = \gamma_n^\theta \bar{\pi}_n^\theta$  with  $\gamma_n^\theta \triangleq j_n^\theta \cdot j_{n-1}^\theta \cdot \dots \cdot j_1^\theta$  and  $(\bar{p}_n^\theta, 1) = \gamma_n^\theta$  so that

$$\bar{l}(\theta) = \log \gamma_N^\theta = \sum_{n=1}^N \log j_n^\theta.$$

In the same way, the computation of  $\bar{\alpha}_n^\theta$  defined by the relation  $\bar{w}_n^\theta = \gamma_n^\theta \bar{\alpha}_n^\theta$  can be achieved along the following steps:

- (i)  $\bar{\alpha}_{n+1/2}^\theta = \Psi_n^\theta / j_{n+1}^\theta$ ,
- (ii)  $(I - \Delta t L_\theta^*) \bar{\alpha}_{n+1}^\theta = \bar{\alpha}_{n+1/2}^\theta + J_\theta^* \bar{\pi}_{n+1}^\theta \Delta t + [\nabla h_\theta]^* \bar{\pi}_{n+1/2}^\theta r^{-1} (\Delta Y_n - h_\theta \Delta t)$ .

Note that, although  $\bar{w}_n^\theta$  is the gradient of  $\bar{p}_n^\theta$ ,  $\bar{\alpha}_n^\theta$  is *not* the gradient of  $\bar{\pi}_n^\theta$ . Actually  $\bar{\alpha}_n^\theta = \bar{w}_n^\theta / (\bar{p}_n^\theta, 1)$  so that

$$\nabla \bar{l}(\theta) = (\bar{\alpha}_N^\theta, 1).$$

5.3. The EM algorithm

Although it is rather straightforward, in the discrete time case, to obtain the expression of the auxiliary function  $\bar{Q}(\theta, \theta')$  in terms of nonlinear smoothing, it is nevertheless worth presenting a derivation that follows the same lines as in the continuous time case. Indeed, there are two different methods—one based on nonlinear filtering, the other on nonlinear smoothing—for the computation of (5.4).

*Filtering:* Define

$$\bar{\lambda}_n^{\theta, \theta'} \triangleq \log \frac{P_0^\theta}{p_0^{\theta'}}(x_0) + \sum_{i=0}^{n-1} \log f_{\theta, \theta'}(x_i, x_{i+1}) + \sum_{i=0}^{n-1} \log \frac{\Psi_i^\theta}{\Psi_i^{\theta'}}(x_i).$$

The idea again is to find an equation for  $\{\bar{w}_n^{\theta, \theta'}, 0 \leq n \leq N\}$  defined by

$$(\bar{w}_n^{\theta, \theta'}, \phi) \triangleq \bar{E}_{\theta'}^+(\phi(x_n) \bar{\lambda}_n^{\theta, \theta'} \bar{Z}_n^{\theta'} | \mathcal{Y}_{t_n}).$$

First

$$\bar{w}_0^{\theta, \theta'} = p_0^{\theta'} \log \frac{P_0^\theta}{p_0^{\theta'}}.$$

Next, by definition

$$\begin{aligned} (\bar{w}_{n+1}^{\theta, \theta'}, \phi) &= \bar{E}_{\theta'}^+(\phi(x_{n+1}) \bar{\lambda}_{n+1}^{\theta, \theta'} \bar{Z}_{n+1}^{\theta'} | \mathcal{Y}_{t_{n+1}}) \\ &= \bar{E}_{\theta'}^+ \left( \phi(x_{n+1}) \Psi_n^{\theta'}(x_n) \right. \\ &\quad \left. \times \left[ \bar{\lambda}_n^{\theta, \theta'} + \log f_{\theta, \theta'}(x_n, x_{n+1}) + \log \frac{\Psi_n^\theta}{\Psi_n^{\theta'}}(x_n) \right] \bar{Z}_n^{\theta'} | \mathcal{Y}_{t_{n+1}} \right) \\ &= \bar{E}_{\theta'}^+(\Psi_n^{\theta'}(x_n) [\Pi_{\theta'} \phi](x_n) \bar{\lambda}_n^{\theta, \theta'} \bar{Z}_n^{\theta'} | \mathcal{Y}_{t_{n+1}}) \\ &\quad + \bar{E}_{\theta'}^+(\Psi_n^{\theta'}(x_n) [\kappa_{\theta, \theta'} \phi](x_n) \bar{Z}_n^{\theta'} | \mathcal{Y}_{t_{n+1}}) \\ &\quad + \bar{E}_{\theta'}^+ \left( \Psi_n^{\theta'}(x_n) \log \frac{\Psi_n^\theta}{\Psi_n^{\theta'}}(x_n) [\Pi_{\theta'} \phi](x_n) \bar{Z}_n^{\theta'} | \mathcal{Y}_{t_{n+1}} \right) \\ &= (\bar{w}_n^{\theta, \theta'}, \Psi_n^{\theta'} [\Pi_{\theta'} \phi]) + (\bar{p}_n^{\theta'}, \Psi_n^{\theta'} [\kappa_{\theta, \theta'} \phi]) \\ &\quad + \left( \bar{p}_n^{\theta'}, \Psi_n^{\theta'} \log \frac{\Psi_n^\theta}{\Psi_n^{\theta'}} [\Pi_{\theta'} \phi] \right), \end{aligned}$$

