August 1986 LIDS-P-1587

IDENTIFICATION AND FILTERING: OPTIMAL RECURSIVE MAXIMUM LIKELIHOOD APPROACH[§]

José M. F. Moura[†]
and
Sanjoy K. Mitter[‡]

Abstract: The paper studies the combined estimation of the parameters and filtering of the state of a stochastic nonlinear dynamical system. It circumvents the two basic limitations found in the litterature on the subject: i) the lack of recursibility of the optimal solution, and ii) the approximations involved when authors discuss recursive algorithms. To derive the *optimal recursive* joint identification algorithm, the problem is formulated in the context of stochastic nonlinear filtering theory. Key to the results is the use of a dynamic programming type approach to the problem, whereby what is important is not the integral form of the likelihood function, but the partial differential equation (pde) it satisfies. The joint likelihood functional of the state and parameters is established as a Feyman-Kac path integral solution to a stochastic pde. The maximum likelihood (ML) estimates are defined as the roots of the likelihood equation, i.e., as the stationarity points of the (negative) log-likelihood functional. By application of Ito's differential rule, the pde satisfied by the latter is derived, and then used to obtain formally recursive equations for the maximum likelihood estimates. These are applied to the important case where the underlying state model is linear. The resulting structure provides a recursive scheme for joint estimation of the state and parameters. In some sense, it is for continuous time the optimal version of the approximate stochastic (discrete time) identification algorithms found in the litterature. In general, the nonlinear structure of the problem precludes a finite dimensional implementation of the optimal estimator. This reflects a tradeoff between recursibility and complexity. Practical implementation requires again approximations, but now directly on the solution, not on the statement of the problem. The optimal algorithm provides a guideline on how to achieve a balance between reduced dimension and acceptable transient behavior.

§ Work supported by the Army Research Office under contract **DAAG-29-84-K-005.**

† Massachusetts Institute of Technology and Laboratory for Information and Decision Systems, on leave from Instituto Superior Técnico, Lisbon, Portugal.

‡ Massachusetts Institute of Technology and Laboratory for Information and Decision Systems.

1 Statement of the Problem

In numerous problems in Stochastic Control and Statistical Communications, the underlying phenomena are modeled as the output of systems driven by external random disturbances. A crucial step in these applications is the adequate identification of a suitable model that replicates the essential features of the phenomena under study. The topic is addressed here from the following point of view. The system structure is known to belong to a given class of models, that of finite dimensional diffusion processes. This knowledge greatly simplifies the identification problem, since it assumes before hand that questions regarding the dimension of the system have been successfully answered, and that the stochastic differential equations describing the processes have been specified up to a finite number of unknown parameters. Within this restricted scope, a very general perspective is taken, namely that the models of interest are nonlinear and stochastic. The framework includes the important problem of identification of parameters in linear systems, where the unknowns may represent noise power levels, pole locations, significant time constants, equivalent bandwidths, correlation times, or other structural parameters defining for example the state variable description of the processes.

In the above context, identification corresponds to the determination of parameters in diffusion processes. The approach to be taken considers it as a stochastic nonlinear filtering problem of specialized structure. As such, the solution requires the *joint* i) filtering of the processes and ii) identification of the unknown parameters. The paper presents the optimal *recursive* algorithm and analyses its structure.

The present work is now compared to the existing results published in the litterature. Two main concerns are distinguished:

i) The first relates to the presentation of the likelihood function for the identification of parameters in processes. Starting with Schweppe [26], many authors, e.g., Balakrishnan [6], Bagchi [5], Borgar and Bagchi [7], Tugnait [30], address the question of formulating the likelihood function. The results describe the likelihood functional in integral form. The maximum likelihood (ML) parameters are then found by maximization of this functional. These approaches are intrinsically *nonrecursive*, lacking expedite methods to update the parameter estimates.

ii) The second concern is along the direction of obtaining recursive identification structures. Here the main thrust constructs *approximate* solutions to the associated nonlinear optimization problem. Depending on the type of simplifications made, different algorithms result, see Sandell and Yared [25], Ljung [18], Caines [10], Astrom and Wittenmark [4], Soderstrom [27],

Young [32], for typical references. Recurrent arguments in these methods use truncated Taylor series approximations to the nonlinear model, or apply stochastic approximation algorithms as introduced in the Statistics litterature by Robbins and Monro [24]. A large effort has been invested in proving asymptotic convergence results for these suboptimal recursions. Practical experience shows that the part of the estimator concerned with the process filtering usually converges at a faster rate than the parameter estimate itself. If one thinks of the parameters as slowly drifting, this behavior is intuitively understood as being a consequence of the two drastically different time constants present in the problem. To improve on the behavior of the recursive algorithms requires an understanding of the associated transients, for which the above approximate techniques provide no clues.

What the work accomplishes, in the context of continuous time diffusions, is the presentation of the *optimal recursive* solution for the (joint) identification of the state and parameters of a possibly nonlinear finite dimensional system. This is done by first deriving a stochastic partial differential equation (pde) for the likelihood function. This equation propagates *recursively* the (joint) ML likelihood function. In fact, working with the (negative) log-likelihood function, the equation that results is of the Hamilton-Jacobi type. Said in other words, the approach obtains the recursive ML-estimates by first imbedding the combined identification problem into a dynamic programming perspective. Following a different approach, the paper extends to the identification question the formal approach of Mortensen [21] to ML-state estimation and is motivated by the suggestion of Mitter [19] of obtaining dynamical equations for the MAP state estimates via a duality between stochastic control and nonlinear filtering. The estimation structure involves two sets of stochastic differential equations, one that filters the state, and the other that estimates the parameters. The proposed algorithm corrects and couples directly *both*. It is a nonlinear stochastic optimal filter, but its structure reflects the specific nature of the underlying model.

The following steps are taken:

i) Determination of the joint likelihood function for both the state (at present time t) and the unknown model parameters (section 3).

ii) Derivation of the stochastic partial differential equation (SPDE) for the likelihood function (section 4).

iii) Application of the so called logarithmic transformation, converting the problem of maximizing the likelihood function into that of minimizing the (negative) log-likelihood function (section 6).

iv) Tracking of the stationarity points of the (negative) log-likelihood function, leading to a set of dynamical equations for the ML-estimates of the state and of the parameters (section 7).

v) Particularization of the general estimator equations to the important problem of a linear state model (section 8). This is made simple, by use of an interpretation of the likelihood-function as a ratio of two densities (section 5).

To carry out the above involves the application of Ito Stochastic Calculus and of Ito differential rule. Suitable references include Bucy and Joseph [9], Lipster and Shiryayev [17], Kunita [16].

For the discrete time case, only the first three steps above can be accomplished exactly. Step iii) still provides a recursive update of the (log)-likelihood function, but step iv) is substituted by a global optimization procedure. Finally, note that the symmetric point of view of simultaneous ML-estimation of the state and parameters is not essential to the results described. The approach can be reformulated in terms of a double criterion where the state estimate minimizes the mean square error, and the parameter estimate minimizes the likelihood function. For the linear state model, this follows immediately, since the two state estimates, the conditional mean and the ML estimate, are related by a suitable normalization.

2 Model: Preliminaries

As mentioned in section 1, the problem concerns the recursive identification of parameters in diffusion processes. The model is now set up. On the complete probability space $(\Omega, \mathcal{F}, P)$, $\{\mathcal{F}_t, t \geq 0\}$ is a (complete) nondecreasing family of right continuous sub-σ-algebras of $\mathcal{F}$. Given two independent Wiener processes $U_t = (u_t,\ \mathcal{F}_t,\ P_t)$ and $W_t = (w_t,\ \mathcal{F}_t,\ P_t)$, $t \geq 0$ with intensities q and r respectively, and an independent $\mathcal{F}_t$-measurable random variable x_0, $P(|x_0| < \infty) = 1$, with probability density function (pdf) $p_0(x, \theta)$, let $x(t)$ and $Y_t = (y_t,\ \mathcal{F}_t,\ P_t)$ be two Ito processes governed by the Ito stochastic differential equations

$$dx_t = f(t, \theta, x_t)dt + g(t, \theta, x_t)du_t, \qquad 0 \leq t \leq T \tag{2-1}$$

$$dy_t = h(t, \theta, x_t)dt + dw_t, \qquad 0 \leq t \leq T. \tag{2-2}$$

Equation (2–1) is labeled the state equation, while (2–2) is the observation equation.

The dependence of the model on some possibly unknown parameter vector θ, taking values in $\Theta \subseteq \Re^m$ is explicitly stated. Most of the time, the processes and the parameter are considered one-dimensional. This does not restrict in a fundamental way the main thrust of the paper. The vector signal and the multidimensional

parameter cases are recovered by a suitable matrix and implied tensor notation reinterpretation of the results. On occasion, a short hand is used for functions of several variables, e.g.,

$$\alpha_t^\theta(x) = \alpha(t, \theta, x). \tag{2-3}$$

When no ambiguity arises, arguments may also be ommitted. In sections 6–8, subscripts will indicate partial derivatives.

