

Stability and Approximation of Nonlinear Filters : an Information Theoretic Approach*

François LeGland
IRISA / INRIA
Campus de Beaulieu
35042 RENNES Cédex, France
legland@irisa.fr

Abstract

It has recently been proved by Clark, Ocone and Coumarbatch that the relative entropy (or Kullback-Leibler information distance) between two nonlinear filters with different initial conditions is a supermartingale, hence its expectation can only decrease with time. This result was obtained for a very general model, where the unknown state and observation processes form jointly a continuous-time Markov process.

The purpose of this paper is (i) to extend this result to a large class of f -divergences, including the total variation distance, the Hellinger distance, and not only the Kullback-Leibler information distance, and (ii) to consider not only robustness w.r.t. the initial condition of the filter, but also w.r.t. perturbation of the state generator. On the other hand, the model considered here is much less general, and consists of a diffusion process observed in discrete-time.

Keywords : nonlinear filtering, stability, relative entropy, Kullback-Leibler information, Hellinger distance, total variation distance, f -divergence.

1 Introduction

The dependency of a nonlinear filter w.r.t. its initial condition, has been recently the subject of considerable attention. Pathwise exponential decay has been proved under restrictive assumptions which in practice hold provided the state space is either finite, see Atar and Zeitouni [4], Le Gland and Mevel [14, 15], or compact, see Atar and Zeitouni [3], Del Moral and Guionnet [11], Budhiraja and Kushner [7]. Similar results have been obtained for special cases with noncompact state space and sufficiently small observation noise, see Atar [1], Budhiraja and Ocone [8]. A slightly different approach has been taken by Atar, Viens and Zeitouni [2]. On

*This work was partially supported by the Army Research Office, under grant DAAH04-95-1-0164, by the CNRS programme *Modélisation et Simulation Numérique*, under grant 97N23/0019, and by the French *Ministère des Affaires Etrangères*.

the other hand, mean-square decay without compactness assumption has been proved by Ocone and Pardoux [17], under ergodicity of the state process. Stability of the relative entropy has been proved for a very general model, without compactness or ergodicity assumption, by Clark, Ocone and Coumarbatch [9]. Some generalizations of the last mentioned work are obtained here, for the following less general model however.

The unobserved state process $\{X_t, t \geq 0\}$ is the solution of the following stochastic differential equation (SDE) on \mathbb{R}^m

$$dX_t = b_t(X_t) dt + \sigma_t(X_t) dW_t, \quad X_0 \sim \mu_0(dx), \quad (1)$$

where $\{W_t, t \geq 0\}$ is a Wiener process of appropriate dimension, with identity covariance matrix, independent of the initial state X_0 . With (1) is associated the following time-dependent second-order partial differential operator

$$L_t = \frac{1}{2} a_t^{i,j} \frac{\partial^2}{\partial x_i \partial x_j} + b_t^i \frac{\partial}{\partial x_i}, \quad (2)$$

where $a_t = (a_t^{i,j}) = \sigma_t \sigma_t^*$, and where the convention of summation on repeated indices is used. Let $\{P_t^s, t \geq s \geq 0\}$ be the Markov semigroup generated by $\{L_t, t \geq 0\}$.

At discrete time instants

$$0 = t_0 < \dots < t_n < \dots \quad (3)$$

a d -dimensional noisy observation z_n of the state X_{t_n} becomes available, with conditional probability distribution

$$P\{z_n \in dz \mid X_{t_n} = x\} = g_n(x, z) dz.$$

The time-dependent transition probability kernel for the sampled Markov chain $\{X_{t_n}, n \geq 0\}$ is defined by $Q_n = P_{t_n}^{t_{n-1}}$. The likelihood function for the estimation of the state X_{t_n} based on the observation z_n alone is defined by

$$\Psi_n(x) = g_n(x, z_n).$$

The *memoryless channel* assumption holds here, under which the observations $\{z_0, \dots, z_n\}$ are independent given the sequence of states $\{X_{t_0}, \dots, X_{t_n}\}$. This assumption holds for example in the case of observations in additive white noise, i.e.