where the operator  $\kappa_{\theta, \theta'}$  is defined by

$$[\kappa_{\theta, \theta'} \phi](x) \triangleq \int \phi(y) \log f_{\theta, \theta'}(x, y) \Pi_{\theta'}(x, dy). \tag{5.7}$$

Therefore, the resulting equation is

$$\begin{aligned} \bar{w}_{n+1}^{\theta, \theta'} &= \Pi_{\theta'}^*(\Psi_n^{\theta'} \bar{w}_n^{\theta, \theta'}) + \kappa_{\theta, \theta'}^*(\Psi_n^{\theta'} \bar{p}_n^{\theta'}) + \Pi_{\theta'}^* \left( \Psi_n^{\theta'} \log \frac{\Psi_n^\theta}{\Psi_n^{\theta'}} \bar{p}_n^{\theta'} \right), \\ \bar{w}_0^{\theta, \theta'} &= p_0^{\theta'} \log \frac{P_0^\theta}{p_0^{\theta'}}. \end{aligned}$$

It follows from (5.4) that the auxiliary function  $Q(\theta, \theta')$  is approximated by

$$\bar{Q}(\theta, \theta') = (\bar{w}_N^{\theta, \theta'}, 1) / (\bar{p}_N^{\theta'}, 1).$$

**Smoothing:** Introduce the backward equation, dual to (5.5),

$$\bar{v}_n^{\theta'} = \Psi_n^{\theta'}[\Pi_{\theta'} \bar{v}_{n+1}^{\theta'}], \quad \bar{v}_N^{\theta'} \equiv 1. \quad (5.8)$$

Then

$$\begin{aligned} (\bar{w}_{n+1}^{\theta, \theta'}, \bar{v}_{n+1}^{\theta'}) &= (\Pi_{\theta'}^* (\Psi_n^{\theta'} \bar{w}_n^{\theta, \theta'}), \bar{v}_{n+1}^{\theta'}) + (\kappa_{\theta, \theta'}^* (\Psi_n^{\theta'} \bar{p}_n^{\theta'}), \bar{v}_{n+1}^{\theta'}) \\ &\quad + \left( \Pi_{\theta'}^* \left( \Psi_n^{\theta'} \log \frac{\Psi_n^{\theta'}}{\Psi_{\theta'}^{\theta'}} \bar{p}_n^{\theta'} \right), \bar{v}_{n+1}^{\theta'} \right) \\ &= (\bar{w}_n^{\theta, \theta'}, \bar{v}_n^{\theta'}) + (\Psi_n^{\theta'} \bar{p}_n^{\theta'}, [\kappa_{\theta, \theta'} \bar{v}_{n+1}^{\theta'}]) + \left( \bar{p}_n^{\theta'} \bar{v}_n^{\theta'}, \log \frac{\Psi_n^{\theta'}}{\Psi_{\theta'}^{\theta'}} \right). \end{aligned}$$