If $\widetilde{\Omega}$ is a set of smooth functions, e.g., the set of continuous functions on $[0,T]$

$$\widetilde{\Omega} = C([0,T]\,;\, \Re^1), \tag{2-4}$$

endow it with the topology of uniform convergence on bounded t-intervals. Let $\alpha(t,\widetilde{\omega})$ denote the standard coordinate function

$$\alpha(t,\widetilde{\omega}) = \widetilde{\omega}(t). \tag{2-5}$$

The map $\alpha(t)$ given by $\widetilde{\omega} \mapsto \alpha(t,\widetilde{\omega})$ is continuous, therefore Borel-measurable. It can be shown, see Stroock and Varadhan [28], that the Borel σ-field $\mathcal{F}^\alpha$ is

$$\mathcal{F}^\alpha = \sigma - \{\, \alpha_s : 0 \le s \le T \,\} \quad P - a.s. \tag{2-6}$$

where σ-$\{\cdot\}$ denotes the least σ-algebra generated by the enclosed variables. Even if not stated, all σ-algebras are assumed P-complete. We define

$$\mathcal{F}_t^\alpha = \sigma - \{\, \alpha_s : 0 \le s \le t \,\}. \tag{2-7}$$

The restriction of the underlying measure P to $\mathcal{F}^\alpha$ (or $\mathcal{F}_t^\alpha$) is represented by P^α (or P_t^α). The measure induced by the Wiener process is called the Wiener Measure. Expectation with respect to a measure μ is noted by $E^\mu(\cdot)$. Again, although not directly indicated, expectations are defined uniquely within sets of μ-measure zero. When the measure is clear from the context, the superscript is ommitted. With the above notation, the nondecreasing family of σ-algebras $\mathcal{F}_t$ is usually the cartesian product

$$\mathcal{F}_t = \sigma - \{x_0\} \times \mathcal{F}_t^u \times \mathcal{F}_t^w. \tag{2-8}$$

The model (2–1)–(2–2) is assumed well posed. For all $\theta \in \Theta$, the drift f, the diffusion g, and the sensor h satisfy measurability, smoothness, and Lipschitz growth conditions on t, θ, and x that guarantee the existence, uniqueness, and smoothness of the strong sense solution of (2–1)–(2–2). Sufficient conditions are readily available in standard texts, e.g., Lipster and Shiryayev [17], Stroock and Varadhan [28].

The paper studies the combined maximum-likelihood (ML) estimation of the state and model parameter: Given a record of observations

$$Y_t = (y_s, 0 \le s \le t)$$

construct for each t the ML-estimate of the state $x(t)$ and of the parameter θ. As t ranges from 0 to T, we are further interested in a *recursive* solution. For f, g, and h linear in x, the combined state-parameter estimation is part of the parameter identification in linear systems problem, highly relevant in adaptive control, communication systems, ARMA modeling. Because the parameters appear nonlinearly in the model, it still is a nonlinear problem. As such, and except for trivial situations, the optimal solution is non-feasible. From a practical point of view, its implementation requires some sort of approximation, the optimal algorithm providing a guideline for shortcuts to be taken in actual practical problems. However, the fact that the model is linear in the state, provides the identification problem under consideration with a peculiar structure. This structure is reflected in the solution, and will be noted upon later. It is worth to point out here, that the algorithm to be discussed intrinsically differs from a common strategy found in the litterature, where the parameters are treated as auxiliary state variables with trivial dynamics. Extended Kalman-Bucy filtering arguments are then applied to the extended state vector model. The optimal version of this technique would construct the optimal filter for the extended state vector (nonlinear) model, e.g., the Zakai equation, see Zakai [33]. The difficulty with this approach lies in that the diffusion operator of the extended state vector process is degenerate, the optimal filter becoming numerically highly sensitive.

3 The Likelihood Function

On the measure space $(\Omega, \mathcal{F}, P)$, assume given the family $\mathbf{P} = \{ P_\theta : \theta \in \Theta \}$ of probability measures indexed by the parameter $\theta \in \Theta$. Further, let there exists a (σ-finite) measure λ that dominates every $P_\theta \in \mathbf{P}$. The Radon-Nikodym derivative, see Lipster and Shiryayev [17],

$$L_\omega(\theta) = \frac{dP_\theta(\omega)}{d\lambda(\omega)} \tag{3-1}$$

considered as a function of θ (with ω fixed) is called a likelihood function. The subindex ω will frequently be ommitted. $L_\omega(\theta)$ is unique within stochastic equivalence, i.e., it is defined $\lambda-$a.s.. The statistic that maximizes $L_\omega(\theta)$ is the maximum likelihood estimate.

We apply the above concept to the model (2–1)–(2–2). Loosely speaking, the likelihood function $L(\theta)$ is the Radon-Nikodym derivative of the measure induced by the observation process on the set of continuous functions $\left(C[0,T]; \Re^1 \right)$ with respect to the Wiener measure, as evaluated along the observed record Y_t.

Define

$$Z_t^\theta = \exp\left[\frac{1}{r} \int_0^t h(s,\theta,x_s) dy_s - \frac{1}{2r} \int_0^t h^2(s,\theta,x_s) ds \right]. \tag{3–2.a}$$

Assuming the finite energy condition

$$E^P\left(\int_0^T h^2(s,\theta,x_s)ds\right) < \infty, \tag{3–2.b}$$

the supermartingale Z_t^θ is in fact a martingale (e.g., Theorem 6.1, p. 216, and example 4, p. 221, Lipster and Shiryayev [17]) with

$$EZ_t^\theta = 1. \tag{3–2.c}$$

Then Girsanov's Theorem, see Lipster and Shiryayev [17], applies. Under the measure $\widetilde{P}$ whose Radon-Nikodym derivative

$$\frac{d\widetilde{P}}{dP}(\omega) = \left(Z_T^\theta\right)^{-1} \quad (P - \text{a.s.}) \tag{3–3}$$

i) the process Y_t is a Wiener process; ii) Y_t is independent of X_t; and iii) the probability measure of X_t remains unchanged. Girsanov's theorem implies that given an $\mathcal{F}$-measurable function φ

$$E^P[\varphi] = E^{\widetilde{P}}[\varphi Z_T^\theta]. \tag{3–4}$$

Define the innovations process $V_t = (\nu_t,\ \mathcal{F}_t,\ P_t)$ by

$$d\nu(t) = dy(t) - \widehat{h}_t^\theta(x_t)dt \tag{3–5.a}$$

where

$$\widehat{h}_t^\theta(x_t) = E_\theta^P\left[\, h(t,\theta,x_t) \mid \mathcal{F}_t^y \,\right]. \tag{3–5.b}$$

We assume the innovations property, i.e., that for all t

$$\mathcal{F}_t^\nu = \mathcal{F}_t^y. \tag{3–6}$$

Under the assumed independence of X_t and W_t, Allinger and Mitter [1] have shown that (3–6.a) is true under the general condition (3–2b).

Using (3–3) and the Girsanov's Theorem, the family of measures

$$\exp\left[-\frac{1}{r}\int_0^t \widehat{h}(s,\theta,x_s)d\nu_s - \frac{1}{2r}\int_0^t \widehat{h}^2(s,\theta,x_s)ds\right] P_{\theta,t}^y(d\omega) \tag{3–7}$$

is Wiener measure. A likelihood function for θ is then

$$L_t(\theta) = \exp\left[\frac{1}{r}\int_0^t \widehat{h}(s,\theta,x_s)dy_s - \frac{1}{2r}\int_0^t \widehat{h}^2(s,\theta,x_s)ds\right]. \tag{3–8}$$

Equation (3–8) follows by evaluating the Radon-Nikodym derivative of $P^y_{\theta,t}$ with respect to the Wiener measure (3–7), and upon substitution of the innovations process by its definition (3–5). Application of Girsanov's Theorem is justified, see notes 1 and 3 to Theorem 7-13, Lipster and Shiryayev [17].

It remains to modify (3–8) to include the state $x_t = X$ at present time t as a parameter. In what follows, let

$$\Omega = \Omega^x \times \Omega^w \tag{3–9.a}$$

$$\mathcal{F} = \mathcal{F}^x \times \mathcal{F}^w \tag{3–9.b}$$

$$P = P^x \times P^w \tag{3–9.c}$$

where Ω^x, Ω^w are $\left(C[0,T];\Re^1\right)$, the set of continuous functions defined on $[0,T]$, and all remaining objects have previously been defined. Further, let

$$\mathcal{F}(x_t) = \sigma - \left\{x_s : s \leq t,\ x(t) = X \text{ held fixed}\right\} \tag{3–10.a}$$

$$\mathcal{F}^{yx_t}_s = \mathcal{F}^y_s \vee \mathcal{F}(x_t) \tag{3–10.b}$$

$$P^{yx_t}_s = P \mid \mathcal{F}^{yx_t}_s. \tag{3–10.c}$$

From (3–10.c),

$$P^{yx_t}_s = P^x \times P^w \mid \mathcal{F}^y_s \vee \mathcal{F}(x_t). \tag{3–11}$$

Invoking the independence between W_t and X_t (see begining paragraph of section 2)

$$P^{yx_t}_s = \left(\left[P^x \mid \mathcal{F}(x_t)\right] \times P^w\right) \Big| \mathcal{F}^y_s \tag{3–12}$$

where $P^x \mid \mathcal{F}(x_t)$ is the measure induced by the X_t trajectories terminating at time t at $x_t = X$. Equation (3–12) provides the correct modified measure with respect to which conditional expectations are to be taken, when, besides the record Y_t, the terminal state x_t is also known. But (3–12) shows that the problem is conceptually equivalent to the original filtering problem, except for a change of the a priori measure induced by the (state) process. Also note that

$$E^{P^x \mid \mathcal{F}(x_t)}[\cdot] = E^{P^x}\left[\cdot \mid \mathcal{F}(x_t)\right] \tag{3–13}$$

$$= E^{P^x}\left[\cdot \mid x_t = X\right] \tag{3–14}$$

where (3–14) is a notational convenience.