$$z_n = h_n(X_{t_n}) + v_n,$$

where $\{v_n, n \geq 0\}$ is a white noise sequence, independent of the Markov chain. In the special case of a Gaussian white noise sequence with identity covariance matrix, it holds

$$\Psi_n(x) = (2\pi)^{-d/2} \exp \left\{ -\frac{1}{2} |z_n - h_n(x)|^2 \right\}.$$

Let $\mathcal{Z}_n = (z_0, \dots, z_n)$ denote the sequence of observations up to time t_n , and introduce the following conditional probability distributions of the state X_{t_n}

$$\mathbf{P}[X_{t_n} \in dx \mid \mathcal{Z}_{n-1}] = \mu_n^-(dx)$$

and

$$\mathbf{P}[X_{t_n} \in dx \mid \mathcal{Z}_n] = \mu_n(dx).$$

The transition from μ_{n-1} to μ_n , is conveniently described by the following two steps :

$$\begin{aligned} \mu_{n-1} &\xrightarrow{\text{prediction}} \mu_n^- = Q_n^* \mu_{n-1} \\ &\xrightarrow{\text{correction}} \mu_n = \Psi_n \cdot \mu_n^-, \end{aligned}$$

where \cdot denotes the projective product, see (5) below. In the prediction step, $\mu_n^- = Q_n^* \mu_{n-1}$ is the value taken at time t_n by the solution (in weak sense) of the Fokker-Planck equation

$$\frac{\partial \mu_t^n}{\partial t} = L_t^* \mu_t^n, \quad t \geq t_{n-1} \quad (4)$$

starting from μ_{n-1} at time t_{n-1} . In the correction step, μ_n is simply given by the Bayes rule

$$\mu_n = \Psi_n \cdot \mu_n^- = \frac{\Psi_n \mu_n^-}{\langle \mu_n^-, \Psi_n \rangle}. \quad (5)$$

The purpose of this paper is to investigate the Lipschitz property of the two mappings

$$\mu \mapsto Q_n^* \mu \quad \text{and} \quad \mu \mapsto \Psi_n \cdot \mu,$$

for various information theoretic distances on the set of probability measures. The general concept of f -divergence was introduced by Csiszár [10], and has been studied thoroughly by Liese and Vajda [16], Vajda [18]. For any *convex* function f defined on $[0, \infty)$, and such that $f(1) = 0$, the f -divergence between two probability measures μ and ν absolutely continuous w.r.t. μ , is defined by

$$f(\nu \parallel \mu) = \int f\left(\frac{d\nu}{d\mu}\right) d\mu = \langle \mu, f\left(\frac{d\nu}{d\mu}\right) \rangle.$$

Typical choices for the function f are

$$f(r) = |r - 1|, \quad f(r) = (\sqrt{r} - 1)^2, \quad f(r) = -\log r,$$

for which the corresponding f -divergences are the total variation distance, the (square of the) Hellinger distance, and the Kullback-Leibler information distance (or relative entropy) respectively. If both ν and μ are absolutely continuous w.r.t. a measure λ , with density q and p respectively, then the f -divergence takes the form

$$f(\nu \parallel \mu) = \int f\left(\frac{q}{p}\right) p d\lambda.$$

By definition — and in opposition with the distance between the densities p and q in $L^\alpha(\lambda)$ for $\alpha \neq 1$ — the above formula does not depend upon the dominating measure λ . Another interesting feature of f -divergences is their invariance upon change of variable. Indeed, let ϕ be a diffeomorphism of \mathbb{R}^m . If $\nu \ll \mu$, then $\nu \circ \phi^{-1} \ll \mu \circ \phi^{-1}$ with density

$$\frac{d(\nu \circ \phi^{-1})}{d(\mu \circ \phi^{-1})} = \frac{d\nu}{d\mu} \circ \phi^{-1},$$

hence

$$\begin{aligned} f(\nu \circ \phi^{-1} \parallel \mu \circ \phi^{-1}) &= \langle \mu \circ \phi^{-1}, f\left(\frac{d\nu}{d\mu} \circ \phi^{-1}\right) \rangle \\ &= \langle \mu, f\left(\frac{d\nu}{d\mu}\right) \rangle = f(\nu \parallel \mu). \end{aligned}$$

