ON STOCHASTIC DIFFERENTIAL EQUATIONS ARISING IN STATE ESTIMATION PROBLEMS

by

Frank Paul Romeo

Memorandum No. ERL-M 246

7 May 1968

ELECTRONICS RESEARCH LABORATORY

College of Engineering
University of California, Berkeley
94720

ON STOCHASTIC DIFFERENTIAL EQUATIONS ARISING IN STATE ESTIMATION PROBLEMS

Abstract

Two problems of importance concerning the estimation of state in dynamical systems are considered. First, a stochastic partial differential equation for the conditional density of the state given the observation is derived. Secondly, a new formulation of the state estimation problem is given. This formulation explicitly incorporates as a constraint the recursive form of the nonlinear filter. The resulting equations for finding the optimal filter are reasonably tractable and are investigated in a number of examples.

Research sponsored by the National Science Foundation under Grant GK-716.

CONTENTS

CHAPTER I

INTRODUCTION AND PRELIMINARIES

Introduction

This thesis is principally concerned with state estimation problems for systems whose dynamics can be modeled by stochastic differential equations, i.e., systems whose state $x(t)$ satisfies:

$$1.1 \qquad dx(t) = m(x(t),t)dt + \sigma(x(t),t)d\xi(t),$$

where $\xi(t)$ is a standard Brownian motion. The precise interpretation of 1.1 will be given later. Suffice it to remark at this point that $d\xi(t)$ plays the role of a white noise and is not inconsistent with the actual situation often encountered in practice. Although for simplicity the state is assumed to be scalar-valued in most of this thesis, almost all of the results can be generalized to higher dimensions. For the estimation problem it is assumed that the state $x(t)$ can not be directly observed. Instead, the observation $y(t)$ is related to $x(t)$ via an observation equation.

$$1.2' \qquad dy(t) = \mu(y(t),t)dt + n(x(t),t)dt + d\eta(t),$$

where $\eta(t)$ is a second Brownian motion. It should be noted that little loss of generality is incurred by assuming that the drift term $\mu(y(t),t)$ in 1.2' is zero, i.e.,

1.2 $dy(t) = n(x(t),t)dt + dq(t)$.

This is because whenever

$$dy(t) = \mu(y(t),t)dt + dq(t)$$

has a unique solution $y(t)$ in terms of $q(t)$ 1.2' can be transformed into 1.2 with no drift term. The existence of a unique solution imposes only mild restrictions on $\mu(y(t),t)$. The basic problem that will be considered is to find the "best" estimator $\hat{x}(t)$ of $x(t)$ using as data $y(s)$, $t_0 \leq s \leq t$. This estimation problem underlies much of the statistical analysis of dynamical systems. It is basic to stochastic control theory, identification theory, and detection theory.

It is well known that the conditional expectation:

$$\hat{x}(t) = E\{x(t) \mid y(s),\ t_0 \leq s \leq t\}$$

minimizes the mean square error $E[x(t) - x(t)]^2$ among all estimators which depend only on the observed data $\{y(s),\ t_0 \leq s \leq t\}$. In general, the conditional mean $\hat{x}(t)$ cannot be computed recursively. That is, to compute $\hat{x}(t+\Delta)$ it is not enough to know $\hat{x}(t)$ and $y(s)$, $t \leq s \leq t+\Delta$. Indeed, in general there exists no finite-dimensional vector $\underline{z}(t)$ such that $\underline{z}(t)$ can be recursively computed and in which $\hat{x}(t)$ can be imbedded.

If one demands recursive computation (and this is a practical requirement) then the conditional density of $x(t)$ given $y(s)$, $t_0 \leq s \leq t$ is often the best quantity to be

computed. Stratonovich (Ref. 2) appears to have been the first to suggest that the conditional density of x(t), given the observation, satisfies a stochastic partial differential equation which bears superficial similarity to the Fokker-Planck equation. He was followed by Kushner and others (Ref. 4,5,6,7). Because of a lack of clarity with respect to the stochastic calculas (in the Ito sense), these early papers were not entirely satisfactory even as heuristic expositions. Mortensen's thesis (Ref. 5) contains a precise formulation of the problem of determining the recursive equation which is satisfied by the conditional density. Unfortunately his theorem requires a strong hypothesis which is not satisfied even by the case where l.l is linear. In Chapters II and III Mortensen's results are improved upon. First, in Chapter II, the existence of the density is proved using a result of Prokhorov (Ref. 1). In Chapter III, it is proved that the density satisfies a stochastic partial differential equation. The result of Chapter III represents an improvement over the corresponding result of Mortensen. In course of the research of this thesis, the work of Duncan and the work of Zakai on the same problem appeared. The last named work appears to have resolved all the outstanding difficulties attending the problem, and definitely represents an improvement over his (Mortensen's) thesis. However, since the form of the equation for the density has long been conjectured, technique of proof acquires an independent interest. In this respect, the results of

Chapter III are sufficiently different from other approaches to warrant one more exposition.

Any nonlinear filter, if it is to be implemented, must necessarily be a compromise between the best estimator and a realizable device. Neither the conditional mean (because it is not recursive) nor the conditional density (because it is infinite dimensional) is an implementable device. One common technique is simply to linearize the equations and calculate the appropriate Kalman filter. A very interesting method was recently proposed by Kushner (Ref. 9). By a clever truncation of the system of moment equations he is able to approximate the conditional expectation dynamics with a finite dimensional differential equation.

The filter proposed in Chapter IV is based on the philosophy that if one can't implement the dynamics of the best estimator, then use the best dynamic estimator. What is meant by "dynamic estimator" is a filter with a recursive property so the estimator is continuously updated with the reception of new data. The structure will be:

$$1.3 \qquad dz(t) = g(z(t),t)dt + f(z(t),t)dy(t),$$

where of course $z(t)$ is the estimate of $x(t)$. Heuristically, equation 1.3 says: The estimator at time $(t+\Delta)$ is a function of the estimator at time t plus a multiple of the new information, $(y(t+\Delta) - y(t))$. The problem of designing the optimal filter is now transformed into the problem of

specifying $f(\cdot,\cdot)$ and $g(\cdot,\cdot)$ according to some criterion.

To simply say that the loss is proportional to the square of the error is no longer sufficient to properly define the problem. The error at time t, i.e. $E[x(t) - z(t)]^2$, is of course a function of $f(z(\tau),\tau)$ and $g(z(\tau),\tau)$ only for $\tau \in [0,t]$. The filter that minimizes the error at time t may be much different from the one that minimizes the average error at time t and at time t/2. The nonexistence of a uniformly best estimator of the form 3.1 (except in special cases) necessitates a more explicit definition of optimality. In keeping with the spirit of updating the estimator, the notion of "sequentially best" is introduced.

The main result of Chapter IV is an algorithm for generating the sequentially best estimator. That is to say, the theorem of Chapter IV prescribes three equations, the simultaneous solution of which yields two functions $f(\cdot,\cdot)$ and $g(\cdot,\cdot)$ and the transition density for x(t) and z(t). The two functions are such that if a filter is constructed so that the output satisfies equation 3.1, then z(t) will be the sequentially best estimator.

Although the equations specifying the sequentially best estimator can seldom be solved analytically, the point to be emphasized is that they can be computed off-line. Once computed, they completely determine a feedback realization for the optimal estimator. In a very real sense this achieves the goal of recursive filtering.

The estimator defined by equation 3.1 is general in that it is nonlinear and time varying, but restrictive in being one dimensional. The estimator can be made more accurate by imbedding it in a vector of increasing dimensionality which satisfies an equation similar to 3.1. Doing it appears to pose no great difficulty and results similar to those of one dimension can be expected.

It is unfortunate that no example other than the linear case was found to be tractable analytically. With linearity the distributions become Gaussian and the algorithm coincides exactly with the method for finding the Kalman filter (Ref. 8). The reduction of the "sequentially best" recursive estimator in the linear case to the "uniformly best" recursive estimator (i.e., Kalman) gives considerable weight to the belief that "sequentially best" estimators are indeed good estimators in the general case.

In search of an example, numerical calculations were undertaken with varying degrees of success. These results appear in Chapter V. They remove all doubts whether numerical techniques are feasible for solving the algorithm and generating $g(\cdot,\cdot)$ and $f(\cdot,\cdot)$ off-line: at the same time they show that careful attention must be paid to the numerical analysis to avoid instabilities and approximation errors.

Preliminaries

A stochastic process $\{x(t,\omega),\ t \in [0,T], \omega \in \Omega\}$ is a parameterized family of random variables on a fixed probability space $(\Omega, \mathcal{A}, P)$. When explicit indication of the ω dependence is not essential, it will be suppressed. A standard Brownian motion is a stochastic process satisfying the following conditions:

(a) $w(t)$ has independent and Gaussian distributed increments

(b) $E[w(t) - w(s)] = 0, \quad w(0) = 0$

(c) $E[w(t) - w(s)]^2 = |t - s|$

A Brownian motion, if separable, is almost surely sample continuous, a fact first discovered by Norbert Wiener. We shall consider only separable Brownian motions. The sample functions of a separable Brownian motion, though continuous, are very irregular, as is demonstrated by the following theorem.

Theorem 1

Let $T_n = \{0 = t_0^n < t_1^n < t_2^n < \cdots < t_n^n = T\}$ be a sequence of nested partitions of $[0,T]$ such that $\max_{1 \le k \le n} |t_k^n - t_{k-1}^n| \longrightarrow 0$ as $n \longrightarrow 0$. Then,

$$1.4 \qquad \lim_{n \to \infty} \sum_{i=0}^{n-1} [w(t_{i+1}^n) - w(t_i^n)]^2 = T \text{ almost surely.}$$

From this theorem follows the fact that almost all sample paths are of unbounded variation. This precludes interpretation of functionals of the form:

1.5 $\int_0^T f(t)dw(t,\omega)$

as ordinary Stieltjes integrals. Wiener (Ref. 16) was the first to give meaning to 1.5, but Itô (Ref. 11) enlarged the theory to include in particular the case when the intergrand also depends on ω. A great deal of sophisticated development is summarized by the following theorems.

Throughout the theorems stated below $\{a_t,\ t \geq 0\}$ denotes a monotone increasing sequence of sub-σ-algebras of a with the property that w(t) is measurable with respect to a_t and $(w(t+\Delta) - w(t))$ is independent of a_t.

Theorem 2

For the same partition used in theorem 1, let $f(t,\omega)$ be measurable with respect to Lebesgue measure for each ω and measurable with respect to a_t for each t, then:

1.6 $$\lim_{n\to\infty} \sum_{i=0}^{n-1} f(t_i^n,\omega)(w(t_{i+1}^n,\omega) - w(t_i^n,\omega)) = \int_0^T f(t,\omega)dw(t,\omega)$$

The convergence is in probability if :

$\int_0^T f^2(t,\omega)dt < \infty$ with probability one, and in quadratic mean if:

1.7 $$E\left\{\int_0^T f^2(t,\omega)dt\right\} < \infty.$$

The limiting random variable in 1.6 is called a stochastic integral and displays some important properties to be used in this sequel.

Theorem 3

If $\int_0^T (f_n(t) - f(t))^2 dt \longrightarrow 0$ in probability as $n \to \infty$,

then $\sup_{0 \le t \le T} \left| \int_0^t f_n(\tau) dw(\tau) - \int_0^t f(\tau) dw(\tau) \right| \longrightarrow 0$ in probability.

As a function of the upper limit the following result will be useful.

Theorem 4

If 1.7 holds, then: $I(t) = \int_0^t f(\tau)\, dw(\tau)$ is a martingale with almost surely continuous sample paths.

A function of two variables $m(\cdot,\cdot)$ is said to obey a uniform Lipshitz condition in the first variable if there is some positive constant K such that:

$$|m(x_1,t) - m(x_2,t)| < K|x_1 - x_2| .$$

Theorem 5

If $m(\cdot,\cdot)$ and $\sigma(\cdot,\cdot)$ obey a uniform Lipshitz condition in the first variable, then

$$1.8 \qquad x(t) = x(0) = \int_0^t m(x(\tau),\tau)d\tau + \int_0^t \sigma(x(\tau),\tau)dw(\tau)$$

has a unique solution.