For an $\mathcal{F}_t^y$-measurable φ, and applying the Girsanov transformation (3–3), one has successively

$$E^{P_t^{yx_t}}[\varphi] = E^P\left[\varphi \mid \mathcal{F}_t^{yx_t}\right] \tag{3–15}$$

$$= E^{\widetilde{P}}\left[\varphi Z_t^\theta \mid \mathcal{F}_t^{yx_t}\right] \tag{3–16}$$

$$= E^{P^x|\mathcal{F}(x_t)}\left[\varphi Z_t^\theta\right] \tag{3–17}$$

$$= E^{P^x}\left[\varphi Z_t^\theta \mid x_t = X\right] \tag{3–18}$$

The step from (3–16) to (3–17) is justified by conditions ii) and iii) of the Girsanov theorem (see under equation (3–3)), i.e., because under $\widetilde{P}$, X_t and Y_t are independent, and $\widetilde{P}$ leaves the law of X_t unchanged.

Let now

$$\widetilde{h}_s^\theta(X) = E^P\left[h(s,\theta,x_s) \mid \mathcal{F}_s^{yx_t},\theta\right] \tag{3–19}$$

which transforms to

$$\widetilde{h}_s^\theta(X) = E^{P^x|\mathcal{F}(x_t)\times P^w}\left[h(s,\theta,x_s) \mid \mathcal{F}_s^y,\theta\right]. \tag{3–20}$$

Under (3–2.b), measure

$$\exp\left[-\frac{1}{r}\int_0^t \widetilde{h}(s,\theta,x_s)d\widetilde{\nu}_s - \frac{1}{2r}\int_0^t \widetilde{h}^2(s,\theta,x_s)ds\right] P_{\theta,t}^{yx_t}(d\omega) = \lambda(dw) \tag{3–21}$$

is Wiener measure. The innovations $\widetilde{V}_t$ are defined as in (3–5) with $\widehat{h}(s,\theta,x_s)$ substituted by $\widetilde{h}(s,\theta,x_s)$. We have proved the following result.

Theorem 3-1. *Under the above setup, the joint likelihood function for the terminal state $x_t = X$ and the parameter θ is*

$$L_t(\theta,X) = \frac{dP_{\theta,t}^{yx_t}}{d\lambda}(\omega)$$

$$= \exp\left[\frac{1}{r}\int_0^t \widetilde{h}(s,\theta,x_s)dy_s - \frac{1}{2r}\int_0^t \widetilde{h}^2(s,\theta,x_s)ds\right]. \tag{3–22}$$

■

Alternatively, (3–22) can be reinterpreted as

$$L_t(\theta,X) = E^{P^x}\left[Z_t^\theta \mid x_t = X,\theta\right] \tag{3–23}$$

where in (3–23), the trajectory Y_t is fixed. Equivalence of (3–22) and (3–23) follows from Theorems 7-1 and 7-13, see also note 1, p. 266, Lipster and Shiryayev [17].

The following remarks are intended to enlighten (3–22) and (3–23):

i) The expectation in (3–23) is over the measure induced by the state trajectories X_t when the terminal state is kept fixed, i.e., $x_t = X$. Of all possible trajectories, one is only considering those terminating at X at time t. The likelihood function is now dependent on X (and on θ), and may be maximized with respect to X (and with respect to θ).

ii) If in (2–1), we set $u_t \equiv 0$, then $(x_s,\ 0 \leq t \leq T)$ is deterministically obtained by integration of (2–1). The set of trajectories terminating at $x_t = X$ collapses in one trajectory (integrate (2–1) backwards in time from $x_t = X$), i.e., $\mathcal{P}^x$ is a delta measure (in function space). Then, (3–23) becomes

$$L_t(\theta, X) = \exp\left[\frac{1}{r}\int_0^t h(s,\theta,x_s(X))dy_s - \frac{1}{2r}\int_0^t h^2(s,\theta,x_s(X))ds\right] \quad (3\text{–}24)$$

where $x_s(X)$ is notation for the solution at time s of (2–1) integrated backwards from $x_t = X$. No expectation is involved in (3–24). Equation (3–24) is nothing but the classical likelihood function of nonrandom waveform estimation when the measurement noise is white, see Van Trees [31], chapter V.

4 Stochastic Partial Differential Equation for the Likelihood Function

To obtain filtering structures updating recursively the ML-estimates of $x_t = X$ and θ, a stochastic partial differential equation (SPDE) for $L_t(\theta, X)$ is needed. To accomplish this, the likelihood function (3–23) is interpreted as the solution in a probabilistic sense of a Cauchy problem associated with a forward parabolic differential equation (FPDE) (Feyman-Kac's formula). The corresponding differential operator is obtained by writing a backward Ito stochastic differential equation (SDE) for the forward Ito SDE (2–1). For details on the probabilistic interpretation of the solution of Cauchy problems associated with forward and backward PDEs, and on the connections between these and forward and backward Ito SDEs, see Kunita [16].

By Feyman-Kac's formula, equation (3–23) is the solution of the forward PDE

$$\begin{cases} dv_t^\theta(X) = \widetilde{\mathcal{L}}_t^\theta v_t^\theta(X)dt + \dfrac{1}{r}h(t,\theta,X)v_t^\theta(X)dy_t \\ \lim\limits_{t\to 0} v_t^\theta(X) = 1. \end{cases} \quad (4\text{–}1)$$

where $\widetilde{\mathcal{L}}_t^\theta$ is a second order elliptic operator. It is the infinitesimal generator of the Ito backward SDE associated with the process (2–1) given by

$$\widetilde{\mathcal{L}}_t^\theta[\cdot] = \widetilde{f}(t,\theta,X)\frac{\partial}{\partial X}[\cdot] + \frac{1}{2}g(t,\theta,X)^2\frac{\partial^2}{\partial X^2}[\cdot] \tag{4–2}$$

where the modified (backward) drift

$$\widetilde{f}(t,\theta,X) = \left(p(t,X)^+\right)\frac{\partial}{\partial X}\left[g(t,\theta,X)^2 p(t,X)\right] - f(t,\theta,X). \tag{4–3}$$

In (4–3), f and g are the drift and diffusion coefficients of (2–1), $p(t,X)$ is the probability density function of P^x (the prior measure of the process) assumed to exist, and

$$p(t,X)^+ = \begin{cases} 0 & \text{if} \quad p(t,X) = 0 \\ \dfrac{1}{p(t,X)} & \text{otherwise.} \end{cases} \tag{4–4}$$

The drift of the backward process as given by (4–3) was obtained in Anderson[2], and used in Anderson and Rhodes [3] to derive smoothing formulae for nonlinear diffusion models. Conditions exist for the time reversed process to be again a diffusion, see Anderson[2], Elliot [12], Haussman and Pardoux [15], Follmer [14]. By the support Theorem in Stroock [29] (Theorem 4.20 for the degenerate case, or Corollary 3.9 for the nondegenerate case), $p(t,X) > 0, \forall X$, so that (4–4) is irrelevant under the present conditions.

5 The Likelihood Function as a Ratio of Two Densities

In section 8, it will be useful to have the interpretation of the (conditional) likelihood function $L(t,\theta,X)$ as a ratio of two densities. Let $p(t,X)$ be the probability density function of the process measure P_t^x, satisfying the Forward Kolmogorof equation

$$dp_t^\theta(X) = \mathcal{L}_t^{\theta *}\left(p_t^\theta(X)\right)dt, \tag{5–1}$$

where

$$\mathcal{L}_t^{\theta *}(\cdot) = -\frac{\partial}{\partial X}\left(f(t,\theta,X)\,\cdot\right) + \frac{1}{2}\frac{\partial^2}{\partial X^2}\left(g(t,\theta,X)^2\,\cdot\right) \tag{5–2}$$

is the adjoint of the operator $\mathcal{L}_t^\theta$ associated with the original diffusion process (2–1). Also, let $q_t^\theta(x)$ satisfy

$$dq_t^\theta(x) = \mathcal{L}_t^{\theta *}\left(q_t^\theta(x)\right)dt + \frac{1}{r}h(t,\theta,X)q_t^\theta(x)dy_t. \tag{5–3}$$

Equation (5–3) is known in the filtering litterature as the Zakai equation, see Zakai [33], also Bucy [8]. General conditions for (5–3) to have a solution are given for example in Kunita [16]. When it exists, the solution of (5–3) is known as the unnormalized probability density function.