Finally, the following *monotonicity* property — which is well-known in the special case of the total variation distance — has been proved by Csiszár [10], see also Vajda [18, Theorem 9.9]. If $\nu \ll \mu$, then $K^* \nu \ll K^* \mu$ for any Markov kernel K , and

$$f(K^* \nu \parallel K^* \mu) \leq f(\nu \parallel \mu), \quad (6)$$

i.e. f -divergences cannot increase under the action of a Markov kernel.

2 Prediction step

Consider the following two equations

$$\dot{\mu}_t = L_t^* \mu_t \quad \text{and} \quad \dot{\nu}_t = L_t^* \nu_t + \varepsilon_t,$$

where the time-dependent linear partial differential operator L_t is defined in (2), hence

$$\langle u, L_t v \rangle = -\frac{1}{2} \int a_t^{i,j} \frac{\partial u}{\partial x_i} \frac{\partial v}{\partial x_j} dx + \int a_t^i u \frac{\partial v}{\partial x_i} dx,$$

for any test functions u and v , where

$$a_t^i = b_t^i - \frac{1}{2} \frac{\partial a_t^{i,j}}{\partial x_j}.$$

2.1 General estimate

Proposition 2.1 Assume that $\nu_t \ll \mu_t$ for any $t \geq s$. If the convex function f is differentiable, then

$$\frac{d}{dt} f(\nu_t \parallel \mu_t) \leq \langle \varepsilon_t, f'(\frac{d\nu_t}{d\mu_t}) \rangle,$$

for any $t \geq s$.

Remark 2.2 In the nonperturbed case where $\varepsilon_t \equiv 0$, the above result can be extended to any f -divergence, using a limiting argument, and could also be obtained directly, as a consequence of the monotonicity property (6).

Proof: Assume that f has two derivatives, and that μ_t is absolutely continuous w.r.t. the Lebesgue measure, with density p_t . Then the density $r_t = \frac{d\nu_t}{d\mu_t}$ solves

$$\dot{r}_t = \frac{L_t^*(p_t r_t) + \varepsilon_t}{p_t} - r_t \frac{L_t^* p_t}{p_t}.$$

If the function f is smooth enough, then the transformed ratio $u_t = f(r_t)$ solves

$$\dot{u}_t = f'(r_t) \dot{r}_t,$$

hence

$$\begin{aligned} \frac{d}{dt} f(\nu_t \parallel \mu_t) &= \langle p_t, \dot{u}_t \rangle + \langle \dot{p}_t, u_t \rangle \\ &= \langle p_t, f'(r_t) \frac{L_t^*(p_t r_t) + \varepsilon_t}{p_t} \rangle - \langle p_t, f'(r_t) r_t \frac{L_t^* p_t}{p_t} \rangle \\ &\quad + \langle L_t^* p_t, f(r_t) \rangle \\ &= \langle L_t^*(p_t r_t), f'(r_t) \rangle + \langle L_t^* p_t, f(r_t) - r_t f'(r_t) \rangle \\ &\quad + \langle \varepsilon_t, f'(r_t) \rangle. \end{aligned}$$