This integral equation makes precise the meaning of the solution of the stochastic differential equation,

$$1.9 \qquad dx(t) = m(x(t),t)dt + \sigma(x(t),t)dw(t).$$

The following is frequently called the Ito differential rule.

Theorem 6

Let $F(u,v)$ be twice differentiable in u and once in v. If $y(t) = F(x(t),t)$, where $x(t)$ is defined by 1.8 then:

$$dy(t) = F_1(x(t),t)dx(t) + \tfrac{1}{2}F_2(x(t),t)\sigma^2(x(t),t)dt + F'(x(t),t)dt,$$

where $F_1(u,v) = \frac{\partial}{\partial u}F(u,v)$, $F_2(u,v) = \frac{\partial^2}{\partial u^2}F(u,v)$, and $F'(u,v) = \frac{\partial}{\partial v}F(u,v)$.

A heuristic justification of theorem 6 is not difficult if one replaces $(dw(t))^2$ by dt. This of course is not legal, but is intuitively correct in light of equation 1.4. The fact that $dw(t)$ acts like $(dt)^{\frac{1}{2}}$ is a source of some confusion in the earlier literature on applying stochastic calculus to physical problems. It necessitates the inclusion of higher order terms in each series expansion and must be handled delicately to obtain the exact stochastic differential equation representing the behavior of a particular quantity.

Equation 1.9 defines a map R from the space of Brownian motions to the space of solutions; both spaces are $C_{[0,T]}$, the space of continuous functions of t for t in the interval $[0,T]$.

Let α be the σ-algebra of sets in C generated by cylinder sets of the form:

$$A = \{x(\cdot) \in C \mid a_1 < x(t_1) < b_1, \cdots a_n < x(t_n) < b_n,$$
$$0 \leq t_1 < \cdots < t_n \leq T\}$$

The finite dimensional distributions of the Brownian motion extends uniquely to a probability measure W on (C,α) such that for cylinder sets W reduces to:

$$W(A) = [2^n \pi^n t_1(t_2-t_1) \cdots (t_n-t_{n-1})]^{-\frac{1}{2}} \int_{a_1}^{b_1}\int_{a_2}^{b_2} \cdots \int_{a_n}^{b_n}$$
$$\exp\left[-\frac{x_1^2}{2t_1} - \frac{(x_2-x_1)^2}{2(t_2-t_1)} - \cdots \frac{(x_n-x_{n-1})^2}{2(t_n-t_{n-1})}\right] dx_1 \cdots dx_n.$$

The measure W is universally referred to as Wiener measure. The map $R:C \to C$ takes sets of α into sets of α again, i.e. it is measurable. C now has two measures, the Wiener measure W and W^R, the measure induced by R; $W^R(A) = W(R^{-1}(A))$.

The following theorem is due to Prokhorov (Ref. 1). (See reference 15 for a thorough discussion of the topic.)

Theorem 7

If equation 1.8 is modified to read:

$$x(t) = \int_0^t m(x(\tau),\tau)d\tau + w(\tau),$$

and $m(\cdot,\cdot)$ is continuous in both variables and satisfies a uniform Lipshitz condition in the first, then

$$W^R(A) = \int_A p^R(x)dW, \quad \text{where x denotes a point in C, and}$$

1.10 $$p^R(x) = \exp\left\{\int_0^T m(x(t),t)dx(t) - \tfrac{1}{2}\int_0^T m^2(x(t),t)dt\right\}.$$

Theorem 7 states that under very weak conditions the measure induced by R is absolutly continuous with respect to Wiener

measure and the Radon-Nikodym derivative is the functional 1.10. In Chapter III it will be demonstrated that the transition density function $P(b,t|a,s)$ may be represented as a function space integral with respect to Wiener measure.

Condition (e) in the following theorem is not needed in certain cases, but is included for the sake of accuracy.

Theorem 8

If a Markov process $x(t)$ with transition density function $P(b,t|a,s)$ has the properties:

(a) $\lim_{h \downarrow 0} E\left\{ \frac{x(s+h) - x(s)}{h} \,\middle|\, x(s) = a \right\} = m(a,s)$

(b) $\lim_{h \downarrow 0} E\left\{ \frac{[x(s+h) - x(s)]^2}{h} \,\middle|\, x(s) = a \right\} = \sigma^2(a,s)$

(c) $m(\cdot,\cdot)$ and $\sigma^2(\cdot,\cdot)$ are twice differentiable and uniformly Lipshitz in the first variable

(d) for every $\varepsilon > 0$, the probability of the event $\{|x(t) - x(s)| > \varepsilon \,|\, x(s) = a\} = o(t - s)$

(e) $P(b,t|a,s)$ is three times differentiable in the space coordinates a and b,

then the transition density function satisfies:

$$\frac{\partial}{\partial t} P(b,t|a,s) = -\frac{\partial}{\partial b}[m(b,t)P(b,t\ a,s)] + \tfrac{1}{2}\frac{\partial}{\partial b^2}[\sigma^2(b,t)P(b,t\ a,s)]$$

The above differential equation was proposed and proved to hold under certain of the above conditions by A. N. Kolmogorov. It is classically known as the Fokker-Planck equation, but is frequently referred to as the forward Kolmogorov equation.

Doob (Ref. 13) provides consistency among the afore mentioned works by pointing out that the solution of the stochastic differential equations studied by Ito, et alii, do indeed satisfy the hypothesis of Kolmogorov's theorem. Thus a consistent and rich structure is provided which enmeshes the theory of stochastic differential equations with the more classic results from diffusion processes and Wiener integrals.

CHAPTER II

EXISTENCE OF A CONDITIONAL DENSITY

C will denote the space of continuous functions of t on $[0,T]$, all of which vanish at zero. Let CXC denote the product space of C with itself. (WXW) is the product Wiener measure over the σ-algebra a_∞ generated by the products of cylinder sets in C (Ref. 12).

Define a mapping R from CXC into itself by:

$R(\xi(\cdot),\eta(\cdot)) = (x(\cdot),y(\cdot))$, where

$$2.1 \qquad x(t) = \int_0^t m(x(\tau),\tau)d\tau + \xi(t), \quad \text{and}$$

$$2.2 \qquad y(t) = \int_0^t n(x(\tau),\tau)d\tau + \eta(t).$$

Let $(WXW)^R$ denote the measure induced by R.

In this chapter and the next $\sigma(x(t),t)$ in equation 1.1 is assigned the value of unity, hence 1.1 reduces to 2.1. Allowing a general $\sigma(\cdot,\cdot)$ would require special conditions bounding it away from zero or a piecewise approach to the problem. It is felt that the added complications are not warranted in this treatment of the topic.

Equations 2.1 and 2.2 point out another specialization, $x(0) = y(0) = 0$. There is no loss of generality here because appending initial conditions on $x(\cdot)$ and $y(\cdot)$ is a simple matter (Ref. 5) and not worth further attention. The main result of this chapter stems from the following theorem.

Theorem

If $m(\cdot,\cdot)$ and $n(\cdot,\cdot)$ are continuous in both variables and there exists a K such that:

$$|m(x_1,t) - m(x_2,t)| < K|x_1 - x_2| \quad \text{and}$$

$$|n(x_1,t) - n(x_2,t)| < K|x_1 - x_2|, \quad \text{then}$$

$$(WXW)^R(A) = \int_A p^R(x,y)d(WXW), \quad \text{where}$$

$$p^R(x,y) = \exp\left\{\int_0^T n(x(t),t)dy(t) + \int_0^T m(x(t),t)dx(t) - \tfrac{1}{2}\int_0^T n^2(x(t),t)dt - \tfrac{1}{2}\int_0^T m^2(x(t),t)dt\right\}.$$

Corollary

The distribution of x(t), conditioned on y(t) for all of t in the interval $[0,T]$, is absolutly continuous with respect to Lebesgue measure.

proof

For any Borel set A of the real line define:

$$C_A = \{x(\cdot) \in C \mid x(T) \in A\}.$$

By Fubini's theorem,

$$2.3 \qquad \int_{C_A} p^R(x,y)dW$$

exists for almost all $y(\cdot)$ sections. A conditional probability is assigned the real line by dividing 2.3 by:

$$\int_{C_B} p^R(x,y)dW, \quad \text{where } B = (-\infty,\infty).$$

If the Lebesgue measure of A is zero, the Wiener measure of C_A is zero, thus 2.3 is zero. This proves the corollary.

The theorem will be proven with the aid of several lemmas.

Lemma 1

R is a measurable map.

proof

The inverse image of sets of the form:

$$\{(x(\cdot),y(\cdot)) \mid x(t_i) < a \text{ for some } t_i \in [0,T]\}$$

are known to be measurable (Ref. 13). It remains only to consider half planes below some $y(\cdot)$ coordinate.

$$\{(\xi(t),\eta(t)) \mid y(t_i) < b \text{ for some } t_i \in [0,T]\}$$

$$= \bigcup_i \{(\xi(\cdot),\eta(\cdot)) \mid \eta(t_i) < r_i\} \cap \{(\xi(\cdot),\eta(\cdot)) \mid \int_0^{t_i} n(x(t),t)dt < (b - r_i)\}$$

where the union is taken over all retionals.

If $\int_0^t n(x(t),t)dt$ is viewed independently as a measurable map from C to C, then clearly each set in the union is measurable, which proves lemma 1.

We shall construct R^n, an approximation of R, as follows:

Let $T_n = \{0 = t_1^n, t_2^n, \cdots t_n^n = T\}$ be a sequence of nested partitions of $[0,T]$ with the property that

$$\max_k |t_{k+1}^n - t_k^n| \longrightarrow 0 \quad \text{as} \quad n \longrightarrow \infty .$$

If the partition points are always equally spaced, $(t^n_{i+1} - t^n_i) = \Delta t^n$ does not depend on i. Let this be the case; the resulting reduction of Δt^n is not essential to the proof, but simplifies the notation somewhat.

Define:

$$\Delta\xi^n_i = \xi(t^n_{i+1}) - \xi(t^n_i), \quad \Delta\eta^n_i = \eta(t^n_{i+1}) - \eta(t^n_i),$$

$$\Delta x^n_i = x(t^n_{i+1}) - x(t^n_i), \quad \Delta y^n_i = y(t^n_{i+1}) - y(t^n_i).$$

The continuous Markov process is approximated by:

$$x^n_{i+1} = x^n_i + m(x^n_i, t^n_i)\Delta t^n + \Delta\xi^n_i$$

$$y^n_{i+1} = y^n_i + n(x^n_i, t^n_i)\Delta t^n + \Delta\eta^n_i ,$$

where $x^n_0 = y^n_0 = 0$ for all n. The continuous version of the above Markov chain is achieved by:

$$x^n(t) = x^n_i + m(x^n_i, t^n_i)(t - t^n_i) + \xi(t) - \xi(t^n_i)$$

$$y^n(t) = y^n_i + n(x^n_i, t^n_i)(t - t^n_i) + \eta(t) - \eta(t^n_i),$$

for $t^n_i < t < t^n_{i+1}$ and $i = 1,2,\cdots n-1$.

Now $R(n)(\xi(\cdot),\eta(\cdot)) = (x^n(\cdot), y^n(\cdot))$.

If α_n denotes the σ-algebra generated by cylinder sets whose time indices are members of T_n, then the α_n's are monotonically increasing and α_∞ is generated by the sets of the union $\bigcup_1^\infty \alpha_n$.

Denote the measure induced by R(n) on the members of the algebra α_n by:

$$(WXW)^{R(n)}(F) = K_n \int_F \exp\{-(2\Delta t^n)^{-1} \sum_{i=0}^{n-1} [(x_{i+1}^n - x_i^n - m(x_i^n,t_i^n)\Delta t^n)^2 - (y_{i+1}^n - y_i^n - n(x_i^n,t_i^n)\Delta t^n)^2]\} dx_1 \cdot dx_n dy_1 \cdot dy_n$$

where $K_n = (2\pi\Delta t^n)^{-n/2}$ and $F \in \mathcal{A}_n$.