Introducing the unnormalized smoothing density  $\bar{q}_i^{\theta'} = \bar{p}_i^{\theta'} \bar{v}_i^{\theta'}$ , gives

$$\begin{aligned} (\bar{w}_N^{\theta, \theta'}, 1) &= (\bar{w}_0^{\theta, \theta'}, \bar{v}_0^{\theta'}) + \sum_{i=0}^{N-1} [(\bar{w}_{i+1}^{\theta, \theta'}, \bar{v}_{i+1}^{\theta'}) - (\bar{w}_i^{\theta, \theta'}, \bar{v}_i^{\theta'})] \\ &= \left( \bar{q}_0^{\theta'}, \log \frac{p_0^{\theta}}{p_0^{\theta'}} \right) + \sum_{i=0}^{N-1} (\Psi_i^{\theta'} \bar{p}_i^{\theta'}, [\kappa_{\theta, \theta'} \bar{v}_{i+1}^{\theta'}]) + \sum_{i=0}^{N-1} \left( \bar{q}_i^{\theta'}, \log \frac{\Psi_i^{\theta}}{\Psi_{\theta'}^{\theta'}} \right). \end{aligned}$$

**Remark 5.3.** This expression could be derived in a straightforward way, directly from the definitions. This is in opposition with the continuous time case, where the smoothing approach is more complicated to deal with, because of non-adapted stochastic integrals.

Using the notation  $\bar{p}_{i+1/2}^{\theta'} \triangleq \Psi_i^{\theta'} \bar{p}_i^{\theta'}$ , and the identity

$$\log \frac{\Psi_i^{\theta}}{\Psi_i^{\theta'}} = [h_{\theta} - h_{\theta'}]^* r^{-1} (\Delta Y_i - h_{\theta'} \Delta t) - \frac{1}{2} [h_{\theta} - h_{\theta'}]^* r^{-1} [h_{\theta} - h_{\theta'}] \Delta t,$$

gives

$$\begin{aligned} (\bar{w}_N^{\theta, \theta'}, 1) &= \left( \bar{q}_0^{\theta'}, \log \frac{p_0^{\theta}}{p_0^{\theta'}} \right) + \sum_{i=0}^{N-1} (\bar{p}_{i+1/2}^{\theta'}, [\kappa_{\theta, \theta'} \bar{v}_{i+1}^{\theta'}]) \\ &\quad + \sum_{i=0}^{N-1} (\bar{q}_i^{\theta'}, [h_{\theta} - h_{\theta'}]^*) r^{-1} \Delta Y_i \\ &\quad - \sum_{i=0}^{N-1} (\bar{q}_i^{\theta'}, [h_{\theta} - h_{\theta'}]^* r^{-1} h_{\theta'}) \Delta t \\ &\quad - \frac{1}{2} \sum_{i=0}^{N-1} (\bar{q}_i^{\theta'}, [h_{\theta} - h_{\theta'}]^* r^{-1} [h_{\theta} - h_{\theta'}]) \Delta t. \end{aligned}$$

In terms of normalized conditional densities

$$\begin{aligned} \bar{Q}(\theta, \theta') &= \left( \bar{\rho}_0^{\theta'}, \log \frac{p_0^{\theta}}{p_0^{\theta'}} \right) + \sum_{i=0}^{N-1} \left( \bar{\pi}_{i+1/2}^{\theta'}, \kappa_{\theta, \theta'} \left( \frac{\bar{\rho}_{i+1}^{\theta'}}{\bar{\pi}_{i+1}^{\theta'}} \right) \right) \\ &\quad + \sum_{i=0}^{N-1} (\bar{\rho}_i^{\theta'}, [h_{\theta} - h_{\theta'}]^*) r^{-1} \Delta Y_i \\ &\quad - \sum_{i=0}^{N-1} (\bar{\rho}_i^{\theta'}, [h_{\theta} - h_{\theta'}]^* r^{-1} h_{\theta'}) \Delta t \end{aligned}$$

$$-\frac{1}{2} \sum_{i=0}^{N-1} (\bar{\rho}_i^{\theta'}, [h_\theta - h_{\theta'}]^* r^{-1} [h_\theta - h_{\theta'}]) \Delta t,$$

to be compared with (4.4).