After straightforward computations

$$\begin{aligned} \langle L_t^*(p r), f'(r) \rangle &= \langle p r, L_t(f'(r)) \rangle \\ &= -\frac{1}{2} \int a_t^{i,j} \left[\frac{\partial p}{\partial x_j} r + p \frac{\partial r}{\partial x_j} \right] f''(r) \frac{\partial r}{\partial x_i} dx \\ &\quad + \int a_t^i p r f''(r) \frac{\partial r}{\partial x_i} dx, \end{aligned}$$

and since $[f(r) - r f'(r)]' = -r f''(r)$

$$\begin{aligned} \langle L_t^* p, f(r) - r f'(r) \rangle &= \langle p, L_t(f(r) - r f'(r)) \rangle \\ &= -\frac{1}{2} \int a_t^{i,j} \frac{\partial p}{\partial x_j} [-r f''(r)] \frac{\partial r}{\partial x_i} dx \\ &\quad + \int a_t^i p [-r f''(r)] \frac{\partial r}{\partial x_i} dx, \end{aligned}$$

hence

$$\begin{aligned} \langle L_t^*(p r), f'(r) \rangle + \langle L_t^* p, f(r) - r f'(r) \rangle \\ = -\frac{1}{2} \int a_t^{i,j} p \frac{\partial r}{\partial x_j} \frac{\partial r}{\partial x_i} f''(r) dx, \end{aligned}$$

for any test functions r and p . Whether μ_t is absolutely continuous or not w.r.t. the Lebesgue measure, it holds

$$\begin{aligned} \frac{d}{dt} f(\nu_t \parallel \mu_t) \\ = -\langle \mu_t, \frac{1}{2} a_t^{i,j} \frac{\partial r_t}{\partial x_j} \frac{\partial r_t}{\partial x_i} f''(r_t) \rangle + \langle \varepsilon_t, f'(r_t) \rangle \\ \leq \langle \varepsilon_t, f'(r_t) \rangle, \end{aligned}$$

since $f''(r_t) \geq 0$. By a limiting argument, the same estimate holds if f has only one derivative. \square

Remark 2.3 If $a_t = (a_t^{i,j}) \equiv 0$, i.e. if the state equation is a deterministic ODE, then equality holds

$$\frac{d}{dt} f(\nu_t \parallel \mu_t) = \langle \varepsilon_t, f'(r_t) \rangle,$$

and in the nonperturbed case where $\varepsilon_t \equiv 0$, the mapping $t \mapsto f(\nu_t \parallel \mu_t)$ is constant. This was expected, by the invariance property of f -divergences upon change of variable, since in this case ν_t and μ_t are the image of ν_s and μ_s respectively, under the solution map of the ODE.

2.2 Application to the projection filter

The projection filter is a finite dimensional nonlinear filter, based on the differential geometric approach to statistics. It is obtained by projecting the Fokker-Planck equation (4) onto the tangent space of a finite dimensional manifold of (square root of) probability densities, according to the Fisher information metric, and its extension to the infinite dimensional space of square roots of probability densities, the Hellinger distance, see Brigo, Hanzon and LeGland [5, 6]. Recall that the (square of the) Hellinger distance is the f -divergence associated with the convex function $f(r) = (\sqrt{r} - 1)^2$, and notice that $f'(r) = \frac{1}{\sqrt{r}} (\sqrt{r} - 1)$.

Let $S = \{p(\cdot, \theta), \theta \in \Theta\}$, where $\Theta \subset \mathbb{R}^n$, be a parametric family of probability densities, and assume that for any $\theta \in \Theta$, the tangent vectors

$$\frac{\partial \sqrt{p(\cdot, \theta)}}{\partial \theta_1}, \dots, \frac{\partial \sqrt{p(\cdot, \theta)}}{\partial \theta_n},$$

are linearly independent vectors in $L^2(\mathbb{R}^m)$. Then $S^{1/2} = \{\sqrt{p(\cdot, \theta)}, \theta \in \Theta\}$ is an n -dimensional submanifold of $L^2(\mathbb{R}^m)$, and the tangent space to $S^{1/2}$ at

$\sqrt{p(\cdot, \theta)}$ is given by

$$T_{\sqrt{p(\cdot, \theta)}} S^{1/2} = \text{span} \left\{ \frac{\partial \sqrt{p(\cdot, \theta)}}{\partial \theta_1}, \dots, \frac{\partial \sqrt{p(\cdot, \theta)}}{\partial \theta_m} \right\}.$$