Clearly $(WXW)^{R(n)}$ is absolutly continuous with respect to (WXW) restricted to $\mathcal{A}_n$, hence there is a density function on CXC, denoted by:

$$p^{R(n)}(x,y) = \frac{d(WXW)^{R(n)}}{d(WXW)} .$$

From the above discussion it is seen that:

$$p^{R(n)}(x,y) = \exp \sum_{i=0}^{n-1} [m(x_i^n,t_i^n)(x_{i+1}^n - x_i^n) - n(x_i^n,t_i^n)(y_{i+1}^n - y_i^n) - \tfrac{1}{2}m^2(x_i^n,t_i^n)\Delta t^n - \tfrac{1}{2}n^2(x_i^n,t_i^n)\Delta t^n]$$

(the arguements of x(t) and y(t) will occasionally be omitted when there is no possibility of confusion)

With the sequence $\{R(n)\}$ so defined the following lemmas may be proven.

Lemma 2

If $H \subseteq CXC$ is compact in the topology of uniform convergence, then for every $\delta > 0$ there exists an η such that for $n \geq \eta$,

$$\sup_{0 \leq t \leq T} |x^n(t) - x(t)| < \delta \text{ and } \sup_{0 \leq t \leq T} |y^n(t) - y(t)| < \delta$$

for all $(\xi(\cdot),\eta(\cdot))$ H.

proof

If two points in CXC, $(\xi(\cdot),\eta(\cdot))$ and $(\xi'(\cdot),\eta'(\cdot))$ have the property:

$\sup_{0\le t\le T} |\xi(t) - \xi'(t)| < \alpha$ and $\sup_{0\le t\le T} |\eta(t) - \eta'(t)| < \alpha$, then

$(x^n(\cdot),y^n(\cdot)) = R(n)(\xi(\cdot),\eta(\cdot))$ and $(x'^n(\cdot),y'^n(\cdot))$

$R(n)(\xi'(\cdot),\eta'(\cdot))$ have the property:

$\sup_{0\le t\le T} |x^n(t) - x'^n(t)| < 2\alpha$ and $\sup_{0\le t\le T} |y^n(t) - y'^n(t)| < 2\alpha$

for all n such that $K\Delta t^n < (2T)^{-1}$.

(recall that K is the Lipshitz constant in the hypothesis of the theorem)

H is compact and R is continuous in the topology of uniform convergence. Thus for any given δ it is possible to pick β so that $\delta/4 > \beta > 0$ and

$\sup_{0\le t\le T} |x'(t) - x(t)| < \delta/4$ and $\sup_{0\le t\le T} |y'(t) - y(t)| < \delta/4$

whenever $\sup_{0\le t\le T} |\xi'(t) - \xi(t)| < \beta$ and $\sup_{0\le t\le T} |\eta'(t) - \eta(t)| < \beta$,

where $R(\xi'(\cdot),\eta'(\cdot)) = (x'(\cdot),y'(\cdot))$.

Choose a finite set $V = \{(\xi_1(\cdot),\eta_1(\cdot)),(\xi_2(\cdot),\eta_2(\cdot)),$ $\ldots (\xi_V(\cdot),\eta_V(\cdot))\}$ such that the β neighborhoods of the members of V cover H.

If $(x(\cdot),y(\cdot)) \in R(H)$ then there is some $(x_i(\cdot),y_i(\cdot))$ an element of R(V) such that:

$\sup_{0\le t\le T} |x(t) - x_i(t)| < \delta/4$ and $\sup_{0\le t\le T} |y(t) - y_i(t)| < \delta/4$.

Now choose η so large that $K(\Delta t^n)T < \beta$ and

$\sup_{0\le t\le T} |x_i^n(t) - x_i(t)| < \delta/4$ and $\sup_{0\le t\le T} |y_i^n(t) - y_i(t)| < \delta/4$

for $i = 1,2, \cdots ,\nu$ and for $n \ge \eta$.

Then for $n \ge \eta$, $\sup_{0\le t\le T} |x^n(t) - x(t)|$

$\le \sup_{0\le t\le T} |x^n(t) - x_i^n(t)| + \sup_{0\le t\le T} |x_i^n(t) - x_i(t)|$

$+ \sup_{0\le t\le T} |x_i(t) - x(t)| \le 2\beta + \delta/4 + \delta/4 \le \delta$.

The same inequality may be derived for $y(t)$, which proves lemma 2.

Lemma 3

As $n \to \infty$, $(WXW)^{R(n)} \to (WXW)^R$ in the sense of weak convergence of measures, i.e., for any bounded continuous real valued function $g(x,y)$ on CxC,

$$\int_{CXC} g(x,y)d(WXW)^{R(n)} \to \int_{CXC} g(x,y)d(WXW)^R \quad \text{as} \quad n \to \infty .$$

proof

Choose H compact and with the property:

$(WXW)(CXC - H) < M(\varepsilon/3)$, where $\sup_{(x,y)\in CXC} |g(x,y)| = M$.

That this is possible is proven in reference 1.

By the continuity of $g(\cdot,\cdot)$ and lemma 2, there exists an η so large that for $n \ge \eta$, $|g(R(n)(\xi,\eta)) - g(R(\xi,\eta))| < \varepsilon/3$ for all $(\xi,\eta) \in H$.

$$\left| \int_{CXC} g(x,y)d(WXW)^{R(n)} - \int_{CXC} g(x,y)d(WXW)^R \right|$$

is bounded by:

$$\left|\int_H g(R(n)(\xi,\eta))d(WXW) - \int_H g(R(\xi,\eta))d(WXW)\right|$$

$$+ \left|\int_{CXC-H} g(R(n)(\xi,\eta)d(WXW) - \int_{CXC-H} g(R(\xi,\eta))d(WXW)\right| < \varepsilon ,$$

since ε is arbitrary, this proves lemma 3.

Lemma 4

$$p^{R(n)}(x,y) \rightarrow p^{R}(x,y) \quad \text{almost surely (WXW).}$$

proof

$\sum_{i=0}^{n-1} -\frac{1}{2}m^2(x_i^n,t_i^n)\Delta t^n - \frac{1}{2}n^2(x_i^n,t_i^n)\Delta t^n$ converges to $-\frac{1}{2}\int_0^T m^2(x(t),t)dt - \frac{1}{2}\int_0^T n^2(x(t),t)dt$ pointwise in CXC.

$\sum_{i=0}^{n-1} m(x_i^n,t_i^n)(x_{i-1}^n - x_i^n)$ converges in quadratic mean to $\int_0^T m(x(t),t)dx(t)$. For every $x(t)$, $\sum_{i=0}^{n-1} n(x_i^n,t_i^n)(y_{i-1}^n - y_i^n)$ converges in mean to $\int_0^T n(x(t),t)dy(t)$. These facts are sufficient to insure that if a pointwise limit of $p^{R(n)}(x,y)$ exists, it must truly equal $p^{R}(x,y)$.

Now $p^{R(n)}(x,y)$ is measurable $\mathcal{A}_n$ and

$$\int_{F_m} p^{R(n)}(x,y)d(WXW) = \int_{F_m} p^{R(m)}(x,y)d(WXW),$$

where $F_m \in \mathcal{A}_m$ and $m < n$. Hence $\{p^{R(n)}(x,y), \mathcal{A}_n \mid n = 1,2,..\}$ is a martingale. Since $\int_{CXC} p^{R(n)}(x,y)d(WXW) = 1$ for all n, the martingale convergence theorem guarantees pointwise convergence almost surely, which proves lemma 4.

Lemma 5

If $A = \{(x(t),y(t)) \mid \sup_{0\le t\le T} |x^n(t)| < M$, and $\sup_{0\le t\le T} |y^n(t)| < M$ for all positive integers $n\}$, then for all $A' \subseteq A$,

$$\int_{A'} p^{R(n)}(x,y)d(WXW) \longrightarrow \int_{A'} p^{R}(x,y)d(WXW).$$

proof

Let $A_m^n = \{(x(t),y(t)) \mid \sup_{0\le t\le m\Delta t^n} |x^n(t)| < M$ and $\sup_{0\le t\le m\Delta t^n} |y^n(t)| < M\}$, and pick $\lambda > 1$.

Notice the monotone class $A_1^n \supseteq A_2^n \supseteq \cdots \supseteq A_n^n = A^n \subseteq A$.

$$\int_{A_m^n} \exp \lambda \Big\{ \sum_{i=0}^{m-1} m(x_i^n,t_i^n)(x_{i-1}^n - x_i^n) + n(x_i^n,t_i^n)(y_{i+1}^n - y_i^n) - \tfrac{1}{2}m^2(x_i^n,t_i^n)\Delta t^n - \tfrac{1}{2}n^2(x_i^n,t_i^n)\Delta t^n \Big\}\, d(WXW)$$

$$\le \int_{A_{m-1}^n} \exp \lambda \Big\{ \sum_{i=0}^{m-2} m(x_i^n,t_i^n)(x_{i+1}^n - x_i^n) + n(x_i^n,t_i^n)(y_{i+1}^n - y_i^n) - \tfrac{1}{2}m^2(x_i^n,t_i^n)\Delta t^n - \tfrac{1}{2}n^2(x_i^n,t_i^n)\Delta t^n \Big\} \exp -\lambda/2\Big\{ m^2(x_{m-1}^n,t_{m-1}^n)\Delta t^n + n^2(x_{m-1}^n,t_{m-1}^n)\Delta t^n \Big\} \exp \lambda \Big\{ m(x_{m-1}^n,t_{m-1}^n)(x_m^n - x_{m-1}^n) + n(x_{m-1}^n,t_{m-1}^n)(y_m^n - y_{m-1}^n) \Big\}\, d(WXW).$$

The above integral may be expressed as a multiple integral over the x_i^n 's and the y_i^n 's. The range of integration of x_m^n and y_m^n will both be $(-\infty,\infty)$. Recalling the fact that if Z is normally distributed , then $E\{\exp(iuZ)\} = \exp(-\tfrac{1}{2}u^2\sigma^2)$, evaluation of the m'th integral and returning the remaining multiple integral

to the Wiener integral notation yields:

$$\int_{A^n_{m-1}} \exp \lambda \Big\{ \sum_{i=0}^{m-2} m(x_i^n, t_i^n)(x_{i+1}^n - x_i^n) - n(x_i^n, t_i^n)(y_{i+1}^n - y_i^n) - \tfrac{1}{2} m^2(x_i^n, t_i^n)\Delta t^n - \tfrac{1}{2} n^2(x_i^n, t_i^n)\Delta t^n \Big\} \exp\Big\{ \tfrac{1}{2}(\lambda^2 - \lambda)(m^2(x_{m-1}^n, t_{m-1}^n) + n^2(x_{m-1}^n, t_{m-1}^n))\Delta t^n \Big\} \, d(WXW).$$

Application of the Lipshitz condition and the continuity as a function of t yields:

$$m^2(x(t),t) \leq \Big[\max_{0 \leq t \leq T} |m(0,t)| + K|x(t)| \Big]^2$$

$$< K'(1 + x^2(t)), \quad \text{for some appropriate } K'.$$

Similarly, $n^2(x(t),t) \leq K'(1 + x^2(t))$.

Thus $\int_{A^n_m} \exp \lambda \Big\{ \sum_{i=0}^{m-1} \cdot \cdot$ (same as above) $\cdot \Big\} d(WXW)$

$$\leq \exp\Big\{ (\lambda^2 - \lambda)K'(1 + M^2)\Delta t^n \Big\} \int_{A^n_{m-1}} \exp \lambda \Big\{ \sum_{i=0}^{m-2} \cdot \text{ (same) } \cdot \Big\} d(WXW).$$

Iteration of the above reduction yields:

$$\int_{A^n} \Big\{ p^{R(n)}(x,y) \Big\}^{\lambda} d(WXW) \leq \exp\Big\{ (\lambda^2 - \lambda)K'(1 + M^2) \Big\}$$

for all n. This condition and pointwise convergence of $p^{R(n)}$ to p^R (Ref. 14) proves lemma 5.

proof of theorem

Pick $\varepsilon > 0$ and define $A_d = \Big\{ (x(t),y(t)) \Big| \sup_{0 \leq t \leq T} |x^n(t)| < d$

and $\sup_{0 \leq t \leq T} |y^n(t)| < d$ for all $n \Big\}$.

Now $\lim_{d\to\infty} A_d = CXC$, so pick d so large that:

$$(WXW)^R(CXC - A_d) < \varepsilon/2.$$

By lemma 5 there is an η so large that

$$\int_{CXC-A_d} p^{R(n)}(x,y)d(WXW) - (WXW)^R(CXC - A_d) < \varepsilon/2.$$

Thus $\int_{CXC-A_d} p^{R(n)}(x,y)d(WXW) < \varepsilon$ for all $n \geq \eta$.