Result 5-1. *Assume (4–1), (5–1), and (5–3), have unique solutions. Then*

$$v_t^\theta(X) = q_t^\theta(x) p_t^\theta(X)^+ \tag{5–4}$$

where $p_t^\theta(X)^+$ *is defined in (4–4).*

■

Proof: Application of Ito's formula to (5–4) leads to

$$dv_t^\theta(X) = p_t^\theta(X)^+ dq_t^\theta(x) - q_t^\theta(x)(p_t^\theta(X)^+)^2 dp_t^\theta(X). \tag{5–5}$$

Evaluating the right hand side of (4–1), after straightforward manipulation, obtain

$$\widetilde{\mathcal{L}}_t^\theta \left(q_t^\theta(x) p_t^\theta(X)^+\right) = p_t^\theta(X)^+ \mathcal{L}_t^{\theta *} \left(q_t^\theta(x)\right) - q_t^\theta(x)(p_t^\theta(X)^+)^2 \mathcal{L}_t^{\theta *} \left(p_t^\theta(X)\right) \tag{5–6}$$

where $\mathcal{L}_t^{\theta *}$ is given by (5–2). Substitution of (5–5) and (5–6) in (4–1)

$$\begin{aligned} p_t^\theta(X)^+ dq_t^\theta(x) - q_t^\theta(x)(p_t^\theta(X)^+)^2 dp_t^\theta(X) &= p_t^\theta(X)^+ \mathcal{L}_t^{\theta *} \left(q_t^\theta(x)\right) dt \\ - q_t^\theta(x)(p_t^\theta(X)^+)^2 \mathcal{L}_t^{\theta *} \left(p_t^\theta(X)\right) dt &+ \left(p_t^\theta(X)^+\right) \frac{1}{r} h(t,\theta,X) q_t^\theta(x) dy_t. \end{aligned}$$

Finally, (5–1) and (5–3) show the desired result.

■

Equation (5–4) factors the likelihood function as the product of the unnormalized probability density function by the inverse of the prior density. Alternatively,

$$q_t^\theta(x) = v_t^\theta(X) p_t^\theta(X)$$

corresponds to the numerator of the Representation Theorem of Bucy, see Bucy and Joseph [9], Theorem 4.1. See Pardoux [23] for a similar discussion. In a sense, factoring (5–4) is not surprising. The parameter θ having no probabilistic structure, the likelihood function for the terminal state $x_t = X$, or the joint likelihood function for the terminal state $x_t = X$ *and* the parameter θ, lead both to the same object.

6 The Negative Log-Likelihood Function

Due to the monotonic behavior of the logarithmic function, it is a standard procedure to substitute the maximization of the likelihood function by the minimization of its negative logarithm. In the context of stochastic filtering and control, the logarithmic transformation has been introduced by Flemming and Mitter [13],

see also Mitter [19], to establish a certain duality between problems in the two fields. Consider

$$L(t,\theta,X) = \exp\left[-S(t,\theta,X)\right] \tag{6-1}$$

or

$$S(t,\theta,X) = -\ln L(t,\theta,X) \tag{6-2}$$

Remark: By (3–2.b), $S(t,\theta,X)$ is well defined.

Notation: In the rest of the paper, subscripts x and θ stand for differentiation with respect to the subscripted variable.

Theorem 6-1. *The (negative) log-likelihood function satisfies*

$$dS = \widetilde{\mathcal{L}}_t^\theta(S)\,dt - \frac{1}{2}g(t,\theta,X)^2 S_x^2 dt + \frac{1}{2r}h(t,\theta,X)^2 dt - \frac{1}{r}h(t,\theta,X)dy(t), \tag{6-3}$$

where the operator $\widetilde{\mathcal{L}}_t^\theta$ has been defined in (4–2).

■

Proof: Application of Ito stochastic differential rule to (6–2) leads directly to (6–3).

■

To facilitate the interpretations later on, (6–3) is explicitly evaluated

$$dS = \sigma\left(S_{xx} - S_x^2\right)dt + \alpha S_x dt + \frac{1}{2r}h(t,\theta,X)^2 dt - \frac{1}{r}h(t,\theta,X)dy(t) \tag{6-4}$$

where

$$\sigma = \frac{1}{2}g(t,\theta,X)^2 \tag{6-5}$$

$$\alpha = 2\sigma_x - 2\sigma\overline{S}_x - f(t,\theta,X) \tag{6-6}$$

$$\overline{S} = -\ln p_t^\theta(X) \tag{6-7}$$

with $p_t^\theta(X)$ being the a priori probability density function of the process which is supposed to exist. For conditions on the coefficients of equations of the type of (6–4) to have a solution which has continuous time derivatives and continuous first and second order spatial derivatives see Flemming and Mitter [13], Theorem 4.1 and section 6.

7 Filtering and Identification: Joint ML-Estimation Equations

The joint maximum likelihood (ML) estimates of θ and $x_t = X$ are

$$\begin{bmatrix} \widehat{x}(t) \\ \widehat{\theta}(t) \end{bmatrix} = \arg \min_{\theta \in \Theta, X \in \Re^1} S(t, \theta, X). \tag{7-1}$$

Alternatively, and in the sequel, the ML-estimates are defined implicitly as the roots of the log-likelihood equation

$$\nabla S(t, \theta, X) = 0 \tag{7-2}$$

where

$$\nabla = \begin{bmatrix} \dfrac{\partial}{\partial X} & \dfrac{\partial}{\partial \theta} \end{bmatrix}^{\mathrm{T}}$$

is the gradient operator. Explicitly, (7–2) is

$$\begin{aligned} \widehat{S}_x(t) &= 0 \\ \widehat{S}_\theta(t) &= 0. \end{aligned} \tag{7-3}$$

The hat notation represents evaluation along the ML-trajectory, while subscripts stand for partial derivatives. Let $\nabla^2 S$ be the Hessian matrix

$$\nabla^2 S = \begin{bmatrix} S_{xx} & S_{x\theta} \\ S_{x\theta} & S_{\theta\theta} \end{bmatrix}. \tag{7-4}$$

Result 7-1. *Assume that for all $t \geq 0$, $\theta \in \Theta$, and $x \in \Re^1$:*

i) $S(t, \theta, X)$ is measurable and smooth up to fourth order partial derivatives;

ii) The Hessian $\nabla^2 S$ exists and is invertible;

iii) For each $t \geq 0$, there exists a countable number of solutions to (7–2).

Then, the joint ML state and parameter estimates satisfy:

$$\begin{bmatrix} d\widehat{x}(t) \\ d\widehat{\theta}(t) \end{bmatrix} = \widehat{A}(t)dt + \widehat{\kappa}(t)d\eta(t) \tag{7-5}$$

with the (pseudo) innovations

$$d\eta(t) = dy(t) - \widehat{h}(t)dt, \tag{7-6}$$

the filter gain

$$\widehat{\kappa}(t) = \begin{bmatrix} \widehat{\kappa}_1(t) \\ \widehat{\kappa}_2(t) \end{bmatrix} \tag{7-7}$$

$$= \frac{1}{r} \frac{1}{\widehat{S}_{xx}\widehat{S}_{\theta\theta} - \widehat{S}_{x\theta}^2} \begin{bmatrix} \widehat{h}_x(t)\widehat{S}_{\theta\theta} - \widehat{h}_\theta(t)\widehat{S}_{x\theta} \\ -\widehat{h}_x(t)\widehat{S}_{x\theta} + \widehat{h}_\theta(t)\widehat{S}_{xx} \end{bmatrix},$$

and the drift

$$\widehat{\mathcal{A}}(t) = -\begin{bmatrix} \widehat{\alpha} \\ 0 \end{bmatrix} \tag{7-8}$$

$$+\frac{1}{\widehat{S}_{xx}\widehat{S}_{\theta\theta} - \widehat{S}_{x\theta}^2} \begin{bmatrix} \widehat{\sigma}\left(\widehat{S}_{\theta\theta}\widehat{S}_{xxx} - \widehat{S}_{x\theta}\widehat{S}_{xx\theta}\right) + \left(\widehat{\sigma}_x\widehat{S}_{\theta\theta} - \widehat{\sigma}_\theta\widehat{S}_{x\theta}\right)\widehat{S}_{xx} \\ \left(-\widehat{\sigma}_x\widehat{S}_{x\theta} + \widehat{\sigma}_\theta\widehat{S}_{xx}\right)\widehat{S}_{xx} + \widehat{\sigma}\left(-\widehat{S}_{xxx}\widehat{S}_{x\theta} + \widehat{S}_{xx}\widehat{S}_{xx\theta}\right) \end{bmatrix}$$

$$+\widehat{\varsigma}(t),$$

where the term corresponding to the Ito correction

$$\widehat{\varsigma}(t)dt = -\frac{1}{2}\left(\widehat{\nabla^2 S}\right)^{-1} \begin{bmatrix} \widehat{S}_{xxx} & 2\widehat{S}_{xx\theta} & \widehat{S}_{x\theta\theta} \\ \widehat{S}_{\theta xx} & 2\widehat{S}_{\theta\theta x} & \widehat{S}_{\theta\theta\theta} \end{bmatrix} \begin{bmatrix} \widehat{\kappa}_1(t)^2 \\ \widehat{\kappa}_1(t)\widehat{\kappa}_2(t) \\ \widehat{\kappa}_2(t)^2 \end{bmatrix} dt. \tag{7-9}$$