For any $\theta \in \Theta$, let Π_θ denote the orthogonal projection of vectors of $L^2(\mathbb{R}^m)$ onto the tangent space $T_{\sqrt{p(\cdot, \theta)}} S^{1/2}$. Assume that the density p_t satisfies

$$\dot{p}_t = L_t^* p_t.$$

Then $\sqrt{p_t}$ satisfies

$$\frac{d}{dt} \sqrt{p_t} = \frac{1}{2\sqrt{p_t}} L_t^* p_t = \mathcal{P}_t(\sqrt{p_t}),$$

and the projection filter density $q_t = p(\cdot, \theta_t)$ satisfies, by definition

$$\frac{d}{dt} \sqrt{q_t} = \Pi_{\theta_t} \circ \mathcal{P}_t(\sqrt{q_t}) = \mathcal{P}_t(\sqrt{q_t}) + \mathcal{R}_t(\sqrt{q_t}), \quad (7)$$

where $\mathcal{R}_t = \mathcal{P}_t - \Pi_{\theta_t} \circ \mathcal{P}_t$ is the projection residual operator. This operator, when applied to $\sqrt{q_t}$, yields a vector of $L^2(\mathbb{R}^m)$ called the projection residual. The norm $\|\mathcal{R}_t(\sqrt{q_t})\|$ is called the projection residual norm, and it is an *a posteriori* estimate — in the sense that it can be computed from the approximate solution q_t only, and does not require the exact solution p_t — of the instantaneous error resulting from the projection.

Remark 2.4 The equation (7) looks like a PDE, but it is truly an ODE in the n -dimensional manifold $S^{1/2}$, which can be equivalently written, using different coordinates, as an ODE in $\Theta \subset \mathbb{R}^n$ for the parameter θ_t .

The following result relates the global projection error, for the Hellinger distance, to the instantaneous projection error.

Proposition 2.5 For any $t \geq s$

$$\frac{d}{dt} H(q_t, p_t) \leq \|\mathcal{R}_t(\sqrt{q_t})\|.$$

Proof: The density q_t solves

$$\begin{aligned} \dot{q}_t &= 2\sqrt{q_t} \frac{d}{dt} \sqrt{q_t} \\ &= 2\sqrt{q_t} \mathcal{P}_t(\sqrt{q_t}) + 2\sqrt{q_t} \mathcal{R}_t(\sqrt{q_t}) = L_t^* q_t + \varepsilon_t, \end{aligned}$$

where

$$\varepsilon_t = 2\sqrt{q_t} \mathcal{R}_t(\sqrt{q_t}).$$

Applying Proposition 2.1 yields

$$\begin{aligned} \frac{d}{dt} H^2(q_t, p_t) &\leq \langle \varepsilon_t, f'(\frac{p_t}{q_t}) \rangle \\ &= \langle 2\sqrt{q_t} \mathcal{R}_t(\sqrt{q_t}), \frac{\sqrt{p_t}}{\sqrt{q_t}} (\frac{\sqrt{q_t}}{\sqrt{p_t}} - 1) \rangle \\ &= 2 \langle \mathcal{R}_t(\sqrt{q_t}), (\frac{\sqrt{q_t}}{\sqrt{p_t}} - 1) \sqrt{p_t} \rangle. \end{aligned}$$

The Cauchy–Schwartz inequality yields

$$\begin{aligned} H(q_t, p_t) \frac{d}{dt} H(q_t, p_t) &= \frac{1}{2} \frac{d}{dt} H^2(q_t, p_t) \\ &\leq \int \mathcal{R}_t(\sqrt{q_t}) (\frac{\sqrt{q_t}}{\sqrt{p_t}} - 1) \sqrt{p_t} dx \\ &\leq \left\{ \int |\mathcal{R}_t(\sqrt{q_t})| dx \right\}^{1/2} \left\{ \int (\frac{\sqrt{q_t}}{\sqrt{p_t}} - 1)^2 p_t dx \right\}^{1/2} \\ &= \|\mathcal{R}_t(\sqrt{q_t})\| H(q_t, p_t), \end{aligned}$$