By Fatou's lemma and lemma 4, $\int_{CXC-A_d} p^R(x,y)d(WXW) < \varepsilon$.

For an arbitrary $A \subseteq CXC$ another application of Fatou's lemma yields:

$$\int_A p^R(x,y)d(WXW) \leq \liminf_{n\to\infty} \int_A p^{R(n)}(x,y)d(WXW).$$

By lemma 5 and the choice of A_d,

$$\limsup_{n\to\infty} \int_A p^{R(n)}(x,y)d(WXW) \leq \limsup_{n\to\infty} \int_{A\cap A_d} p^{R(n)}(x,y)d(WXW) + \varepsilon$$

$$= \int_{A\cap A_d} p^R(x,y)d(WXW) + \varepsilon \leq \int_A p^R(x,y)d(WXW) + \varepsilon,$$

hence $\int_A p^{R(n)}(x,y)d(WXW) \longrightarrow \int_A p^R(x,y)d(WXW)$;

in light of lemma 3,

$$\int_A p^R(x,y)d(WXW) = (WXW)^R(A),$$

which proves the theorem.

CHAPTER III

DYNAMICS OF THE DENSITY

The objective of this chapter is to find a dynamical equation governing the evolution in time of the probability density of $x(t)$ conditioned on $x(s)$ and $\{y(\tau),\ s \le \tau \le t\}$. Motivation for solving this problem in the context of non-linear filtering was discussed in the introduction.

Use will be made of what Gelfand and Yaglom (Ref. 12) call a "conditional Wiener Measure." Generally speaking, Wiener demonstrated that a large class of functionals could be integrated over C, the space of continuous functions on $[0,T]$ with $x(0) = 0$, with respect to Wiener measure. The method of calculation is as follows: The interval $[0,T]$ is partitioned by $T_n = \{0 = t_1^n,\ t_2^n,\ \cdots\ t_n^n = T\}$, where $T/n = \Delta$ and $t_k^n = k\Delta$; $x(\cdot) \in C$ is replaced by a step function $x^n(\cdot)$ which coincides with $x(\cdot)$ at the partition points, i.e. $x^n(t_i^n) = x(t_i^n)$, and $x^n(t)$ is constant between sample points. Furthermore, let x_i^n denote $x^n(t_i^n)$.

The functional to be integrated $F(x(\cdot))$ becomes a function of n variables, $F(x^n(\cdot)) = F(x_1^n, x_2^n, \cdots x_n^n)$. The integral is defined as:

$$\int_C F(x(t))W(dx) = \lim_{n\to\infty} (2\pi\Delta)^{-n/2} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} F(x_1^n, \cdots x_n^n)$$

$$\exp\left[-\frac{x_1^2}{2\Delta} \cdots - \frac{(x_n^n - x_{n-1}^n)^2}{2\Delta}\right] dx_1^n \cdots dx_n^n.$$

(This is roughly what was done in Chapter II)

Suppose the integration over x_n^n is not carried out, i.e. one takes the limit of the n-1 fold integral, keeping $x_n^n = a$ for all n. It follows from the theory of the Daniell integral that this induces a new measure on C, parameterized by a. If this measure is denoted by W(dx;a,T), then,

$$3.1 \quad \int_C F(x(\cdot))W(dx) = \int_{-\infty}^{\infty}\int_C F(x(\cdot))W(dx;a,T)da.$$

Equation 3.1 says that integration over C can be accomplished by first integrating over those members which end on a and then integrating over a.

With this notation the conditional probability that $x(t) \in A$, conditioned on $\{y(\tau), 0 \leq \tau \leq s\}$ can be represented succinctly by:

$$3.2 \quad \frac{\int_A P(a,t,y,s)\, da}{\int_{-\infty}^{\infty} P(a,t,y,s)\, da},$$

where $P(a,t,y,s) = \int_C p(x,t,y,s)\, W(dx;a,t)$

and $p(x,t,y,s) = \exp\left\{\int_0^t m(x(\tau),\tau)dx(\tau) - \frac{1}{2}\int_0^t m^2(x(\tau),\tau)d\tau \right.$

$$\left. + \int_0^s n(x(\tau),\tau)dy(\tau) - \frac{1}{2}\int_0^s n^2(x(\tau),\tau)d\tau \right\}.$$

The above only makes sense for $s \leq t$. That requirement can be obviated but this is not needed in this discussion.

P(a,t,y,s) is the joint density of x(t) and the observation $\{y(\tau),\ s \leq \tau \leq t\}$ evaluated at (a,y); the density is with respect to the product of Lebesgue and Wiener measures on $(-\infty X \infty)$ X C.

To be consistent requires that: $P(a,0,y,s) = \delta(a)$, the Dirac delta function, if $x(0) = 0$ with certainty. If the initial value of $x(\cdot)$ has a density $P(a)$, then of course $P(a,0,y,0) = P(a)$.

The contingency that $x(0) \neq 0$ will be incorporated into the notation in the following way: $W(dx;a,t)$ has been defined to be a measure on those functions of C with $x(0) = 0$ and $x(t) = a$. If the measure is to be further conditioned on $x(s) = b$, $s < t$, then the measure will be denoted by: $W(dx;a,t|b,s)$. The only functionals that will be integrated against this measure will be functionals with domain $C_{[s,t]}$, i.e. the space of continuous functions over the interval $[s,t]$ with $x(s) = b$. Clearly,

$$\int_{-\infty}^{\infty} W(dx;a,t|b,s)(2\pi s)^{-\frac{1}{2}}\exp\left\{-\frac{b^2}{2s}\right\}db = W(dx;a,t)$$

As an example consider the functional: $q(t,s,x)$

$$= \exp\left\{\int_s^t m(x(\tau),\tau)dx(\tau) - \tfrac{1}{2}\int_s^t m^2(x(\tau),\tau)d\tau\right\}.$$

By equation 1.10 the notation above may be used to represent the probability that the diffusion process of equation 1.1 falls in linear set A at time t, given it has value b at time s by:

$$\int_A Q(a,t|b,s)da, \text{ where } Q(a,t|b,s) = \int_C q(t,s,x)W(dx;a,t|b,s).$$

Furthermore, the Chapman - Kolmogorov equation may be written:

$$Q(a,t|b,s) = \int_{-\infty}^{\infty} Q(a,t\ c,u)Q(c,u|b,s)dc.$$

The time rate of change of the integrand of the numerator of 3.2 will be sought, so in truth the dynamics of the joint density and not the conditional density will be obtained. Obviously, the first is sufficient to find the second; the difference is that the conditional density is divided by the marginal density with respect to Wiener measure for $\{y(\tau),\ 0 \leq \tau \leq s\}$. Here, as in Chapter II, the diffusion coefficient $\sigma(\cdot,\cdot)$ is assigned the value of unity. Again this is not a necessary condition, but a simplifying one.

Theorem

If the Lipschitz and continuity conditions of the theorem of Chapter II are satisfied by $m(\cdot,\cdot)$ and $n(\cdot,\cdot)$, and $P(a,t,y,t)$ is three times differentiable in a for all $t \in [0,T]$ and almost all $y \in C_{[0,T]}$, then,

$$3.3 \quad P(a,T,y,T) - P(a,0,y,0)$$
$$= \int_0^T \left[\tfrac{1}{2}\frac{\partial^2}{\partial a^2} P(a,t,y,t) - \frac{\partial}{\partial a}(m(a,t)P(a,t,y,t))\right] dt$$
$$+ \int_0^T n(a,t)P(a,t,y,t)dy(t),$$

where the last integral is the limit in probability of:

$$3.4 \quad \sum_{i=0}^{n} n(a,i\Delta)P(a,i\Delta,y,i\Delta)(y(i\Delta+\Delta) - y(i\Delta)) \quad \text{as } n \to \infty.$$

proof

The difficult parts of the proof are dealt with in several lemmas; they will be assumed here and proved later.

Notice that the first integral contains terms that look like the Fokker-Planck equation; they come from an updating of the $x(\cdot)$ density alone. The second integral vanishes if $n(\cdot,\cdot)$ does and is a linear functional on $y(\cdot)$. This separability is exploited in the proof. By writing:

3.5 $P(a,T,y,T) - P(a,0,y,0)$

$$= \sum_{i=1}^{n} \left[P(a,i\Delta,y,i\Delta-\Delta) - P(a,i\Delta-\Delta,y,i\Delta-\Delta)\right]$$

$$+ \sum_{i=1}^{n} \left[P(a,i\Delta,y,i\Delta) - P(a,i\Delta y,i\Delta-\Delta)\right] ,$$

the incremental behavior is seen to be the sum of effects of first changing $x(t)$ to $x(t+\Delta)$, then changing the length of the observation curve. It is tempting to only look at the incremental behavior, i.e. find $\frac{\partial}{\partial t}P(a,t,y,t)$ and then assert $P(a,T,y,T) = \int_0^T \frac{\partial}{\partial t}P(a,t,y,t)$. This has been the pitfall of others. The non-existence of $dy(t)/dt$ requires that in order to have a precise interpretation, the second integral in 3.3 must be shown to be the limit of a quantity like 3.4.

By lemma 1, $\sum_{i=1}^{n} \left[P(a,i\Delta,y,i\Delta-\Delta) - P(a,i\Delta-\Delta,y,i\Delta-\Delta)\right]$ converges to $\int_0^T \left[\tfrac{1}{2}\frac{\partial^2}{\partial a^2}P(a,t,y,t) - \frac{\partial}{\partial a}(m(a,t)P(a,t,y,t))\right]dt$.

The second sum can be rewritten as:

$$\sum_{i=1}^{n} \int_C p(x,i\Delta,y,i\Delta-\Delta)\left(\exp\left\{\int_{i\Delta-\Delta}^{i\Delta} n(x(t),t)dy(t) - \tfrac{1}{2}\int_{i\Delta-\Delta}^{i\Delta} n^2(x(t),t)dt\right\} - 1\right)W(dx;a,i\Delta)$$

With a tedious expansion of the exponential the sum becomes:

$$3.6 \quad \sum_{i=1}^{n} \int_C p(x,i\Delta,y,i\Delta-\Delta)J\ W(dx;a,i\Delta)$$
$$+ \frac{1}{2!} \sum_{i=1}^{n} \int_C p(x,i\Delta,y,i\Delta-\Delta)J^2\ W(dx;a,i\Delta)$$
$$+ \frac{1}{3!} \sum_{i=1}^{n} \int_C p(x,i\Delta,y,i\Delta-\Delta)J^3\ W(dx;a,i\Delta)$$
$$+ \frac{1}{4!} \sum_{i=1}^{n} \int_C p(x,i\Delta,y,\sigma_i(x))J^4\ W(dx;a,i\Delta) \quad , \text{ where}$$
$$J = \int_{i\Delta-\Delta}^{i\Delta} n(x(t),t)dy(t) - \tfrac{1}{2}\int_{i\Delta-\Delta}^{i\Delta} n^2(x(t),t)dt,$$

and use is made of the fact:

$$p(x,i\Delta,y,i\Delta-\Delta)\exp\left\{\int_{i\Delta-\Delta}^{\sigma_i(x)} n(x(t),t)dy(t) - \tfrac{1}{2}\int_{i\Delta-\Delta}^{\sigma_i(x)} n^2(x(t),t)dt\right\}$$
$$= p(x,i\Delta,y,\sigma_i(x))$$ to absorbe the remainder term of the Taylor theorem into the integrand of the fourth integral.

Generally speaking, if e^a is expanded about e^0, the remainder term of the Taylor expansion is $\frac{a^n}{n!}e^b$ where $|b| \in [0,|a|]$. By the monotonicity of the exponential,

$$3.7 \qquad |e^0 - e^b| \leq |e^0 - e^a|.$$

In the case under consideration, the exponential is a functional on CXC so the remainder can be different for each point $(x(\cdot),y(\cdot))$. For each fixed point,

$$\int_{i\Delta-\Delta}^{t} n(x(s),s)dy(s) - \tfrac{1}{2}\int_{i\Delta-\Delta}^{t} n^2(x(s),s)ds$$

is a continuous function of the upper limit of the integral, i.e. a continuous function of t for $t \in [i\Delta-\Delta,i\Delta]$.