**Remark 5.4** (normalization). Here again, one should rather use normalized quantities, along the following steps:

- (i)  $\bar{\zeta}_{n+1/2}^{\theta'} = \Pi_{\theta'} \bar{\zeta}_{n+1}^{\theta'}$ ,
- (ii)  $k_n^{\theta'} = (\bar{\pi}_n^{\theta'}, \Psi_n^{\theta'} \bar{\zeta}_{n+1/2}^{\theta'})$ ,
- (iii)  $\bar{\zeta}_n^{\theta'} = \Psi_n^{\theta'} \bar{\zeta}_{n+1/2}^{\theta'} / k_n^{\theta'}$ ,

in such a way that  $(\bar{\pi}_n^{\theta'}, \bar{\zeta}_n^{\theta'}) = 1$ . It is easily seen that  $k_n^{\theta'} = j_{n+1}^{\theta'}$ , and that  $\bar{v}_n^{\theta'} = \delta_n^{\theta'} \bar{\zeta}_n^{\theta'}$  with  $\delta_n^{\theta'} \triangleq k_n^{\theta'} \cdot k_{n+1}^{\theta'} \cdot \dots \cdot k_N^{\theta'}$ . Moreover, the normalized smoothing density satisfies  $\bar{\rho}_n^{\theta'} = \bar{\pi}_n^{\theta'} \bar{\zeta}_n^{\theta'}$ .

**Remark 5.5.** It is now possible to give a new formulation of the E-step and M-step of the algorithm. Indeed,  $\theta'$  being fixed:

*Step 3* (E-step). Compute the normalized smoothing density  $\{\bar{\rho}_n^{\theta'}, 0 \leq n \leq N\}$ : this requires in particular to compute the normalized filtering density  $\{\bar{\pi}_n^{\theta'}, 0 \leq n \leq N\}$ .

*Step 4* (M-step). Maximize  $\bar{Q}(\theta, \theta')$  with respect to  $\theta$ -where for each  $\theta \in \Theta$  the computation of  $\bar{Q}(\theta, \theta')$  requires (i) at each time step  $n$ , to integrate some functions depending on  $(\theta, \theta')$  against the normalized smoothing density  $\bar{\rho}_n^{\theta'}$ , and (ii) to sum the resulting discrete time processes from  $n = 0$  to  $n = N - 1$ .

**Remark 5.6.** With the time discretization scheme introduced above, the numerical implementation (including discretization with respect to the space variable) of the EM algorithm requires in the M-step, the explicit evaluation of the transition probabilities kernel  $\Pi_\theta = (I - \Delta t L_\theta)^{-1}$ . On the other hand, the numerical implementation of the direct maximization algorithm requires only the solution of linear equations with operator  $(I - \Delta t L_\theta^*)$ , which is a much faster task.

5.4. Relation with hidden Markov models

The discrete time EM algorithm, as described above, can be seen as a *parametric* version of the Baum–Welch algorithm used in statistical estimation of probabilistic functions of Markov processes. This algorithm was introduced by Baum (1971), and has found interesting applications in the field of acoustic speech recognition, as reported in Levinson, Rabiner and Sondhi (1983).

A hidden Markov model (HMM) is defined by (i) a Markov chain with initial probability distribution  $\pi$  and transition probabilities kernel  $A$ , and (ii) for each possible state  $x$  of the non-observed Markov chain, a probability function  $B(x, \cdot)$  which represents the conditional distribution of the observation given that the chain is in state  $x$ . Such a model will be denoted by  $M = (\pi, A, B)$ . Then (under the

additional assumption that both the Markov chain and the observation sequence take their values in finite sets), the maximum likelihood estimation of the parameters  $(\pi, A, B)$  of the hidden Markov model  $\mathbf{M}$  can be achieved by the EM algorithm, involving an auxiliary function  $Q(\mathbf{M}, \mathbf{M}')$ . What makes the approach interesting is that maximizing  $Q(\mathbf{M}, \mathbf{M}')$  with respect to  $\mathbf{M}$  provides explicit formulas—*reestimation formulas* (Baum, 1971; Levinson, Rabiner and Sondhi, 1983)—for  $(\pi, A, B)$  in terms of  $(\pi', A', B')$ .