hence

$$\frac{d}{dt} H(q_t, p_t) \leq \|\mathcal{R}_t(\sqrt{q_t})\|. \quad \square$$

3 Correction step (Bayes rule)

Notice that if $\nu \ll \mu$, then $\Psi_n \cdot \nu \ll \Psi_n \cdot \mu$ with density

$$\frac{d(\Psi_n \cdot \nu)}{d(\Psi_n \cdot \mu)} = \frac{d\nu}{d\mu} / \langle \Psi_n \cdot \mu, \frac{d\nu}{d\mu} \rangle. \quad (8)$$

Averaged error estimates will be obtained below, making use of the following elementary result.

$$\mathbf{E}[\Psi_n \cdot \mu_n^- | \mathcal{Z}_{n-1}] = \mu_n^-. \quad (9)$$

Indeed, by definition

$$\langle \Psi_n \cdot \mu_n^-, \phi \rangle = \mathbf{E}[\phi(X_{t_n}) | \mathcal{Z}_n],$$

hence

$$\mathbf{E}[\langle \Psi_n \cdot \mu_n^-, \phi \rangle | \mathcal{Z}_{n-1}] = \mathbf{E}[\phi(X_{t_n}) | \mathcal{Z}_{n-1}] = \langle \mu_n^-, \phi \rangle,$$

for any test function ϕ defined on \mathbb{R}^m .

3.1 Total variation distance

Proposition 3.1 For any probability distributions μ and ν

$$\|\Psi_n \cdot \mu - \Psi_n \cdot \nu\|_{\text{TV}}$$

$$\leq \frac{\sup_{x \in \mathbb{R}^m} \Psi_n(x)}{\max[\langle \mu, \Psi_n \rangle, \langle \nu, \Psi_n \rangle]} \|\mu - \nu\|_{\text{TV}}.$$

Proof: The proof is taken from Hürzeler [13, Lemma 9.5], see also Diaconis and Freedman [12, Appendix B] for a similar argument. For any μ and ν , assuming w.l.o.g. that

$$\langle \mu, \Psi_n \rangle \geq \langle \nu, \Psi_n \rangle,$$

the following decomposition holds

$$\begin{aligned} \Psi_n \cdot \mu - \Psi_n \cdot \nu &= \frac{\Psi_n \mu}{\langle \mu, \Psi_n \rangle} - \frac{\Psi_n \nu}{\langle \nu, \Psi_n \rangle} \\ &= R_n \left[\mu - \frac{\langle \mu, \Psi_n \rangle}{\langle \nu, \Psi_n \rangle} \nu \right] = R_n \left[p - \frac{\langle \mu, \Psi_n \rangle}{\langle \nu, \Psi_n \rangle} q \right] \lambda, \end{aligned}$$

where by definition

$$R_n = \frac{\Psi_n}{\langle \mu, \Psi_n \rangle},$$

and where λ dominates μ and ν , with density

$$p = \frac{d\mu}{d\lambda} \quad \text{and} \quad q = \frac{d\nu}{d\lambda},$$

respectively. Then

$$\begin{aligned} \|\Psi_n \cdot \mu - \Psi_n \cdot \nu\|_{\text{TV}} &= \int R_n \left| p - \frac{\langle \mu, \Psi_n \rangle}{\langle \nu, \Psi_n \rangle} q \right| d\lambda \\ &= 2 \int R_n \left[p - \frac{\langle \mu, \Psi_n \rangle}{\langle \nu, \Psi_n \rangle} q \right]^+ d\lambda \\ &\leq 2 \int R_n [p - q]^+ d\lambda \\ &\leq \sup_{x \in \mathbb{R}^m} R_n(x) \int |p - q| d\lambda \\ &= \frac{\sup_{x \in \mathbb{R}^m} \Psi_n(x)}{\langle \mu, \Psi_n \rangle} \|\mu - \nu\|_{\text{TV}}. \quad \square \end{aligned}$$