There must be some $\sigma_i(x)$ such that the remainder term will look like:

$$3.8 \qquad \frac{J^n}{n!} \exp\left\{\int_{i\Delta-\Delta}^{\sigma_i(x)} n(x(t),t)dy(t) - \tfrac{1}{2}\int_{i\Delta-\Delta}^{\sigma_i(x)} n^2(x(t),t)dt\right\}$$

and $\sigma_i(x) \in [i\Delta-\Delta, i\Delta]$ for each $x(\cdot)$. ($y(\cdot)$ never changes in the above discussion) The exponential in 3.8 is swept into the $p(x,i\Delta,y,\sigma_i(x))$ term in the last sum in equation 3.6. Measurability is no problem because the term in question is the difference of functionals which are known to be measurable.

By lemmas 2 and 3 the first sum converges in probability to $\int_0^T n(a,t)P(a,t,y,t)dy(t)$. By lemma 4 the second sum goes to zero. By lemma 5 the third sum vanishes; the proof is almost complete.

Because of 3.7 it must be true that:

$|p(x,i\Delta,y,\sigma_i(x) - p(x,i\Delta,y,i\Delta-\Delta)|$ is bounded by $|p(x,i\Delta,y,i\Delta) - p(x,i\Delta,y,i\Delta-\Delta)|$; as a result,

$$p(x,i\Delta,y,\sigma_i(x)) \leq p(x,i\Delta,y,i\Delta) + p(x,i\Delta,y,i\Delta-\Delta).$$

Thus the fourth sum in 3.6 is trapped and lemma 6 squeezes it to zero. This completes the proof of the theorem.

Lemma

$$\lim_{n\to\infty} \sum_{i=1}^{n} [P(a,i\Delta,y,i\Delta-\Delta) - P(a,i\Delta-\Delta,y,i\Delta-\Delta)]$$

$$= \int_0^T \left[\tfrac{1}{2}\frac{\partial^2}{\partial a^2} P(a,t,y,t) - \frac{\partial}{\partial a}(m(a,t)P(a,t,y,t)\right] dt.$$

proof

The proof mimics the alternative proof Kolmogorov (Ref.10) gave for the forward differential equation.

For each i, one may write the Chapman-Kolmogorov equation: $P(a,i\Delta,y,i\Delta-\Delta) - P(a,i\Delta-\Delta,y,i\Delta-\Delta)$

$$= \int_{-\infty}^{\infty} Q(a,i\Delta|b,i\Delta-\Delta)P(b,i\Delta-\Delta,y,i\Delta-\Delta)db - P(a,i\Delta-\Delta,y,i\Delta-\Delta),$$

where $Q(\cdot,\cdot\ \cdot,\cdot)$ is as defined earlier in this chapter. If the second term of the integrand is expanded, the above becomes:

$$P(a,i\Delta-\Delta,y,i\Delta-\Delta)\left\{\frac{1}{\Delta}\int_{-\infty}^{\infty} Q(a,i\Delta|b,i\Delta-\Delta)db - \frac{1}{\Delta}\right\}\Delta$$

$$+ \frac{\partial}{\partial a}P(a,i\Delta-\Delta,y,i\Delta-\Delta)\,\Delta\int_{-\infty}^{\infty}\frac{(b-a)}{\Delta}\,Q(a,i\Delta|b,i\Delta-\Delta)db$$

$$+ \frac{\partial^2}{\partial a^2}P(a,i\Delta-\Delta,y,i\Delta-\Delta)\,\Delta\int_{-\infty}^{\infty}\frac{(b-a)^2}{2!\,\Delta}\,Q(a,i\Delta|b,i\Delta-\Delta)db$$

$$+ \Theta\int_{-\infty}^{\infty}\frac{(b-a)^3}{3!\,\Delta}\,Q(a,i\Delta\,b,i\Delta-\Delta)db\ ,$$

where Θ is the appropriate remainder coefficient.

From Kolmogorov's paper it follows that:

$$\left\{\frac{1}{\Delta}\int_{-\infty}^{\infty} Q(a,i\Delta|b,i\Delta-\Delta)db - \frac{1}{\Delta}\right\} \to -\frac{\partial}{\partial a}m(a,i\Delta-\Delta)$$

$$\int_{-\infty}^{\infty}\frac{(b-a)}{\Delta}\,Q(a,i\Delta|b,i\Delta-\Delta)db \to -m(a,i\Delta-\Delta)$$

$$\int_{-\infty}^{\infty}\frac{(b-a)^2}{\Delta}\,Q(a,i\Delta|b,i\Delta-\Delta)db \to 1, \text{ and}$$

$$\int_{-\infty}^{\infty}\frac{(b-a)^3}{\Delta}\,Q(a,i\Delta|b,i\Delta-\Delta)db \to 0 \quad \text{as} \quad n\to\infty.$$

Thus making making the above expansion for each term in the sum and taking the limit proves lemma 1.

Lemma 2

$$\int_C p(x,i\Delta,y,i\Delta-\Delta) \int_{i\Delta-\Delta}^{i\Delta} n(x(t),t)dy(t)W(dx;a,i\Delta)$$

$$= \int_{i\Delta-\Delta}^{i\Delta} \int_C p(x,i\Delta,y,i\Delta-\Delta)n(x(t),t)W(dx;a,i\Delta)dy(t)$$

in probability.

proof

Partition $[i\Delta-\Delta,i\Delta]$ with $T_m = \{t_1^m = i\Delta-\Delta, t_2^m, \cdots t_m^m = i\Delta\}$,

then $\int_{i\Delta-\Delta}^{i\Delta} \int_C p(x,i\Delta,y,i\Delta-\Delta)n(x(t),t)W(dx;a,i\Delta)dy(t)$

$$= \lim_{m\to\infty} \sum_{j=1}^{m-1} \int_C p(x,i\Delta,y,i\Delta-\Delta)n(x(t_j^m),t_j^m)W(dx;a,i\Delta)$$

$$(y(t_{j+1}^m) - y(t_j^m))$$

$$= \lim_{m\to\infty} \int_C p(x,i\Delta,y,i\Delta-\Delta) \sum_{j=1}^{m-1} n(x(t_j^m),t_j^m)(y(t_{j+1}^m) - y(t_j^m))$$

$$W(dx;a,i\Delta).$$

Now the sum: $\sum_{j=1}^{m-1} n(x(t_j^m),t_j^m)(y(t_{j+1}^m) - y(t_j^m))$

converges in quadratic mean W(dy) to:

$\int_{i\Delta-\Delta}^{i\Delta} n(x(t),t)dy(t)$ for each x, hence the convergence is

at least in probability on CXC.

The expectation of the square of the finite sum,

$$\int_C \int_C p(x,i\Delta,y,i\Delta-\Delta) \left[\sum_{j=1}^{m-1} n(x(t_j^m),t_j^m)(y(t_{j+1}^m) - y(t_j^m))\right]^2$$

$$W(dx;a,i\Delta)W(dy)$$

$$= \int_C q(i\Delta,0,x) \sum_{j=1}^{m-1} n^2(x(t_j^m),t_j^m)(t_{j+1}^m - t_j^m)W(dx;a,i\Delta)$$

$$= \sum_{j=1}^{m-1} \int_C q(i\Delta,0,x)n^2(x(t_j^m),t_j^m)W(dx;a,i\Delta)(t_{j+1}^m - t_j^m),$$

converges to: $\int_{i\Delta-\Delta}^{i\Delta} \int_C q(i\Delta,0,x)n^2(x(t),t)W(dx;a,i\Delta)dt,$

so is a bounded sequence of numbers.

By corollary 2 on page 164 of reference 14,

$$\lim_{m\to\infty} \int_C p(x,i\Delta,y,i\Delta-\Delta) \sum_{j=1}^{m-1} n(x(t_j^m),t_j^m)(y(t_{j+1}^m) - y(t_j^m))W(dx;a,i\Delta)$$

$$= \int_C p(x,i\Delta,y,i\Delta-\Delta) \int_{i\Delta-\Delta}^{i\Delta} n(x(t),t)dy(t)W(dx;a,i\Delta),$$

which proves lemma 2.

Lemma 3

$$\sum_{i=1}^{n} \int_C p(x,i\Delta,y,i\Delta-\Delta) \int_{i\Delta-\Delta}^{i\Delta} n(x(t),t)dy(t)W(dx;a,i\Delta)$$

converges in probability to:

$$\int_0^T n(a,t)P(a,t,y,t)dy(t) \quad \text{as} \quad n \to \infty .$$

proof

By lemma 2 each term in the sum can exchange its order of integration, i.e.,

$$3.9 \quad \sum_{i=1}^{n} \int_C p(x,i\Delta,y,i\Delta-\Delta) \int_{i\Delta-\Delta}^{i\Delta} n(x(t),t)dy(t)W(dx;a,i\Delta)$$

$$= \sum_{i=1}^{n} \int_{i\Delta-\Delta}^{i\Delta} \int_C p(x,i\Delta,y,i\Delta-\Delta)n(x(t),t)W(dx;a,i\Delta)dy(t)$$

in probability.

Now substitution of $n(x(t),t) = n(a,i\Delta) - (n(x(t),t) - n(a,i\Delta))$ into 3.9 yields:

$$3.10 \quad \sum_{i=1}^{n} \int_{i\Delta-\Delta}^{i\Delta} \int_C p(x,i\Delta,y,i\Delta-\Delta)n(a,i\Delta)W(dx;a,i\Delta)dy(t)$$
$$+ \sum_{i=1}^{n} \int_{i\Delta-\Delta}^{i\Delta} \int_C p(x,i\Delta,y,i\Delta-\Delta)(n(x(t),t) - n(a,i\Delta))$$
$$W(dx;a,i\Delta)dy(t)$$

The first sum in 3.10,

$$\sum_{i=1}^{n} \int_{i\Delta-\Delta}^{i\Delta} \int_C p(x,i\Delta,y,i\Delta-\Delta)n(a,i\Delta)W(dx;a,i\Delta)dy(t)$$
$$= \sum_{i=1}^{n} \int_{i\Delta-\Delta}^{i\Delta} n(a,i\Delta)P(a,i\Delta,y,i\Delta-\Delta)dy(t)$$

converges to $\int_0^T n(a,t)P(a,t,y,t)dy(t)$ in quadratic mean as $n \to \infty$ by the definition of the stochastic integral.

Ito proves in reference 11 that if the supremum of the absolute value of the integrand goes to zero in probability, then the supremum (over the range of the upper limit) of the stochastic integral also converges to zero in probability. This property will be referred to as "continuity of the stochastic integral."

The supremum of the integrand of the second summation in 3.10 is:

$$\sup_{1\le i\le n} \sup_{i\Delta-\Delta\le t\le i\Delta} \int_C p(x,i\Delta,y,i\Delta-\Delta)(n(x(t),t) - n(a,i\Delta))$$
$$W(dx;a,i\Delta)$$

This sup. must go to zero in probability if it gets small in L_1 norm.

$$\int_C \left| \int_C p(x,i\Delta,y,i\Delta-\Delta)(n(x(t),t) - n(a,i\Delta))W(dx;a,i\Delta) \right| W(dy)$$

$$\leq \int_C \int_C p(x,i\Delta,y,i\Delta-\Delta)|n(a,t) - n(a,i\Delta)|W(dx;a,i\Delta)W(dy)$$

$$+ \int_C \int_C p(x,i\Delta,y,i\Delta-\Delta)|n(x(t),t) - n(a,t)|W(dx;a,i\Delta)W(dy)$$

$$\leq |n(a,t) - n(a,i\Delta)| + K\int_C q(i\Delta,0,x)|x(t) - a|W(dx;a,i\Delta),$$

where use is made of the fact that $n(\cdot,\cdot)$ is uniformly Lipschitz in the first variable. Since $n(a,t)$ is a continuous function of t, it is uniformly continuous on $[0,T]$, so $|n(a,t) - n(a,i\Delta)| \to 0$ as $n \to \infty$.

$$\int_C q(i\Delta,0,x)|x(t) - a|W(dx;a,i\Delta) \quad \text{converges to zero}$$

because $x(t)$ is a continuous process. Thus the second sum in 3.10 converges to zero in probability by the continuity of the stochastic integral. This proves lemma 3.

In reference 11 Ito proves two properties of stochastic integrals that will be used several times in the remaining proofs. These properties are related by integration by parts and are stated here for convenience.