Consider now the parametric model described above. It is possible to turn it into a parametric hidden Markov model  $\mathbf{M}_\theta = (p_\theta^0, \Pi_\theta, B_\theta)$  with

$$B_\theta(x, z) = (2\pi)^{-d/2}(\det r)^{-1/2} \exp\{-\frac{1}{2}[h_\theta(x) - z]^* r^{-1}[h_\theta(x) - z]\Delta t\}.$$

Comparing with (5.2) shows that

$$B_\theta(x, z_n) = (2\pi)^{-d/2}(\det r)^{-1/2} \exp\{-\frac{1}{2}z_n^* r^{-1} z_n \Delta t\} \Psi_n^\theta(x).$$

It is easily seen that the auxiliary function defined in (5.4) satisfies  $\bar{Q}(\theta, \theta') = Q(\mathbf{M}_\theta, \mathbf{M}_{\theta'})$ , and that equations (5.5) and (5.8) are parametrized versions of Baum’s forward and backward equations (Baum, 1971; Levinson, Rabiner and Sondhi, 1983). On the other hand, maximization of  $\bar{Q}(\theta, \theta')$  with respect to  $\theta$  is not explicit in general, in opposition to the non-parametric case.

### 6. Numerical example

The continuous time model is described by

$$\begin{aligned} dX_t &= -\theta_2 X_t dt + \theta_3 \frac{X_t}{1 + X_t^2} dt + \sqrt{a} dW_t, \quad X_0 \sim \mathcal{N}(\theta_1, \Sigma), \\ dY_t &= \theta_4 \arctan(X_t/\theta_4) dt + \sqrt{r} dV_t, \end{aligned} \tag{6.1}$$

and the unknown parameter is  $\theta = (\theta_1, \theta_2, \theta_3, \theta_4)$ . The noise covariances in the problem are  $\Sigma$ ,  $a$  and  $r$ , and can be associated with the parameters  $\theta_1, (\theta_2, \theta_3)$  and  $\theta_4$  respectively.

Although the unknown parameter is actually four-dimensional, results are presented in Campillo and Le Gland (1988) for the estimation of one component of  $\theta$  at a time, and the influence of the “associated” noise covariance is investigated.

The numerical experiment can be described as follows:

*Parameters:* The “true” value of the parameter—i.e. the value used for simulating sample paths of the observation process—is  $(\theta_1^*, \theta_2^*, \theta_3^*, \theta_4^*) = (1.0, 0.25, 5.0, 2.0)$ .

*Simulations:* The time interval is  $[0, T]$  with  $T = 10.0$  and time step  $\Delta t = 0.1$ . Observation process sample paths are simulated in the following way. First, the

Euler time discretization scheme (equivalent on this particular example to the Milshtein scheme) is used to simulate the signal process in (6.1),

$$x_{n+1} = x_n + \left[ -\theta_2 x_n + \theta_3 \frac{x_n}{1 + x_n^2} \right] \Delta t + w_n,$$

with  $x_0 \sim \mathcal{N}(\theta_1, \Sigma)$  and  $\{w_n, 0 \leq n \leq N\}$  a Gaussian white noise sequence with covariance matrix  $a\Delta t$ . Next, discrete measurements are generated according to

$$z_n = \theta_4 \arctan(x_n / \theta_4) + v_n,$$

with  $\{v_n, 0 \leq n \leq N\}$  a Gaussian white noise sequence with covariance matrix  $r\Delta t^{-1}$  independent of  $\{w_n, 0 \leq n \leq N\}$ .

*Algorithms:* The discrete measurements are used to solve equations (5.5) and (5.8), and therefore to compute the approximations  $\bar{I}(\theta)$  and  $\bar{Q}(\theta, \theta')$  defined by (5.3) and (5.4) respectively. Actually, equations (5.5) and (5.8) are discretized with respect to the space variable, using the finite difference schemes described in Kushner (1977).

In order to find the MLE, the log-likelihood function is maximized using either the direct approach or the EM algorithm based on nonlinear smoothing, relying on the minimization routine `e04jbf` from the NAG library, which uses a quasi-Newton algorithm and does not require the user to provide a routine for the computation of the gradient.

**Remark 6.1.** In the example introduced above, although the auxiliary function  $Q(\theta, \theta')$  of the continuous time model depends quadratically on the parameters  $\theta_1, \theta_2$  and  $\theta_3$ , the discrete time approximation  $\bar{Q}(\theta, \theta')$  depends quadratically on  $\theta_1$  only. This can be seen on the expression of the operator  $\kappa_{\theta, \theta'}$ , see (5.7).