3.2 Kullback–Leibler information distance

Proposition 3.2 *If ν is \mathcal{Z}_{n-1} -measurable, and absolutely continuous w.r.t. μ_n^- , then*

$$\mathbb{E}[D(\Psi_n \cdot \nu \parallel \Psi_n \cdot \mu_n^-) \mid \mathcal{Z}_{n-1}] \leq D(\nu \parallel \mu_n^-).$$

Proof: For any μ and ν such that $\nu \ll \mu$, the relation (8) yields

$$\begin{aligned} D(\Psi_n \cdot \nu \parallel \Psi_n \cdot \mu) &= -\langle \Psi_n \cdot \mu, \log \frac{d(\Psi_n \cdot \nu)}{d(\Psi_n \cdot \mu)} \rangle \\ &= -\langle \Psi_n \cdot \mu, \log \frac{d\nu}{d\mu} \rangle + \log \langle \Psi_n \cdot \mu, \frac{d\nu}{d\mu} \rangle. \end{aligned}$$

In the special case where $\mu = \mu_n^-$, taking expectation, and using identity (9), yields

$$\begin{aligned} -\mathbb{E}[\langle \Psi_n \cdot \mu_n^-, \log \frac{d\nu}{d\mu_n^-} \rangle \mid \mathcal{Z}_{n-1}] &= -\langle \mu_n^-, \log \frac{d\nu}{d\mu_n^-} \rangle \\ &= D(\nu \parallel \mu_n^-), \end{aligned}$$

and

$$\begin{aligned} &\mathbb{E}[\log \langle \Psi_n \cdot \mu_n^-, \frac{d\nu}{d\mu_n^-} \rangle \mid \mathcal{Z}_{n-1}] \\ &\leq \log \mathbb{E}[\langle \Psi_n \cdot \mu_n^-, \frac{d\nu}{d\mu_n^-} \rangle \mid \mathcal{Z}_{n-1}] = \log \langle \mu_n^-, \frac{d\nu}{d\mu_n^-} \rangle = 0, \end{aligned}$$

using the Jensen inequality, hence the following averaged estimate holds

$$\mathbb{E}[D(\Psi_n \cdot \nu \parallel \Psi_n \cdot \mu_n^-) \mid \mathcal{Z}_{n-1}] \leq D(\nu \parallel \mu_n^-). \quad \square$$

4 Conclusion : supermartingale property

Let the wrongly initialized filter be defined by the same steps as the correctly initialized filter, i.e.

$$\begin{aligned} \nu_{n-1} &\xrightarrow{\text{prediction}} \nu_n^- = Q_n^* \nu_{n-1} \\ &\xrightarrow{\text{correction}} \nu_n = \Psi_n \cdot \nu_n^-. \end{aligned}$$

Combining the results of the previous two sections, the following supermartingale property can be proved.

Proposition 4.1 *If $\nu_0 \ll \mu_0$, then $\nu_n \ll \mu_n$, and*

$$\mathbb{E}[D(\nu_n \parallel \mu_n) \mid \mathcal{Z}_{n-1}] \leq D(\nu_{n-1} \parallel \mu_{n-1}).$$

Remark 4.2 This property was proved by Clark, Ocone and Coumarbatch [9], for a much more general model.

Proof: Using Proposition 3.2 and the monotonicity property (6), yields

$$\begin{aligned} &\mathbb{E}[D(\nu_n \parallel \mu_n) \mid \mathcal{Z}_{n-1}] \\ &= \mathbb{E}[D(\Psi_n \cdot \nu_n^- \parallel \Psi_n \cdot \mu_n^-) \mid \mathcal{Z}_{n-1}] \leq D(\nu_n^- \parallel \mu_n^-) \\ &= D(Q_n^* \nu_{n-1} \parallel Q_n^* \mu_{n-1}) \leq D(\nu_{n-1} \parallel \mu_{n-1}). \quad \square \end{aligned}$$

Acknowledgment

The author gratefully acknowledges Damiano Brigo, Nadia Oudjane and Boris Rozovskii for interesting discussions and comments about earlier versions of this work.