Lemma A

$$\left[\int_0^t f(s)dw(s)\right]\left[\int_0^t g(s)dw(s)\right]$$

$$= \int_0^t f(s)G(s)dw(s) + \int_0^t g(s)F(s)dw(s) + \int_0^t f(s)g(s)ds.$$

Lemma B

$$\left[\int_0^t f(s)dw(s)\right]\left[\int_0^t g(s)ds\right] = \int_0^t f(s)H(s)dw(s) + \int_0^t g(s)F(s)ds,$$

where $F(s) = \int_0^s f(u)dw(u)$, $G(s) = \int_0^s g(u)dw(u)$, $H(s) = \int_0^s g(u)du$.

Lemma 4

$$3.11 \quad \sum_{i=1}^{n} \int_C p(x,i\Delta,y,i\Delta-\Delta)\Big[\int_{i\Delta-\Delta}^{i\Delta} n(x(t),t)dy(t) - \tfrac{1}{2}\int_{i\Delta-\Delta}^{i\Delta} n^2(x(t),t)dt\Big]^2 W(dx;a,i\Delta)$$

$$- \sum_{i=1}^{n} \int_C p(x,i\Delta,y,i\Delta-\Delta)\int_{i\Delta-\Delta}^{i\Delta} n^2(x(t),t)dtW(dx;a,i\Delta)$$

converges to zero in probability as $n \to \infty$.

proof

Let $N(t) = \int_{i\Delta-\Delta}^{t} n(x(s),s)dy(s)$; $N_2(t) = \int_{i\Delta-\Delta}^{i\Delta} n^2(x(s),s)ds$.

By using lemmas A and B, 3.11 can be written:

$$3.12 \quad 2\sum_{i=1}^{n}\int_C p(x,i\Delta,y,i\Delta-\Delta)\int_{i\Delta-\Delta}^{i\Delta} N(t)n(x(t),t)dy(t)W(dx;a,i\Delta)$$
$$- \sum_{i=1}^{n}\int_C p(x,i\Delta,y,i\Delta-\Delta)\int_{i\Delta-\Delta}^{i\Delta} N_2(t)n(x(t),t)dy(t)W(dx;a,i\Delta)$$
$$- \sum_{i=1}^{n}\int_C p(x,i\Delta,y,i\Delta-\Delta)\int_{i\Delta-\Delta}^{i\Delta} N(t)n^2(x(t),t)dtW(dx;a,i\Delta)$$
$$+ \tfrac{1}{4}\sum_{i=1}^{n}\int_C p(x,i\Delta,y,i\Delta-\Delta)\Big[\int_{i\Delta-\Delta}^{i\Delta} n^2(x(t),t)dt\Big]^2 W(dx;a,i\Delta)$$

By exchanging the order of integration (as in lemma 2) the first sum in 3.12 can be written:

$$3.13 \quad 2\sum_{i=1}^{n}\int_{i\Delta-\Delta}^{i\Delta}\int_C p(x,i\Delta,y,i\Delta-\Delta)n(x(t),t)N(t)W(dx;a,i\Delta)dy(t)$$

The Chebyshev inequality applied to the integrand of 3.13 yields:

$$\text{prob}\Big\{\Big|\int_C p(x,i\Delta,y,i\Delta-\Delta)n(x(t),t)N(t)W(dx;a,i\Delta)\Big| > \delta\Big\}$$

$$\leq \frac{1}{\delta}\int_C \int_C p(x,i\Delta,y,i\Delta-\Delta)|n(x(t),t)N(t)|W(dx;a,i\Delta)W(dy),$$

which is bounded (Schwarz's inequality) by:

3.14 $$\frac{1}{\delta}\int_C \int_C p(x,i\Delta,y,i\Delta-\Delta)W(dx;a,i\Delta)W(dy)$$
$$\cdot \int_C \int_C p(x,i\Delta,y,i\Delta-\Delta)\left[n(x(t),t)N(t)\right]^2 W(dx;a,i\Delta)W(dy)$$

The first integral in 3.14 has the value one. The second is:

$$\int_C q(i\Delta,0,x)n^2(x(t),t)\int_{i\Delta-\Delta}^{t} n^2(x(s),s)dsW(dx;a,i\Delta),$$

which is bounded by:

3.15 $$\sup_{i\Delta-\Delta\le t\le i\Delta}\Delta\int_C q(i\Delta,0,x)n^4(x(t),t)W(dx;a,i\Delta).$$

But 3.15 converges to zero as $n\to\infty$.

Thus the supremum of the integrand of equation 3.13 converges to zero in probability; therefore 3.13 (the first sum in 3.12) goes to zero in probability by the continuity of the stochastic integral.

The second, third, and fourth sums in equation 3.12 also converge to zero in probability. The arguements for this fact are of course not <u>exactly</u> the same as those for the first sum, but similar enough (in most cases easier) to leave the details to the reader. This proves lemma 4.

<u>Lemma 5</u>

$$\sum_{i=1}^{n}\int_C p(x,i\Delta,y,i\Delta-\Delta)\left[\left(\int_{i\Delta-\Delta}^{i\Delta} n(x(t),t)dy(t)\right.\right.$$
$$\left.\left.-\tfrac{1}{2}\int_{i\Delta-\Delta}^{i\Delta} n^2(x(t),t)dt\right]^3 W(dx;a,i\Delta)$$ converges to zero in probability as $n\to\infty$.

<u>proof</u>

Expansion of the cubic in the integrand yields:

$$3.16 \quad \sum_{i=1}^{n} \int_C p(x,i\Delta,y,i\Delta-\Delta)\left[\int_{i\Delta-\Delta}^{i\Delta} n(x(t),t)dy(t)\right]^3 W(dx;a,i\Delta)$$

$$- \frac{3}{2}\sum_{i=1}^{n} \int_C p(x,i\Delta,y,i\Delta-\Delta)\left[\int_{i\Delta-\Delta}^{i\Delta} n(x(t),t)dy(t)\right]^2 \left[\int_{i\Delta-\Delta}^{i\Delta} n^2(x(t),t)dt\right]$$

$$W(dx;a,i\Delta)$$

$$- \frac{3}{4}\sum_{i=1}^{n} \int_C p(x,i\Delta,y,i\Delta-\Delta)\left[\int_{i\Delta-\Delta}^{i\Delta} n(x(t),t)dy(t)\right]\left[\int_{i\Delta-\Delta}^{i\Delta} n^2(x(t),t)dt\right]^2$$

$$W(dx;a,i\Delta)$$

$$- \frac{1}{8}\sum_{i=1}^{n} \int_C p(x,i\Delta,y,i\Delta-\Delta)\left[\int_{i\Delta-\Delta}^{i\Delta} n^2(x(t),t)dt\right]^3 W(dx;a,i\Delta)$$

The first sum contains a stochastic integral to the third power. A straightforward application of lemma A results in:

$$3.17 \quad \left[\int_{i\Delta-\Delta}^{i\Delta} n(x(t),t)dy(t)\right]^3 = \int_{i\Delta-\Delta}^{i\Delta} n(x(t),t)dy(t)$$

$$\cdot\left[\int_{i\Delta-\Delta}^{i\Delta} n^2(x(t),t)dt + 2\int_{i\Delta-\Delta}^{i\Delta} N(t)n(x(t),t)dy(t)\right]$$

Substitution of 3.17 into 3.16 makes arguements similar to those of lemma 4 applicable. Thus each sum in 3.16 can be shown to converge to zero in probability. This proves lemma 5.

Lemma 6

$$3.18 \quad \sum_{i=1}^{n} \int_C p(x,i\Delta,y,i\Delta)\left[\int_{i\Delta-\Delta}^{i\Delta} n(x(t),t)dy(t)\right.$$

$$\left. - \tfrac{1}{2}\int_{i\Delta-\Delta}^{i\Delta} n^2(x(t),t)dt\right]^4 W(dx;a,i\Delta) \quad \text{and}$$

$$3.19 \quad \sum_{i=1}^{n} \int_C p(x,i\Delta,y,i\Delta-\Delta)\left[\int_{i\Delta-\Delta}^{i\Delta} n(x(t),t)dy(t)\right.$$

$$\left. - \tfrac{1}{2}\int_{i\Delta-\Delta}^{i\Delta} n^2(x(t),t)dt\right]^4 W(dx;a,i\Delta)$$

both converge to zero as $n \to \infty$.

proof

Equation 3.18 requires special handling because it is the first time that the two parts of the integral are not independent functionals of y. That is to say, $p(\cdot,\cdot,y,i\Delta)$ and $\int_{i\Delta-\Delta}^{i\Delta}(\cdot,\cdot)dy(t)$ are both measurable $\alpha_{i\Delta}$. Convergence will be proven by way of transformation of variables, i.e. the expectation with respect to Wiener measure of 3.18,

$\int_C \int_C p(x,i\Delta,y,i\Delta)J^4W(dx;a,i\Delta)W(dy)$ can be written:

3.19 $$\int_C \int_C J^4W^R(dx;a,i\Delta)W^R(dy),$$

where $W^R(dx;a,i\Delta)$ and $W^R(dy)$ are the measures on C induced by solving:

$$x(t) = \int_0^t m(x(s),s)ds + \xi(t) \quad \text{and}$$

$$y(t) = \int_0^t n(x(s),s)ds + \zeta(t).$$

Now $W^R(dy)$ can be replaced by $W(d\zeta)$ if y(t) is replaced with : $\int_0^t n(x(s),s)ds + \zeta(t)$ in J^4. When this is done, equation 3.19 becomes:

$$\sum_{i=1}^{n} \int_C \int_C \left[\tfrac{1}{2}\int_{i\Delta-\Delta}^{i\Delta} n^2(x(t),t)dt + \int_{i\Delta-\Delta}^{i\Delta} n(x(t),t)d\zeta(t)\right]^4 W^R(dx;a,i\Delta)W(d\zeta)$$

$$= \sum_{i=1}^{n} \int_C \int_C q(i\Delta,0,x)\left[\tfrac{1}{2}\int_{i\Delta-\Delta}^{i\Delta} n^2(x(t),t)dt + \int_{i\Delta-\Delta}^{i\Delta} n(x(t),t)dy(t)\right]^4 W(dx;a,i\Delta)W(dy),$$

where the $q(i\Delta,0,x)$ is the Radon-Nikodym derivative that relates $W^R(dx;a,i\Delta)$ to $W(dx;a,i\Delta)$ and the dummy variable ? has been changed to y. Expansion of the fourth power gives:

$$3.21 \quad \sum_{i=1}^{n}\int_C q(i\Delta,0,x)\Big[\tfrac{1}{2}\int_{i\Delta-\Delta}^{i\Delta} n^2(x(t),t)dt\Big]^4 W(dx;a,i\Delta)$$
$$+ \tfrac{3}{2}\sum_{i=1}^{n}\int_C q(i\Delta,0,x)\Big[\int_{i\Delta-\Delta}^{i\Delta} n^2(x(t),t)dt\Big]^3 W(dx;a,i\Delta)$$
$$+ \sum_{i=1}^{n}\int_C\int_C q(i\Delta,0,x)\Big[\int_{i\Delta-\Delta}^{i\Delta} n(x(t),t)dy(t)\Big]^4 W(dx;a,i\Delta)W(dy)$$

The first two sums clearly vanish as $n\to\infty$. The third sum can be bounded by invoking a result due to Skorokhod (Ref.15).

$$\sum_{i=1}^{n}\int_C\int_C q(i\Delta,0,x)\Big[\int_{i\Delta-\Delta}^{i\Delta} n(x(t),t)dy(t)\Big]^4 W(dx;a,i\Delta)W(dy)$$
$$\leq 36\Delta\sum_{i=1}^{n}\int_C\Big[\int_{i\Delta-\Delta}^{i\Delta} n^4(x(t),t)dt\Big]q(i\Delta,0,x)W(dx;a,i\Delta),$$

which obviously goes to zero as $n\to\infty$.

What has been shown is that the expectation of 3.18 converges to zero; since it is non-negative it must go to zero in probability as well.