■

Proof: This is only a formal argument. A rigorous analysis requires that conditions for smoothness of the required derivatives of S be given. That is not trivial in the general nonlinear case. In the linear stationary problem of section 8, however, these conditions translate into smoothness of the solution of the algebraic Riccati equation for which there are in fact results, see Delchamps [11]. By the measurability and smoothness of S on t, θ, and X, and by completeness of the σ-fields, it is possible to choose $\mathcal{F}_t^y$-(progressively) measurable versions for $\widehat{S}(t)$, $\widehat{S}_x(t)$, $\widehat{S}_\theta(t)$, and for the higher order derivatives, see Lemma 4.7 in Lipster and Shiryayev [17]. To track the time variations of the stationarity points of $S(t,\theta,X)$, the total time derivative of (7-3)

$$d\begin{bmatrix} \widehat{S}_x(t) \\ \widehat{S}_\theta(t) \end{bmatrix} = 0 \tag{7-10}$$

is evaluated by Ito's rule. In symbolic notation,

$$\partial_t \widehat{\nabla S} + \widehat{\nabla^2 S}\begin{bmatrix} d\widehat{x}(t) \\ d\widehat{\theta}(t) \end{bmatrix} + \widetilde{\varsigma}(t)dt = 0, \tag{7-11}$$

where

$$\widehat{\nabla^2 S} = \begin{bmatrix} \widehat{S}_{xx} & \widehat{S}_{x\theta} \\ \widehat{S}_{x\theta} & \widehat{S}_{\theta\theta} \end{bmatrix} \tag{7-12}$$

is the Hessian evaluated along the ML-trajectory. The last term $\widetilde{\zeta}(t)$ represents the Ito correction

$$\widetilde{\zeta}(t)dt = \frac{1}{2}\widehat{\nabla^3 S}\, dm(t) \tag{7–13}$$

with

$$\widehat{\nabla^3 S} = \begin{bmatrix} \widehat{S}_{xxx} & 2\widehat{S}_{xx\theta} & \widehat{S}_{x\theta\theta} \\ \widehat{S}_{\theta xx} & 2\widehat{S}_{\theta\theta x} & \widehat{S}_{\theta\theta\theta} \end{bmatrix}, \tag{7–14}$$

the matrix of third order derivatives, and

$$dm(t) = \begin{bmatrix} \langle \widehat{x}, \widehat{x} \rangle_t \\ \langle \widehat{x}, \widehat{\theta} \rangle_t \\ \langle \widehat{\theta}, \widehat{\theta} \rangle_t \end{bmatrix} dt \tag{7–15}$$

the vector of quadratic variations. To evaluate the first term of (7–11), following Mitter [20], interchange orders of differentiation

$$\partial_t \widehat{\nabla S} = \widehat{\nabla dS} \tag{7–16}$$

and use (6–4) to compute dS. Assuming the Hessian is invertible, obtain from (7–11)

$$\begin{bmatrix} d\widehat{x}(t) \\ d\widehat{\theta}(t) \end{bmatrix} = -\left(\widehat{\nabla^2 S}\right)^{-1} \widehat{\nabla dS} + \widehat{\zeta}(t)dt \tag{7–17}$$

with

$$\widehat{\zeta}(t) = -\left(\widehat{\nabla^2 S}\right)^{-1} \widetilde{\zeta}(t). \tag{7–18}$$

After algebraic manipulations, and using the stationarity condition (7–3), equation (7–17) leads to (7–5) – (7–9).

■

The following comments apply:

i) From (7–8), the ML-filter drift is $-[\,\widehat{\alpha}\ 0\,]^{\mathrm{T}}$ corrected by terms involving higher order derivatives. This is the model following structure of the filter.

ii) Under general conditions, maximum likelihood commutes with nonlinear no memory operations. So,

$$\widehat{h}(t) = h(t, \widehat{\theta}_t, \widehat{x}_t),$$

and similarly for $\widehat{\sigma}$, $\widehat{\sigma}_x$, $\cdots$. For memory functions, care should be taken. In general, expressions like $S_{xx}(t)$ make no sense in the realm of Ito calculus. They would involve notions of stochastic integration with nonantecipative

integrands. As remarked before, this does not however preclude that an $\mathcal{F}_t^y$-measurable version may be chosen for $\widehat{S}_{xx}(t)$ (and for the other higher order derivatives), so that (7–5) is well defined in the sense of Ito.

iii) In face of the previous comment, it is clear that the filter (7–5) is not in closed form, knowledge of the higher order derivatives evaluated along the ML-trajectory being required. Dynamical equations for these may be obtained in much the same way (7–5) was derived. As a further example, the equation for $\widehat{S}_{xx}(t)$ is

$$\begin{aligned} d\widehat{S}_{xx}(t) &= \widehat{S}_{xx\theta}d\widehat{\theta}(t) + \widehat{S}_{xxx}d\widehat{x}(t) + (\widehat{\sigma}_{xx} + 2\widehat{\alpha}_x)\,\widehat{S}_{xx}dt \\ &\quad + (2\widehat{\sigma}_x + \widehat{\alpha})\,\widehat{S}_{xxx}dt + \widehat{\sigma}\widehat{S}_{xxxx}dt - 2\widehat{\sigma}\widehat{S}_{xx}^2dt \\ &\quad + \frac{1}{r}\widehat{h}_x(t)^2dt - \frac{1}{r}\widehat{h}_{xx}(t)d\eta(t) + \text{Ito correction.} \end{aligned} \tag{7–19}$$

In turn, this equation requires knowledge of measurable versions of third and fourth order derivatives of $\widehat{S}$. Well known in nonlinear filtering, this coupling to higher order moments shows that recursibility trades off with dimensionality in ML-estimation.

iv) As noted, the process $(\eta(t),\ t \geq 0)$ is a pseudo innovation. In its definition, $\widehat{h}(t)$ is not the conditional expectation of the sensor but the sensor evaluated along the ML-trajectory. Its bounded variation component also contributes to the drift in (7–5).

v) The structure of (7–5) resembles that of the (optimal) conditional mean filter, except that conditional expectation is herein substituted by ML-estimation. One distinguishes a drift term and an "innovations" weighted by a gain.

vi) Equation (7–5) has assumed the invertibility of the Hessian matrix along the ML-trajectory. If this is not true, (7–11) might then be solved by a least squares technique. Next section sheds light on the meaning of the Hessian matrix. In the linear state problem, it corresponds to the inverse of the state error covariance matrix computed by the Riccati equation associated with the Kalman-Bucy filter.

vii) Filter (7–5) addresses the combined state and parameter ML-estimation. If the parameters are known a priori, the filter structure is correspondingly simplified. The resulting filter corrects (with the Ito correction term) the one in Mortensen [21], who was the first to address formally the question of nonlinear ML state estimation. If there is no plant noise, the state is

a deterministic signal, (7–5) describes the recursive observer/identification algorithm. Finally, if, except for the unknown parameters, the state is a known time function, (7–5) provides the recursive parameter estimator equations.