*Results:* The results presented below are for the estimation of  $\theta_3$ . The other three parameters are frozen:  $(\theta_1, \theta_2, \theta_4) = (\theta_1^*, \theta_2^*, \theta_4^*)$ . Two cases are considered.

- In the first case (Fig. 1 and Fig. 2) the numerical values of the noise covariances are:  $\Sigma = 1.0, a = 1.0$  and  $r = 1.0$ . In this case, the EM algorithm has converged after 5 iterations.

- In the second case (Fig. 3 and Fig. 4) the numerical values of the noise covariances are:  $\Sigma = 1.0, a = 0.01$  and  $r = 1.0$ . In this case, the EM algorithm has not yet converged after 200 iterations. Therefore, only 12 iterations are shown on Fig. 4.

The two figures related to the direct maximization of the likelihood function (Fig. 1 and Fig. 3) show:

- *in solid line:* the log-likelihood function vs. the free parameter;
- *in dashed line:* iterations of the quasi-Newton algorithm for the direct log-likelihood function maximization.

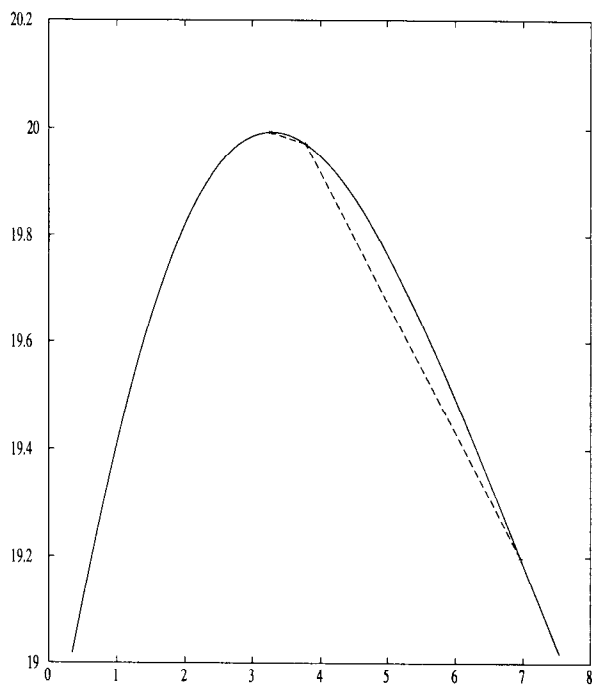


Fig. 1. Direct maximization ( $a = 1.0$ ).

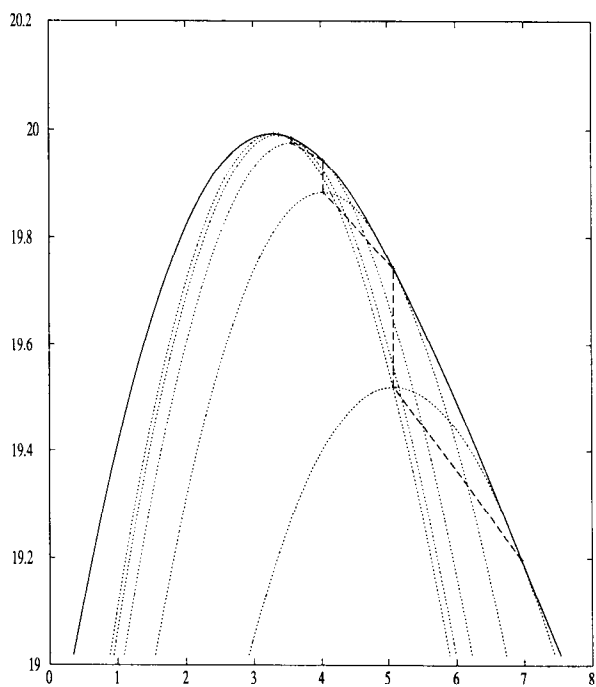


Fig. 2. EM algorithm ( $a = 1.0$ ).