Taking the expectation of 3.19 results in:

$$\sum_{i=1}^{n}\int_C q(i\Delta,0,x)\Big[\int_{i\Delta-\Delta}^{i\Delta} n(x(t),t)dy(t) - \tfrac{1}{2}\int_{i\Delta-\Delta}^{i\Delta} n^2(x(t),t)dt\Big]^4$$
$$W(dx;a,i\Delta)W(dy),$$

which expands to exactly 3.21. Hence the same arguements apply and 3.19 also goes to zero in probability. This proves lemma 6.

CHAPTER IV

BEST DYNAMIC ESTIMATOR

The verb "to track" carries the connotation of continuous modification in an attempt to maintain a fixed relationship with respect to a quantity which itself varies with time. This property is the central theme of this chapter. The approach is best explained by way of the following model:

4.1 $$dx(t) = m(x(t),t)dt - \sigma(x(t),t)d\xi(t)$$

is a nonlinear time varying state equation driven by the noise process $d\xi(t)$. The observation equation has additional noise $d\eta(t)$.

4.2 $$dy(t) = n(x(t),t)dt - d\eta(t).$$

The estimating scheme or filter is limited to a single state device so changes in the estimate can depend only on the new information and the estimate itself.

4.3 $$dz(t) = g(z(t),t)dt - f(z(t),t)dy(t)$$

If equation 4.3 is rewritten:

4.3' $$z(t+\Delta) = z(t) + \int_t^{t+\Delta} g(z(s),s)ds + \int_t^{t+\Delta} f(z(s),s)dy(s),$$

it is clear that while the requirement that an estimator have the form of equation 4.3 is a restriction, it in fact represents an analytic specification of the concept of tracking or "updating" the estimate. The second motivation for equation 4.3 is obvious if 4.2 and 4.3 are combined.

4.4 $$dz(t) = [g(z(t),t) + f(z(t),t)n(x(t),t)]dt + f(z(t),t)d\eta(t)$$

With the requirement that $\xi(t)$ and $\eta(t)$ be Brownian motions, equations 4.1 and 4.4 constitute a two dimensional diffusion process and 4.3' contains a stochastic integral.

Under mild conditions on $f(\cdot,\cdot)$ and $g(\cdot,\cdot)$, the joint density,

$$P(a,b,t) = \frac{\partial}{\partial a}\frac{\partial}{\partial b}\{\text{probability } x(t) < a \text{ and } z(t) < b\},$$

satisfies the forward Kolmogorov equation.

$$4.5 \qquad \frac{\partial}{\partial t}P(a,b,t) = -\frac{\partial}{\partial a}(m(a,t)P(a,b,t)) - \frac{\partial}{\partial b}\left[(f(b,t)n(a,t) - g(b,t))P(a,b,t)\right] + \tfrac{1}{2}\frac{\partial^2}{\partial a^2}(\sigma^2(a,t)P(a,b,t)) + \tfrac{1}{2}\frac{\partial^2}{\partial b^2}(f^2(b,t)P(a,b,t))$$

The boundary value for 4.5, $P(a,b,t)$, is the initial joint density of the state $x(0)$ and the estimate of the state $z(0)$ and will be assumed known throughout the following.

The optimization problem may be expressed precisely now. For instance if it is desired to minimize the mean square error at time T, one selects the pair of functions $f(\cdot,\cdot)$ and $g(\cdot,\cdot)$ on $(-\infty,\infty)X[0,T]$ such that the solution to 4.5 has the property:

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(a - b)^2P(a,b,T)dadb$$

is minimized. Alternatively $f(\cdot,\cdot)$ and $g(\cdot,\cdot)$ may be chosen to minimize either:

$$\frac{1}{T}\int_0^T\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(a - b)^2P(a,b,t)dadbdt, \quad \text{or}$$

$$\max_{0\leq t\leq T} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(a - b)^2 P(a,b,t)dadb.$$

The above are variational calculus problems of terrifying complexity. The first example, mean square error at time T, of course heaps all the cost function on one instant of time. Thus there may be sub-intervals of [0,T] for which z(t) bears little resemblance to x(t) and the error is large. The same criticism may be made of the other two criteria. An estimator above reproach is one that minimizes the error uniformly over the interval [0,T]. Unfortunately such a filter does not in general exist. An optimality criterion that retains the spirit of tracking or updating of the estimator and at the same time makes the problem more tractable than the above will be proposed here.

An estimator will be said to be "sequentially best" if every other estimator that has a smaller error at any time t has a larger error over some interval $(t_1,t_2)\subseteq[0,t]$. Clearly if an estimator is sequentially best there can not be another estimator which is uniformly better. Thus if a uniformly best estimator exists, it must be the sequentially best.

The criterion is best illustrated with a discrete process. Since a quantized version of the model is needed in the proof of the main result anyway, the corresponding Markov chains will be defined here. Keep in mind that besides being the canonical approximation they do indeed converge to the continuous analog.

$$x^n_{i+1} = x^n_i + m(x^n_i,t^n_i)\Delta t^n + \sigma(x^n_i,t^n_i)\Delta\xi^n_i$$

$$y^n_{i+1} = y^n_i + n(x^n_i,t^n_i)\Delta t^n + \eta^n_i$$

$$z^n_{i+1} = z^n_i + g(z^n_i,t^n_i)\Delta t^n + f(z^n_i,t^n_i)(y^n_{i+1} - y^n_i)$$

$$= z^n_i + [f(z^n_i;t^n_i)n(x^n_i,t^n_i) + g(z^n_i,t^n_i)]\Delta t^n + f(z^n_i,t^n_i)\Delta\eta^n_i$$

The superscript n denotes the partition.

$T_n = \{0 = t^n_0, t^n_1, t^n_2, \cdots, t^n_n = T\}$ and $\max_{0\le k\le n-1} |t^n_{k+1} - t^n_k| \to 0$ as $n \to \infty$. Let $(t^n_{k+1} - t^n_k) = \Delta t^n$ for all k (for brevity only) and let $\Delta\xi^n_i = \xi(t^n_{i-1}) - \xi(t^n_i)$ and $\Delta\eta^n_i = \eta(t^n_{i-1}) - \eta(t^n_i)$.

Since $\Delta\xi^n_i$ and $\Delta\eta^n_i$ are increments of Brownian motion, the transition densities are:

$$P(x^n_{i+1}|x^n_i,z^n_i) = (2\pi\sigma^2(x^n_i,t^n_i)\Delta t^n)^{-\frac{1}{2}} \cdot \exp\left\{-\frac{(x^n_{i+1} - x^n_i - m(x^n_i,t^n_i)\Delta t^n)^2}{2\sigma^2(x^n_i,t^n_i)\Delta t^n}\right\}$$

$$P(z^n_{i+1}|x^n_i,z^n_i) = (2\pi f^2(z^n_i,t^n_i)\Delta t^n)^{-\frac{1}{2}} \cdot \exp\left\{-\frac{(z^n_{i+1} - z^n_i - [f(z^n_i,t^n_i)n(x^n_i,t^n_i) - g(z^n_i,t^n_i)]\Delta t^n)^2}{2f^2(z^n_i,t^n_i)\Delta t^n}\right\}.$$

Clearly $P(x^n_2,z^n_2,t^n_2)$ is affected by the choice of $f(z^n_0,t^n_0)$ and $g(z^n_0,t^n_0)$ as well as $f(z^n_1,t^n_1)$ and $g(z^n_1,t^n_1)$. The $f(z^n_0,t^n_0)$ and $g(z^n_0,t^n_0)$ which minimize the above integral will not in general be the same as those which minimize:

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(z_1^n - x_1^n)^2 P(x_1^n, z_1^n)dx_1^n dz_1^n,$$

the error at time t_1^n. To choose the better tracker or the sequentially better (discrete case) filter is to choose the latter, in spite of the fact that this may force a large error at time t_2^n. In the limit as $n \to \infty$ the interpretation is, roughly speaking, having chosen $f(\cdot,\cdot)$ and $g(\cdot,\cdot)$ on $(-\infty,\infty)X[0,t)$, what functions on $(-\infty,\infty)X[t,t+dt)$ will minimize the error at time $(t+dt)$.

It will be shown in this chapter that in the discrete case a unique sequentially best estimator always exists and that the filter can be found by solving a certain set of equations. The form of the equations is invariant under refinements of the partitions of the time interval, so the sequentially best in the continuous case must also satisfy these relations. This is not an assertion of the existence or uniqueness of a continuous sequentially best filter, rather an algorithm for finding it when one does exist.

With the sequential or step by step criterion the complexion of the problem has actually been changed. Instead of searching the space of all $f(\cdot,\cdot)$'s and $g(\cdot,\cdot)$'s that may be plugged into 4.5 to attain some global property, the functions can be ground out beginning on the left of $[0,T]$.

Lemma 1

A sequentially best estimator has the property:

$$E(x(t)|z(t)) = z(t).$$

proof

Suppose $f(\cdot,\cdot)$ and $g(\cdot,\cdot)$ are specified on $(-\infty,\infty)X[0,T]$. The output of the filter $z(t)$ may be treated as a statistic as indeed it is. To operate in $z(t)$ to minimize $E(x(t) - h(t))^2$, where $h(t) = H(z(t),t)$ means that $h(t) = E(x(t)|z(t))$ would be the new estimator for $x(t)$. But:

$$h(t) = \int_{-\infty}^{\infty} aP(a,b,t)da \;/\; \int_{-\infty}^{\infty} P(a,b,t)da,$$

where $P(a,b,t)$ is the solution to 4.5. If $P(a,b,t)$ satisfies the smoothness conditions necessary for equation 4.5 to hold, then $h(t)$ can be differentiated according to the Ito differential rule, i.e.,

$$dh(t) = G(h(t),t)dt + F(h(t),t)dz(t),$$

for some functions $F(\cdot,\cdot)$ and $G(\cdot,\cdot)$. Thus $h(\cdot)$ is of the form of equation 4.3 and has an error which is never greater than that of $z(t)$, a contradiction unless $h(t) = z(t)$ with probability one. This proves lemma 1.

Lemma 1 shows $E(x(t)|z(t)) = z(t)$ which is certainly stronger than unbiasedness, $E(x(t)) = E(z(t))$. Thus to restrict attention to the class of unbiased estimators does not exclude extremal points where the mean square error is involved. The effect of the unbiasedness condition will be developed first.

Lemma 2

The sequentially best (discrete case) estimator has the property:

$$g(z_i^n,t_i^n) = \bar{m}(z_i^n) - f(z_i^n,t_i^n)\bar{n}(z_i^n), \text{ where}$$

$$\bar{m}(z_i^n) = \int_{-\infty}^{\infty} m(x_i^n,t_i^n)P(x_i^n|z_i^n)dx_i^n \;\&\; \bar{n}(z_i^n) = \int_{-\infty}^{\infty} n(x_i^n,t_i^n)P(x_i^n|z_i^n)dx_i^n.$$

proof

By lemma 1, $E(x_{i+1}^n|z_{i+1}^n) = \int_{-\infty}^{\infty} x_{i+1}^n P(x_{i+1}^n|z_{i+1}^n)dx_{i+1}^n = z_{i+1}^n$.

Taking the expectation of both sides and using:

$$P(x_{i+1}^n,z_{i+1}^n) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} P(x_{i+1}^n,z_{i+1}^n|x_i^n,z_i^n)P(x_i^n,z_i^n)dx_i^n dz_i^n, \text{ yields:}$$

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x_i^n - z_i^n)P(x_i^n,z_i^n)dx_i^n dz_i^n$$

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\left[m(x_i^n,t_i^n)\Delta t^n - (f(z_i^n,t_i^n)n(x_i^n,t_i^n) + g(z_i^n,t_i^n))\Delta t^n\right]$$

$$\cdot P(x_i^n,z_i^n)dx_i^n dz_i^n, \text{ which is zero if } z_i^n \text{ is unbiased.}$$

Replacing $P(x_i^n,z_i^n)$ by $P(x_i^n|z_i^n)P(z_i^n)$ and carrying out the x_i^n integration yields: $\int_{-\infty}^{\infty}\left[f(z_i^n,t_i^n)\bar{n}(z_i^n) + g(z_i^n,t_i^n) - \bar{m}(z_i^n)\right]P(z_i^n)dz_i^n$ $= 0$. Now $P(z_i^n)$ may assume a range of values while keeping $\bar{m}(z_i^n)$ and $\bar{n}(z_i^n)$ constant, thus the integrand must vanish pointwise or: $g(z_i^n,t_i^n) = \bar{m}(z_i^n) - f(z_i^n,t_i^n)\bar{n}(z_i^n)$.