8 Identification of Parameters in Linear Systems

Although in general, equation (6–4) has no closed form solution, there is an important class of problems for which one is readily available. This corresponds to the special case where in the model (2–1)-(2–2)

$$\begin{aligned} f(t,\theta,x_t) &= F(t,\theta)x(t) \\ g(t,\theta,x_t) &= G(t,\theta) \\ h(t,\theta,x_t) &= H(t,\theta)x(t). \end{aligned}$$

Substitution in (2–1) and (2–2) leads to the linear (in state) model

$$dx(t) = F(t,\theta)x(t)dt + G(t,\theta)du(t), \qquad x_0 \tag{8–1}$$

$$dy(t) = H(t,\theta)x(t)dt + dw(t), \qquad y(0) = 0. \tag{8–2}$$

The initial condition is also assumed to be

$$x_0 \sim \text{Normal}\left(\overline{x}_0, \overline{P}_0\right). \tag{8–3}$$

The mean and covariance of x_0 may be θ dependent. All remaining hypotheses underlying (2–1) and (2–2) stay in force. For each value of $\theta \in \Theta$, the processes X_t and Y_t described by (8–1) and (8–2) are Gaussian. The prior mean $\overline{x}(t)$ and prior covariance $\overline{P}(t,\theta)$ of $x(t)$ satisfy

$$\dot{\overline{x}}(t,\theta) = F(t,\theta)\overline{x}(t), \qquad \overline{x}(0,\theta) = \overline{x}_0, \tag{8–4}$$

and the Lyapounov equation

$$\dot{\overline{P}}(t,\theta) = F(t,\theta)\overline{P}(t,\theta) + \overline{P}(t,\theta)F(t,\theta)^{\mathrm{T}} + G(t,\theta)G(t,\theta)^{\mathrm{T}}, \quad \overline{P}(0,\theta) = \overline{P}_0. \tag{8–5}$$

The conditional density of $x(t)$ given $\mathcal{F}_t^y$ is also Gaussian. Its moments, the mean

$$\mu(t,\theta) = E\left[x(t)|\mathcal{F}_t^y\right] \tag{8–6}$$

and the covariance

$$P(t,\theta) = \text{Cov}\left[x(t) - \mu(t,\theta)|\mathcal{F}_t^y\right] \tag{8–7}$$

are updated by the Kalman-Bucy filter

$$d\mu(t,\theta) = F(t,\theta)\mu(t,\theta)dt + P(t,\theta)H(t,\theta)r^{-1}\left[\, dy(t) - H(t,\theta)\mu(t,\theta)dt \,\right] \tag{8-8}$$

$$\begin{aligned}\dot{P}(t,\theta) &= F(t,\theta)P(t,\theta) + P(t,\theta)F(t,\theta)^{\mathrm{T}} + G(t,\theta)G(t,\theta)^{\mathrm{T}} \\ &\quad - P(t,\theta)H(t,\theta)^{\mathrm{T}}R^{-1}H(t,\theta)P(t,\theta).\end{aligned} \tag{8-9}$$

The initial condition for the above recursion is

$$\begin{aligned}\mu(0,\theta) &= \overline{x}_0 \\ P(0,\theta) &= \overline{P}(0,\theta).\end{aligned} \tag{8-10}$$

With the insight provided in section 5 that interprets the likelihood function as the ratio of the unnormalized conditional probability density function over the prior probability density function, it is easy to establish the closed form expression for S

$$\begin{aligned}2S(t,\theta,X) &= \left[\, X - \mu(t,\theta) \,\right]^{\mathrm{T}} \left[\, P(t,\theta) \,\right]^{-1} \left[\, X - \mu(t,\theta) \,\right] \\ &\quad + \frac{1}{r}\int_0^t \left|\, H(t,\theta)\mu(t,\theta) \,\right|^2 ds - \frac{2}{r}\int_0^t H(t,\theta)\mu(t,\theta)\, dy(t) \\ &\quad - \left[\, X - \overline{x}(t) \,\right]^{\mathrm{T}} \left[\, \overline{P}(t,\theta) \,\right]^{-1} \left[\, X - \overline{x}(t) \,\right] + \ln\left[\, \left|\, P(t,\theta) \,\right| / \left|\, \overline{P}(t,\theta) \,\right| \,\right]\end{aligned} \tag{8-11}$$

as a solution to (6–4). In (8–11), $|\cdot|$ is the determinant function. Verification of (8–11) follows if it is substituted in (6–4), all derivatives are carried out (taking care of Ito corrections), and terms in x^2, x^1, and x^0 are collected.

To specialize the ML estimator equation (7–5) to the present problem, notice that from (8–11), S is quadratic in X. So

$$\widehat{S}_{xxx} = 0. \tag{8-12}$$

Also, because G is independent of X

$$\widehat{\sigma}_x = \frac{1}{2}(GG^T)_x = 0. \tag{8-13}$$

Obtain

$$\begin{aligned}\begin{bmatrix} d\widehat{x}(t) \\ d\widehat{\theta}(t) \end{bmatrix} &= \begin{bmatrix} \left[F(t,\widehat{\theta}(t)) + G(t,\widehat{\theta}(t))^2 \widehat{\overline{P}}(t)^{-1} \right] \\ 0 \end{bmatrix} \widehat{x}(t)dt + \widehat{B}(t)dt \\ &\quad + \widehat{\varsigma}(t)dt + \widehat{\kappa}(t)d\widehat{I}_t.\end{aligned} \tag{8-14}$$

where $\widehat{\varsigma}(t)$ is as in (7–9)

$$\widehat{B}(t) = \frac{(\sigma\widehat{S_{xx}})_\theta}{\widehat{S}_{xx}\widehat{S}_{\theta\theta} - \widehat{S}^2_{x\theta}} \begin{bmatrix} -\widehat{S}_{x\theta} \\ \widehat{S}_{xx} \end{bmatrix} \tag{8–15}$$

and the filter gain

$$\widehat{\kappa}(t) = \frac{1}{r}\frac{1}{\widehat{S}_{xx}\widehat{S}_{\theta\theta} - \widehat{S}^2_{x\theta}} \begin{bmatrix} H(t,\widehat{\theta}(t))\,\widehat{S}_{\theta\theta} - H_\theta(t,\widehat{\theta}(t))\widehat{x}(t)\,\widehat{S}_{x\theta} \\ -H(t,\widehat{\theta}(t))\,\widehat{S}_{x\theta} + H_\theta(t,\widehat{\theta}(t))\widehat{x}(t)\,\widehat{S}_{xx} \end{bmatrix} \tag{8–16}$$

Rearrange the filter gain as

$$\widehat{\kappa}(t) = \widehat{\kappa}^a(t) + \widehat{\kappa}^b(t) \tag{8–17.a}$$

where

$$\widehat{\kappa}^a(t) = \frac{1}{r}\frac{1}{\widehat{S}_{xx}\widehat{S}_{\theta\theta} - \widehat{S}^2_{x\theta}} \begin{bmatrix} H(t,\widehat{\theta}(t))\widehat{S}_{\theta\theta} \\ H_\theta(t,\widehat{\theta}(t))\widehat{x}(t)\widehat{S}_{xx} \end{bmatrix} \tag{8–17.b}$$

and

$$\widehat{\kappa}^b(t) = -\frac{1}{r}\frac{1}{\widehat{S}_{xx}\widehat{S}_{\theta\theta} - \widehat{S}^2_{x\theta}} \begin{bmatrix} H_\theta(t,\widehat{\theta}(t))\widehat{x}(t) \\ H(t,\widehat{\theta}(t)) \end{bmatrix} \widehat{S}_{x\theta} \tag{8–17.c}$$

Appendix A computes all the required partial derivatives $\widehat{S}_{xx}, \cdots$, as well as the determinant of the Hessian $|\ \widehat{\nabla^2 S}(t)\ |$. The expressions in the Appendix involve knowledge of the prior moments $\overline{x}(t)$ and $\overline{P}(t,\theta)$, and of the conditional moments $\mu(t,\theta)$ and $P(t,\theta)$, as well as higher order partial derivatives, along the ML-trajectory. In general, these quantities are not available in closed form. Instead, they are themselves described by stochastic differential equations, which are in turn obtained by Ito's rule. For example, the equation for $\widehat{\overline{x}}(t)$ is

$$d\widehat{\overline{x}}(t) = \widehat{\partial_t \overline{x}(t)} + \widehat{\overline{x}_\theta}(t)d\widehat{\theta}(t) + \frac{1}{2}\widehat{\overline{x}_{\theta\theta}}(t)\widehat{k}_2(t)^2 dt \tag{8–18}$$

where $\widehat{k}_2(t)$ is the second component of the gain vector, see (8–16) or (8–17). The first term of the right hand side of (8–18) is given by (8–4) evaluated along the ML estimates. Likewise,

$$d\widehat{\mu}(t) = \widehat{\partial_t \mu(t)} + \widehat{\mu_\theta}(t)d\widehat{\theta}(t) + \frac{1}{2}\widehat{\mu_{\theta\theta}}(t)\widehat{k}_2(t)^2 dt \tag{8–19}$$

where, from the Kalman-Bucy filter (8–8),

$$\begin{aligned} \widehat{\partial_t \mu(t)} &= F(t,\widehat{\theta}(t))\widehat{\mu}(t)dt \\ &\quad + \widehat{P}(t)H(t,\widehat{\theta}(t))\frac{1}{r}\big[dy(t) - H(t,\widehat{\theta}(t))\widehat{\mu}(t)dt\big]. \end{aligned} \tag{8–20}$$

Again, we observe that the above equations are not in closed form, higher order partial derivatives of the moments being required.

9 Conclusion

The present work has presented the recursive equations satisfied by the joint maximum likelihood estimates of the state and of the parameters of a dynamical system. This has been accomplished in steps. First, by deriving the joint likelihood function. Secondly, by finding a recursive (stochastic partial differential equation) description for the likelihood function. Thirdly, by writing the stochastic partial differential equation for the log-likelihood function through application of the Ito stochastic differential rule to the logarithmic transformation. Finally, obtaining the desired recursive algorithm through minimization of the log-likelihood function. After discussing the structure of the estimator, section 8 of the paper studied the special problem when the underlying model is linear in the state. Examples of application of the present work and the question of implementation of the schemes proposed are considered in Moura, Mitter, and Ljung [22]. Consistency, for which there are presently partial answers, is studied elsewhere.