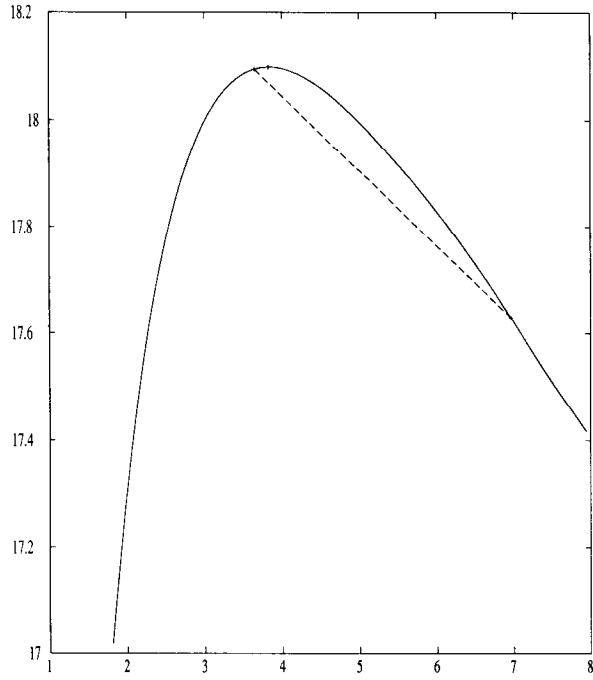


Fig. 3. Direct maximization ( $a = 0.01$ ).

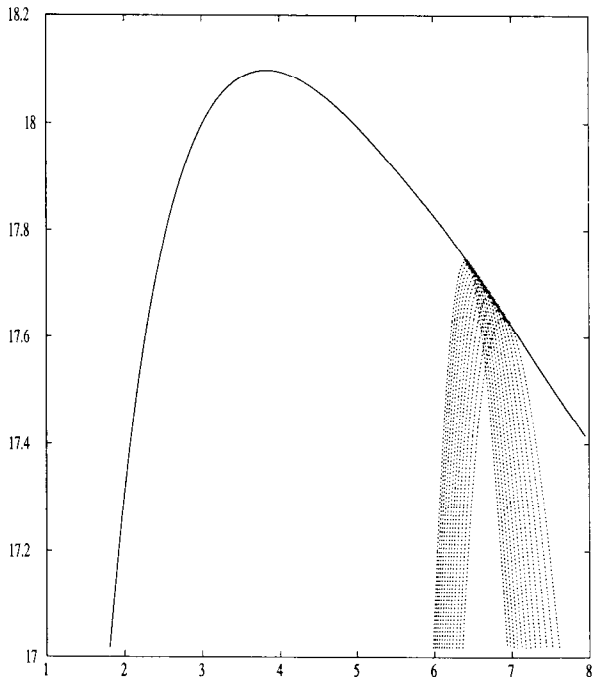


Fig. 4. EM algorithm ( $a = 0.01$ ).

The two figures related to the EM algorithm (Fig. 2 and Fig. 4) show:

- *in solid line*: the log-likelihood function vs. the free parameter;
- *in dotted lines*: the auxiliary functions corresponding to successive estimates;
- *in dashed lines*: iterations of the EM algorithm.

Other results can be found in Campillo and Le Gland (1988) for the estimation of  $\theta_1$  and  $\theta_4$ . The conclusion of these numerical experiments is that the EM algorithm converges very slowly whenever some noise covariances associated with the parameters to be estimated are small. The reason why is that the log-likelihood function is then approximated from below by a set of very sharp auxiliary functions: this situation does not allow to update significantly enough the current estimate at each M-step. Actually, this can be seen directly from (2.3), (2.4), or equivalently from (4.4). Assume for instance that both  $p_0^\theta$  and  $h_\theta$  are independent of  $\theta$ , and that the signal noise covariance is small; then every auxiliary function  $Q(\theta, \theta')$  will certainly be very sharp as a function of  $\theta$ . It should be stressed that in such cases, the slow variation of the estimate should not be interpreted as an indication that the algorithm has already achieved convergence.

## 7. Conclusion

The direct maximization of the log-likelihood function has been compared with the EM algorithm, for the MLE of parameters in partially observed diffusion processes. Some formulas given in Dembo and Zeitouni (1986) have been clarified, and it has been shown that smoothing is necessary to make the EM algorithm approach efficient. On the other hand, formulas have been given in terms of filtering stochastic PDE's for the computation of the original log-likelihood function and its gradient.

It has been shown that

(E) The E-step in the EM algorithm is certainly slower than the direct computation of the log-likelihood function, since it involves nonlinear smoothing instead of nonlinear filtering.