References

- [1] R. Atar. Exponential stability for nonlinear filtering of diffusion processes in a noncompact domain. *The Annals of Probability*, 26(4):1552–1574, Oct. 1998.

- [2] R. Atar, F. Viens, and O. Zeitouni. Robustness of Zakai's equation via Feynman–Kac representations. In W. McEneaney, G. Yin, and Q. Zhang, editors, *Stochastic Analysis, Control, Optimization and Applications : A Volume in Honor of W.H. Fleming*, Systems & Control : Foundations & Applications, pages 339–352. Birkäuser, Boston, 1998.
- [3] R. Atar and O. Zeitouni. Exponential stability for nonlinear filtering. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 33(6):697–725, 1997.
- [4] R. Atar and O. Zeitouni. Lyapunov exponents for finite state nonlinear filtering. *SIAM Journal on Control and Optimization*, 35(1):36–55, Jan. 1997.
- [5] D. Brigo, B. Hanzon, and F. Le Gland. A differential geometric approach to nonlinear filtering : the projection filter. *IEEE Transactions on Automatic Control*, AC-43(2):247–252, Feb. 1998.
- [6] D. Brigo, B. Hanzon, and F. Le Gland. Approximate filtering by projection on the manifold of exponential densities. *Bernoulli*, 5(3):495–534, June 1999.
- [7] A. S. Budhiraja and H. J. Kushner. Robustness of nonlinear filters over the infinite time interval. *SIAM Journal on Control and Optimization*, 36(5):1618–1637, Sept. 1998.
- [8] A. S. Budhiraja and D. L. Ocone. Exponential stability in discrete-time filtering for non-ergodic signals. *Stochastic Processes and their Applications*, 82(2):245–257, 1999.
- [9] J. M. C. Clark, D. L. Ocone, and C. E. Coumarbatch. Relative entropy and error bounds for filtering of Markov process. (preprint).
- [10] I. Csizár. Eine informationstheoretische ungleichung und ihre anwendung auf den beweis der ergodizität von markoffschen ketten. *Publications of the Mathematical Institute of the Hungarian Academy of Science, Series A*, 8:85–108, 1963.
- [11] P. Del Moral and A. Guionnet. On the stability of measure valued processes. Applications to nonlinear filtering and interacting particle systems. Publication du Laboratoire de Statistique et Probabilités 3–98, Université Paul Sabatier, Toulouse, 1998.
- [12] P. Diaconis and D. Freedman. On the consistency of Bayes estimates (with discussion). *The Annals of Statistics*, 14(1):1–67, Mar. 1986.
- [13] M. Hürzeler. *Statistical Methods for General State-Space Models*. Ph.D. Thesis, Department of Mathematics, ETH Zürich, Zürich, 1998.
- [14] F. Le Gland and L. Mevel. Basic properties of the projective product, with application to products of column-allowable nonnegative matrices. *Mathematics of Control, Signals, and Systems*. (accepted, March 1999).
- [15] F. Le Gland and L. Mevel. Exponential forgetting and geometric ergodicity in hidden Markov models. *Mathematics of Control, Signals, and Systems*. (accepted, March 1999).
- [16] F. Liese and I. Vajda. *Convex Statistical Distances*, volume 95 of *Teubner-Texte zur Mathematik*. Teubner B.G., Leipzig, 1987.
- [17] D. L. Ocone and E. Pardoux. Asymptotic stability of the optimal filter with respect to its initial condition. *SIAM Journal on Control and Optimization*, 34(1):226–243, 1996.
- [18] I. Vajda. *Theory of Statistical Inference and Information*, volume 11 of *Theory and Decision Library B : Mathematical and Statistical Methods*. Kluwer Academic Publishers, Dordrecht, 1989.