10 Bibliography

[1] F. Allinger and S. K. Mitter, "New Results on the Innovation Problem for Nonlinear Filtering," *Stochastics* **4** (1981), 339–348.

[2] B. D. O. Anderson, "Reverse Time Diffusion Equation Models," *Stochastic Processes and Their Applications* **12** (1982), 313–326.

[3] B. D. O. Anderson and I. B. Rhodes, "Smoothing Algorithms for Nonlinear Finite-Dimensional Systems," *Stochastics* **9** (1983), 139–165.

[4] K. J. Astrom and B. Wittenmark, "Problems of Identification and Control," *Journal of Mathematical Analysis and Applications* **34** (1971), 90–113.

[5] A. Bagchi, "Consistent Estimates of Parameters in Continuous–Time Systems," in *Analysis and Optimization of Stochastic Systems*, eds. O. Jacobs et alt., Academic Press (1980), New York, 437–450.

[6] A. V. Balakrishnan, "Stochastic Differential Systems," Lecture Notes in Economics and Mathematical Systems (1973), Springer Verlag, New York.

[7] V. Borkar and A. Bagchi, "Parameter Estimation in Continuous–Time Stochastic Processes," *Stochastics* **8** (1982), 193–212.

[8] R. S. Bucy "Nonlinear Filtering," *IEEE Trans. on Automatic Control* **10 no. 1** (1965), 198.

[9] R. S. Bucy and P. D. Joseph "*Filtering for Stochastic Processes with Applications to Guidance*," Wiley Interscience (1968), New York.

[10] P. E. Caines, "Stationary Linear and Nonlinear Systems Identification and Predictor Set Completeness," *IEEE Transactions on Automatic Control* **23 no. 4** (1978), 583–594.

[11] D. F. Delchamps, "A Note on the Analyticity of the Riccati Metric," in *Lectures in Applied Mathematics* **18** (1980), American Mathematical Society, 37–41.

[12] R. J. Elliot, "Reverse-Time Markov Processes," *IEEE Transactions on Information Theory* **32 no.2** (1986), 290–292.

[13] W. Fleming and S. K. Mitter, "Optimal Stochastic Control and Pathwise Filtering of Non-degenerate Diffusions," *Stochastics* **8** (1982), 63–77.

[14] H. Folmer, "An Entropy Approach to the Time Reversal of Diffusion Processes," in "Stochastic Differential Systems: Filtering and Control," M. Métivier and E. Pardoux, eds., Lecture Notes in Control and Information Sciences 69 (1985), Springer Verlag, New York., 156-163.

[15] U. G. Haussman and E. Pardoux, " Time Reversal of Diffusion Processes," in "Stochastic Differential Systems: Filtering and Control," M. Métivier and E. Pardoux, eds., Lecture Notes in Control and Information Sciences 69 (1985), Springer Verlag, New York., 176–182.

[16] H. Kunita, "Stochastic Partial Differential Equations Connected with Nonlinear Filtering," in *Nonlinear Filtering and Stochastic Control*, eds. S. K. Mitter and A. Moro, Lecture Notes in Mathematics 972 (1982), Springer Verlag, New York.

[17] R. S. Lipster and A. N. Shiryayev "*Statistics of Random Processes*," (1977), Springer Verlag, New York.

[18] L. Ljung, "Asymptotic Behavior of the Extended Kalman Filter as a Parameter Estimator for Linear Systems," *IEEE Trans. on Automatic Control* **24 no. 1** (1979), 36–50.

[19] S. K. Mitter "Lectures on Nonlinear Filtering and Stochastic Control," in *Nonlinear Filtering and Stochastic Control*, eds. S. K. Mitter and A. Moro, Lecture Notes in Mathematics 972 (1982), Springer Verlag, New York.

[20] S. K. Mitter "Approximations for Nonlinear Filters," in *Nonlinear Stochastic Problems*, eds. R. S. Bucy and J. M. F. Moura (1982), D. Reidel, Dordrecht, Holland.

[21] R. E. Mortensen, "Maximum Likelihood Recursive Nonlinear Filtering," *Journal of Optimization Theory and Applications* **2** (1968), 386–394.

[22] J. M F. Moura, S. K. Mitter, and L. Ljung, "Recursive Maximum Likelihood Estimation: Case Studies," to be submitted.

[23] E. Pardoux, "The solution of the Nonlinear Equation as a Likelihood Equation," *IEEE 20th Decision and Control Conference* (1981), S. Diego, California, 316–319.

[24] H. Robbins and S. Monro, "A Stochastic Aprroximation Method," *Annales of Mathematical Statistics* **22** (1951), 400–407.

[25] N. R. Sandell, Jr. and K. I. Yared, "Maximum Likelihood Identification of State Space Models for Linear Dynamical Systems," (1976), Report ESL-R-814, Electronic System Laboratory, M. I. T. .

[26] F. C. Schweppe, "Evolution of Likelihood Functions for Gaussian Systems," *IEEE Trans. on Information Theory* **11** (1965),61–70.

[27] T. Soderstrom, "On the Convergence Properties of the Generalized Least Squares Identification Method," *Automatica* **10** (1974), 617–626.

[28] D. W. Stroock and S. R. S. Varadhan, "*Multidimensional Diffusion Processes*," (1979),Springer Verlag, New York.

[29] D. W. Stroock, "Lectures in Stochastic Differential Equations," Tata Institute, Springer Verlag, New York.

[30] J. K. Tugnait, "Continuous Time Systems Identification on Compact Parameter Sets," *IEEE Transactions on Information Theory* **31 no. 5** (1985), 652–659.

[31] H. L. Van Trees, "*Detection, Estimation, and Modulation Theory Part I,*" (1968), John Wiley, New York.

[32] P. C. Young, "An Instrumental Variable Method for Real-Time Identification of a Noisy Process," *Automatica* **6** (1970), 271–287.

[33] M. Zakai, "On the Optimal Filtering of Diffusion Processes," *Z. Wahrscheinlichkeitstheorie Verw. Gebiete* **11** (1969), 230–243.

Appendix A: Identification of Linear Systems

In this appendix, we write the detailed equations obtained when the state equation and the observation equation are linear in the state. To avoid burdening the notation, the processes and the unknown parameter vector are all taken to be scalars. The gradient of the log-likelihood function along the ML-trajectory leads to

$$\widehat{S}_x(t) = \widehat{P^{-1}}(t)\left[\widehat{x}(t) - \widehat{\mu}(t)\right] - \widehat{\overline{P}^{-1}}(t)\left[\widehat{x}(t) - \widehat{\overline{x}}(t)\right] \tag{A-1}$$

and

$$\begin{aligned}\widehat{S}_\theta(t) = &-\widehat{P^{-1}}(t)\left[\widehat{x}(t) - \widehat{\mu}(t)\right]\widehat{\mu_\theta}(t) + \widehat{\overline{P}^{-1}}(t)\left[\widehat{x}(t) - \widehat{\overline{x}}(t)\right]\widehat{\overline{x}_\theta}(t) \\ &+ \frac{1}{2}\left(\widehat{P^{-1}}(t)\right)_\theta\left[\widehat{x}(t) - \widehat{\mu}(t)\right]^2 - \frac{1}{2}\left(\widehat{\overline{P}^{-1}}(t)\right)_\theta\left[\widehat{x}(t) - \widehat{\overline{x}}(t)\right]^2 \\ &+ \frac{2}{r}\int_0^t \widehat{\left(H\mu(s,\theta)\right)_\theta}\, dy(s) - \frac{2}{r}\int_0^t \left[H\mu(s,\theta)\right]\widehat{\left[\left(H\mu(s,\theta)\right)_\theta\right]}\, ds \\ &+ \left(\ln\frac{\mid \widehat{P}(t)\mid}{\mid \widehat{\overline{P}}(t)\mid}\right)_\theta .\end{aligned} \tag{A-2}$$

By stationarity, it follows from (A–1) the useful relations

$$\widehat{P^{-1}}(t)\left[\widehat{x}(t) - \widehat{\mu}(t)\right] = \widehat{\overline{P}^{-1}}(t)\left[\widehat{x}(t) - \widehat{\overline{x}}(t)\right] \tag{A-3.a}$$

$$= \left[\widehat{\overline{P}}(t) - \widehat{P}(t)\right]^{-1}\left[\widehat{\mu}(t) - \widehat{\overline{x}}(t)\right] \tag{A-3.b}$$

and

$$\widehat{x}(t) = \left[\widehat{P^{-1}}(t) - \widehat{\overline{P}^{-1}}(t)\right]^{-1}\left[\widehat{P^{-1}}(t)\widehat{\mu}(t) - \widehat{\overline{P}^{-1}}(t)\widehat{\overline{x}}(t)\right]. \tag{A-3.c}$$