(M) The computation of the auxiliary function  $Q(\theta, \theta')$  in the M-step of the EM algorithm,  $\theta'$  being fixed, requires (i) at each time  $t$ , to integrate some functions depending on  $(\theta, \theta')$  against a normalized smoothing density depending only on  $\theta'$ , and (ii) to integrate the resulting processes over the interval  $[0, T]$ . This gives another evidence that the EM algorithm is more complicated than the direct approach as far as computations are concerned. On the other hand, the maximization of the auxiliary function is generally simple to deal with, and in some cases can even be done explicitly.

(EM) The EM algorithm converges very slowly whenever some noise covariances associated with the parameters to be estimated are small.

However, the EM algorithm should provide an interesting approach for non-parametric estimation in the context of partially observed diffusion processes, i.e.

non-parametric estimation of the initial density, the drift and the observation function. This form of the EM algorithm is used indeed in the context of finite-space Markov chains with finite-state observations (hidden Markov models), and leads to well-known reestimation formulas, which are of practical use e.g. in the field of acoustic speech recognition.

### Acknowledgements

The authors gratefully acknowledge Etienne Pardoux for contributing to the proof of Proposition 3.5, Monique Pontier, Ofer Zeitouni and the anonymous referee for their valuable comments.

### References

- L.E. Baum, An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes, in: O. Shisha, ed., *Inequalities III* (Academic Press, New York, 1971) pp. 1–8.
- F. Campillo and F. Le Gland, MLE for partially observed diffusions: direct maximization vs. the EM algorithm, INRIA Report 884, Valbonne, August 1988.
- A. Dembo and O. Zeitouni, Parameter estimation of partially observed continuous-time stochastic processes via the EM algorithm, *Stochastic Process. Appl.* 23 (1986) 91–113.
- A.P. Dempster, N. M. Laird and D. B. Rubin, Maximum likelihood estimation from incomplete data via the EM algorithm, *J. Roy. Statist. Soc. Ser. B* 39 (1977) 1–38.
- M.R. James and F. Le Gland, Consistent parameter estimation for partially observed diffusions with small noise, in preparation.
- N.V. Krylov and B.L. Rozovskii, On the Cauchy problem for linear stochastic partial differential equations, *Math. USSR Izv.* 11 (1977) 1267–1284.
- H.J. Kushner, *Probability Methods for Approximations in Stochastic Control and for Elliptic Equations* (Academic Press, New York, 1977).
- F. Le Gland, Estimation de paramètres dans les processus stochastiques en observation incomplète, Thèse de Docteur-Ingénieur, Univ. de Paris IX-Dauphine, 1981.
- S.E. Levinson, L.R. Rabiner and M.M. Sondhi, An introduction to the application of the theory of probabilistic functions of a Markov process to automatic speech recognition, *Bell System Tech. J.* 62 (1983) 1035–1074.
- R.S. Liptser and A.N. Shiriyayev, *Statistics of Random Processes: General Theory* (Springer, Berlin, 1977).
- D. Nualart and E. Pardoux, Stochastic calculus with anticipating integrands, *Probab. Theory Rel. Fields* 78 (1988) 535–581.
- D. Ocone, Stochastic calculus of variations for stochastic partial differential equations, *J. Funct. Anal.* 79 (1988) 288–331.
- E. Pardoux, Stochastic PDE's and filtering of diffusion processes, *Stochastics* 3 (1979) 127–167.
- E. Pardoux, Equations du lissage non-linéaire, in: H. Korezlioglu, G. Mazziotto and J. Szpirglas, eds., *Filtering and Control of Random Processes* (Springer, Berlin, 1984) pp. 206–218.
- E. Pardoux, Two-sided stochastic calculus for SPDE's in: G. DaPrato and L. Tubaro, eds., *Stochastic PDE's and Applications* (Springer, Berlin, 1987) pp. 200–207.
- E. Pardoux and Ph. Protter, A two-sided stochastic integral and its calculus, *Probab. Theory Rel. Fields* 76 (1987) 15–49.
- A.S. Sznitman, Martingales dépendant d'un paramètre: une formule d'Itô, *Z. Wahrsch. Verw. Gebiete* 60 (1982) 41–70.
- C.F.J. Wu, On the convergence properties of the EM algorithm, *Ann. Statist.* 11 (1983) 95–103.