From (A–2), and using (A–3),

$$\begin{aligned}&\frac{2}{r}\int_0^t \widehat{\left(H\mu(s,\theta)\right)_\theta}\, d\mathcal{I}(s,\theta) + \left(\ln\frac{\mid \widehat{P}(t)\mid}{\mid \widehat{\overline{P}}(t)\mid}\right)_\theta = \\ &\qquad \left[\widehat{\overline{P}}(t) - \widehat{P}(t)\right]^{-1}\left[\widehat{\mu}(t) - \widehat{\overline{x}}(t)\right]\left[\widehat{\mu_\theta}(t) - \widehat{\overline{x}_\theta}(t)\right] \\ &\qquad -\frac{1}{2}\left[\widehat{P_\theta}(t) - \widehat{\overline{P}_\theta}(t)\right]\left[\widehat{\overline{P}}(t) - \widehat{P}(t)\right]^{-2}\left[\widehat{\mu}(t) - \widehat{\overline{x}}(t)\right]^2.\end{aligned} \tag{A-4}$$

The pseudo innovations are

$$d\mathcal{I}(t,\theta) = dy(t) - H\mu(t,\theta)dt. \tag{A-5}$$

The left hand side of (A–4) is the parameter's only log-likelihood function. The right hand side shows how the prior information on the process $x(t)$ affects the parameter estimation. When

$$\overline{P}^{-1}(0) = 0, \tag{A-6}$$

it results that

$$\widehat{x}(t) = \widehat{\mu}(t), \tag{A–7}$$

and the right hand side of (A–4) is zero. The parameter ML-estimate is then that value of θ for which the inner product

$$< (H\mu)_\theta, I_0^t > \equiv 0, \tag{A–8}$$

i.e., that value of θ for which the θ-gradient of the sensor $H\mu(t)$ is orthogonal to the innovations. Under no prior knowledge on the state process statistics, we recover then the usual ML-parameter interpretation.

Remark: The $\widehat{\cdot}$ notation over the first integral in (A–2) stands for a measurable version of the integral evaluated at the ML-estimates at time t. Over the second integral, it means evaluation of the integral over a Borel measurable version of the integrand evaluated at the ML-estimates at time t. As discussed previously, completeness of the σ-fields and continuity of the several functions on the arguments guarantee the existence of such measurable versions.

Carrying out the higher order partial derivatives, obtain along the ML-trajectory

$$\widehat{S}_{xx}(t) = \widehat{P^{-1}}(t) - \widehat{\overline{P}^{-1}}(t) \tag{A–9}$$

$$\widehat{S}_{x\theta}(t) = -\widehat{P^{-1}}(t)\widehat{\mu_\theta}(t) + \widehat{\overline{P}^{-1}}(t)\widehat{\overline{x}_\theta}(t) \tag{A–10}$$
$$+ \widehat{(P^{-1}(t))_\theta}[\widehat{x}(t) - \widehat{\mu}(t)] - \widehat{(\overline{P}^{-1}(t))_\theta}[\widehat{x}(t) - \widehat{\overline{x}}(t)]$$

$$\widehat{S}_{\theta\theta}(t) = \widehat{P^{-1}}(t)[\widehat{\mu_\theta}(t)]^2 - \widehat{\overline{P}^{-1}}(t)[\widehat{\overline{x}_\theta}(t)]^2 \tag{A–11}$$
$$- 2\widehat{(P^{-1}(t))_\theta}[\widehat{x}(t) - \widehat{\mu}(t)]\widehat{\mu_\theta}(t) + 2\widehat{(\overline{P}^{-1}(t))_\theta}[\widehat{x}(t) - \widehat{\overline{x}}(t)]\widehat{\overline{x}_\theta}(t)$$
$$- \widehat{P^{-1}}(t)[\widehat{x}(t) - \widehat{\mu}(t)]\widehat{\mu_{\theta\theta}}(t) + \widehat{\overline{P}^{-1}}(t)[\widehat{x}(t) - \widehat{\overline{x}}(t)]\widehat{\overline{x}_{\theta\theta}}(t)$$
$$+ \frac{1}{2}\widehat{(P^{-1}(t))_{\theta\theta}}[\widehat{x}(t) - \widehat{\mu}(t)]^2 - \frac{1}{2}\widehat{(\overline{P}^{-1}(t))_{\theta\theta}}[\widehat{x}(t) - \widehat{\overline{x}}(t)]^2$$
$$+ \int_0^t \widehat{(H\mu(s,\theta))_{\theta\theta}}dy(s) - \int_0^t \widehat{[(H\mu(s,\theta))_\theta]^2}ds$$
$$- \int_0^t \widehat{[H\mu(s,\theta)][(H\mu(s,\theta))_{\theta\theta}]}ds + \widehat{\left(\ln \frac{|\widehat{P}(t)|}{|\widehat{\overline{P}}(t)|}\right)_{\theta\theta}}$$

$$\widehat{S}_{xx\theta}(t) = \widehat{(P^{-1}(t))_\theta} - \widehat{(\overline{P}^{-1}(t))_\theta}. \tag{A–12}$$

To compute the filter, we need the determinant of the Hesssian matrix. Combining the terms in $\widehat{S}^2_{x\theta}$ with corresponding terms in the product $\widehat{S}_{xx}$ by $\widehat{S}_{\theta\theta}$ and using relations (A–3), obtain

$$
\begin{aligned}
|\,\widehat{\nabla^2 S}(t)\,| &= -\widehat{P^{-1}}(t)\widehat{P^{-1}}(t)\left[\widehat{\mu_\theta}(t)-\widehat{\bar{x}_\theta}(t)\right]^2 \\
&- \widehat{P^{-1}}(t)\widehat{\overline{P}^{-1}}(t)\left[\widehat{\mu}(t)-\widehat{\bar{x}}(t)\right]\left[\widehat{\mu_{\theta\theta}}(t)-\widehat{\bar{x}_{\theta\theta}}(t)\right] \\
&+ \widehat{P^{-1}}(t)\widehat{\overline{P}^{-1}}(t)\left[\widehat{\bar{P}}(t)-\widehat{P}(t)\right]^{-2}\left[\widehat{\mu}(t)-\widehat{\bar{x}}(t)\right]^2\left\{\frac{1}{2}\left[\widehat{\bar{P}_{\theta\theta}}(t)-\widehat{P_{\theta\theta}}(t)\right]\left[\widehat{\bar{P}}(t)-\widehat{P}(t)\right]\right. \\
&\left.- \left[\widehat{\bar{P}_\theta}(t)-\widehat{P_\theta}(t)\right]^2\right\} + \left[\widehat{P^{-1}}(t)-\widehat{\overline{P}^{-1}}(t)\right]\left\{\frac{2}{r}\int_0^t \widehat{\left(H\mu(s,\theta)\right)_{\theta\theta}}\,d\mathcal{I}(s,\theta) - \frac{2}{r}\int_0^t\left[\widehat{\left(H\mu(s,\theta)\right)_\theta}\right]^2 ds\right. \\
&\left.+\left(\ln\frac{|\,\widehat{\bar{P}}(t)\,|}{|\,\widehat{\bar{P}}(t)\,|}\right)_{\theta\theta}\right\}.
\end{aligned}
\tag{A–13}
$$

Under (A–6), i.e., no prior knowledge on the process statistics, the ML state estimate coincides with the conditional mean as stated in (A–7). Equations (A–9) to (A–12) become

$$\widehat{S}_{xx}(t) = \widehat{P^{-1}}(t) \tag{A–14}$$

$$\widehat{S}_{x\theta}(t) = -\widehat{P^{-1}}(t)\widehat{\mu_\theta}(t) \tag{A–15}$$

$$
\begin{aligned}
\widehat{S}_{\theta\theta}(t) = {} & \widehat{P^{-1}}(t)\left[\widehat{\mu_\theta}(t)\right]^2 + \int_0^t \widehat{\left(H\mu(s,\theta)\right)_{\theta\theta}}\,dy(s) \\
& - \int_0^t\left[\widehat{\left(H\mu(s,\theta)\right)_\theta}\right]^2 ds - \int_0^t\left[H\mu(s,\theta)\right]\widehat{\left[\left(H\mu(s,\theta)\right)_{\theta\theta}\right]}\,ds + \left(\ln|\,\widehat{P}(t)\,|\right)_{\theta\theta}
\end{aligned}
\tag{A–16}
$$

$$\widehat{S}_{xx\theta}(t) = \widehat{\left(P^{-1}(t)\right)_\theta} \tag{A–17}$$

Finally, (A–13) is now

$$
\begin{aligned}
|\,\widehat{\nabla^2 S}(t)\,| = \widehat{P^{-1}}(t)\left\{\frac{2}{r}\int_0^t \widehat{\left(H\mu(s,\theta)\right)_{\theta\theta}}\,d\mathcal{I}(s,\theta) - \frac{2}{r}\int_0^t\left[\widehat{\left(H\mu(s,\theta)\right)_\theta}\right]^2 ds\right. \\
\left.+\left(\ln|\,\widehat{P}(t)\,|\right)_{\theta\theta}\right\}.
\end{aligned}
\tag{A–18}
$$