Hence lemma 2 is true by induction.

It is interesting to note that if one ignores bias and tries to reduce error by manipulating $g(\cdot,\cdot)$ and $f(\cdot,\cdot)$ independently, the result is that $g(z_i^n,t_i^n)$ comes out to be proportional to $(\Delta t^n)^{-1}$. In the limit, as $\Delta t^n \to 0$, $g(\cdot,\cdot) = \pm\infty$, an unacceptable answer but not a surprising one. What the

mathematics is saying is: If the new data causes one to decide the estimate is low, the drift coefficient $g(\cdot,\cdot)$ should be assigned the value that will raise the expectation of the new estimate the fastest. Of course there is no fastest and the derivation leads to a nonsense answer. It is analoguous to control problems with a bounded set of controls where the answer is "bang-bang" or one of two extremal points is always optimal.

Lemma 3

The sequentially best (discrete time) estimator has the property: $f(z_i^n,t_i^n) = \overline{nx}(z_i^n) - \bar{n}(z_i^n)\bar{x}(z_i^n)$, where

$$\overline{nx}(z_i^n) = \int_{-\infty}^{\infty} n(x_i^n,t_i^n)x_i^n P(x_i^n|z_i^n)dx_i^n \quad \& \quad \bar{x}(z_i^n) = \int_{-\infty}^{\infty} x_i^n P(x_i^n|z_i^n)dx_i^n.$$

proof

Minimizing $E(z_{i+1}^n - x_{i+1}^n)^2$ is equivalent to minimizing:

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} P(x_{i+1}^n,z_{i+1}^n)(-2x_{i+1}^n z_{i+1}^n - (z_{i+1}^n)^2)dx_{i+1}^n dz_{i+1}^n.$$

An expansion similar to that used in lemma 2 and the result of lemma 2 yields:

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} P(x_i^n,z_i^n)\big[f^2(z_i^n,t_i^n)\Delta t^n + (z_i^n + (f(z_i^n,t_i^n)\cdot n(x_i^n,t_i^n) + g(z_i^n,t_i^n))\Delta t^n)^2 - 2(x_i^n + m(x_i^n,t_i^n)\Delta t^n)(z_i^n - (f(z_i^n,t_i^n)n(x_i^n,t_i^n) + g(z_i^n,t_i^n))\Delta t^n\big]dx_i^n dz_i^n$$

With a regrouping of terms the expression to be minimized becomes:

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} P(x_i^n, z_i^n)(-2x_i^n z_i^n + (z_i^n)^2)dx_i^n dz_i^n +$$

$$\Delta t^n \int_{-\infty}^{\infty} P(z_i^n)\left[f^2(z_i^n, t_i^n) - 2\overline{nx}(z_i^n)f(z_i^n, t_i^n) + 2\bar{x}(z_i^n)\bar{n}(z_i^n)f(z_i^n, t_i^n)\right] dz_i^n$$

$$+ \Delta t^n \int_{-\infty}^{\infty} P(z_i^n)(2z_i^n \bar{m}(z_i^n) - 2\bar{x}(z_i^n)\bar{m}(z_i^n))dz_i^n + o(\Delta t^n).$$

By lemma 1 the last integral is zero.

Now it is assumed by the sequentially best criterion that the estimator, that is $f(\cdot,\cdot)$ and $g(\cdot,\cdot)$, have been chosen up to t_{i-1}^n, so the distributions at t_i^n are fixed. The second integral is minimized by minimizing the integrand at every point. This is easily seen to occur by setting:

$$f(z_i^n, t_i^n) = \overline{nx}(z_i^n) - \bar{n}(z_i^n)\bar{x}(z_i^n).$$

Thus lemma 3 is true by induction.

Substitution of the minimizing $f(\cdot,\cdot)$ back into the expression of the error at time t_{i+1}^n,

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} P(x_i^n, z_i^n)(-2x_i^n z_i^n + (z_i^n)^2)dx_i^n dz_i^n$$

$$-\Delta t^n \int_{-\infty}^{\infty} P(z_i^n)(\overline{nx}(z_i^n) - \bar{n}(z_i^n)\bar{x}(z_i^n))^2 dz_i^n,$$

shows how the error at time t_{i+1}^n equals the error at time t_i^n plus a small multiple of a complicated moment expression. If two estimators have the same error at time t_i^n, $n(x_i^n, t_i^n)$ can be chosen in such a way that either may have a smaller error at time t_{i+1}^n, further illustrating the non-existance of a dynamic estimator with uniformly smallest error. In the linear case the above reduces to a function of the error at time t_i^n only, thus simultaneous minimization is possible.

Theorem

If a sequentially best estimator exists, it must simultaneously satisfy equations 4.5,

4.6 $$g(z(t),t) = \bar{m}(z(t)) - \bar{f}(z(t),t)\bar{n}(z(t)), \text{ and}$$

4.7 $$f(z(t),t) = \overline{nx}(z(t)) - \bar{n}(z(t))\bar{x}(z(t)),$$

where in general, $\bar{r}(z(t)) = \int_{-\infty}^{\infty} r(x(t),t)P(x(t)|z(t))dx(t)$.

proof

The theorem is almost obvious in light of lemmas 2 and 3. If two estimators agree up to time t (they necessarily agree at $t = 0$) and then differ over an interval $(t,t+\Delta)$, there must be a discrete approximation such that $t_i^n = t$ and $t_{i+1}^n \epsilon (t,t+\Delta)$. Since the algorithm for the discrete case defines $f(\cdot,\cdot)$ and $g(\cdot,\cdot)$ uniquely up to sets of measure zero, the filter that does not satisfy 4.5 and 4.6 (the same as not satisfying lemmas 2 and 3) must have a larger error and is therefore not sequentially best. This proves the theorem.

To prove that a unique sequentially best estimator exists is to prove that the simultaneous solution of equations 4.5, 4.6, and 4.7 exists and is unique. Any restrictions on the problem will necessarily be expressed in terms of $m(\cdot,\cdot)$ and $n(\cdot,\cdot)$ and $\sigma(\cdot,\cdot)$, the only unspecified quantities. Although suitable conditions on these coefficients to insure the existence of a unique set of solutions $f(\cdot,\cdot)$ and $g(\cdot,\cdot)$ to 4.5, 4.6, and 4.7 have not been discovered at this time, examples of the next chapter, especially the linear case,

show that the preceding theorem can not be vacuous. Indeed, it is suspected that it is rather widely applicable. A rigorous proof of existence of a solution would be truly interesting, but will not be persued here.

In the class of estimators of the form of equation 4.3, i.e. dynamic estimators, there are in general many extremal points. The sequentially best is one. The filter that minimizes the error at time T is another, etc.. In the linear case it is possible to minimize error simultaneously over $[0,T]$, so there is only one extremal point. If, one begins with the sequentially best estimator, he has a toe hold on the nonlinear problem. Since it is known that the dynamics of the conditional expectation can be expressed in the form of equation 4.3 if infinite dimensional variables are allowed, two things must be true. One is that increasing the dimensionality of the filter should provide better and better results. Another is that if one has an n dimensional filter and changes the functions in the state and observation equations to linear ones, an "uncoupling" must take place with the dimensions greater than one becoming extraneous.

An uninteresting example is the following. Construct filter 1 by solving equation 4.5 for $f(\cdot,\cdot)$ and $g(\cdot,\cdot)$ on $(-\infty,\infty)X[0,T+\Delta]$ such that $E(z(T+\Delta) - x(T+\Delta))^2$ is minimum. Do the same for filter 2 where $\Delta = 2\Delta$, etc.. The estimator at time t will be the output of filter k when $k\Delta \leq t < (k-1)\Delta$.

The above is uninteresting because not only is building so many separate filters not practical, but the solution of 4.5 for the optimum $f(\cdot,\cdot)$ and $g(\cdot,\cdot)$ is the terribly difficult problem that was to be avoided in the first place. What justifies the discussion is how the uncoupling takes place. The coefficients of filter k coincide exactly with those of filter k + n on $[0,T+k\Delta]$ in the linear case. Thus if the higher dimensions are involved by switching between many different filters, linearity allows the replacement of all by the single filter designed for the longest time interval involved.

Now the sequentially best filter is a functional on $y(\cdot)$, the observation curve, and as such should have an embedding in a higher dimensional estimator. Exactly how this can be achieved is an open question; one that looks promising for further exploration. Only a heuristic discussion is possible at this time.

Consider the incremental behavior of the estimator, where the drift coefficient has been chosen for unbiasedness,

$$z_{i+1}^n = z_i^n + \bar{m}(z_i^n)\Delta t^n + f(z_i^n,t_i^n)[\Delta y_i^n - \bar{n}(z_i^n)\Delta t^n].$$

The new estimate z_{i+1}^n depends on the old z_i^n in three ways. First of course it is proportional to z_i^n because it is on updating of z_i^n. Secondly $\bar{m}(z_i^n)\Delta t^n$ is the expected change in $x(\cdot)$ between t_i^n and t_{i+1}^n, an estimate of the change of state x_i^n based on statistic z_i^n. Thirdly $\bar{n}(z_i^n)$ makes $[\Delta y_i^n - \bar{n}(z_i^n)\Delta t^n]$ into a statistic with zero mean, a fraction

of which is to be added to $z_i^n + \bar{m}(z_i^n)\Delta t^n$, the informationless estimate (predictor) of x_{i+1}^n. In the sequential construction to this chapter it can be seen that the fraction is a reflection of the confidence one has in the estimate, i.e.

$$f(z_i^n, t_i^n) = \overline{nx}(z_i^n) - \bar{n}(z_i^n)\bar{x}(z_i^n)$$

vanishes as the variance of the conditional distribution of x_i^n given z_i^n goes to zero.

Consider the following two dimensional improvement over the estimation scheme proposed in this chapter.

$$dz(t) = g(z(t),z'(t),t) + f(z(t),z'(t),t)dy(t)$$

$$dz'(t) = g'(z(t),z'(t),t) + f'(z(t),z'(t),t)dy(t)$$

$z(t)$ is the estimate of $x(t)$, so $z'(t)$ has the role of supplying supplemental information that isn't needed when a uniformly best exists. By making $z(t)$ a good estimate of $x(t)$, the second variable $z'(t)$ must allow an improvement in $\bar{m}(z(t))$ and/or $\bar{n}(z(t))$. Perhaps setting $z'(t) = E(x^2(t)|z(t),z'(t))$ and letting $\bar{m}(z(t),z'(t)) = E(m(x(t),t)|z(t),z'(t))$ is one way of making a better updating of the estimate. Simultaneously $\bar{n}(z(t),z'(t)) = E(n(x(t),t)|z(t),z'(t))$ must represent an improvement in the extraction of information from dy. It is obvious that all the trouble is caused by the fact that the best estimate of $m(x(t),t)$ is not $m(z(t),t)$ where $z(t)$ is the best estimate of $x(t)$. It <u>is</u> the best in the particular case of $m(x(t),t) = m(t)x(t)$, the linear case. Thus the uncoupling. Having made the best estimate of $x(t)$, one automatically has the best estimate of $m(x(t),t) = m(t)x(t)$ and

$n(x(t),t) = n(t)x(t)$. The $z'(t)$ can then offer no improvement and falls away.

The above discussion is by no means rigorous and certainly not the only reasoning by which the sequentially best estimator may be made more accurate. It is offered rather so the reader may interpret the results of this chapter with respect to the over all problem and perhaps motivate others to persue improvements to the sequentially best estimation scheme.

CHAPTER V

EXAMPLES

Frequent reference has been made to the linear case and the Kalman filter. It is interesting to examine exactly how the algorithm reduces in this instance.

Example 1

First, of course, $\sigma(x(t),t)$, $m(x(t),t)$, and $n(x(t),t)$ must be linear; let $\sigma(x(t),t) = \sigma(t)$, $m(x(t),t) = m(t)x(t)$, and $n(x(t),t) = n(t)x(t)$. By lemma 1,

$$g(z(t),t) = E(m(t)x(t)|z(t)) - f(z(t),t)E(n(t)x(t)|z(t))$$
$$= m(t)z(t) - f(z(t),t)n(t)z(t).$$

now all the coefficients of 4.5 are specified with the exception of $f(z(t),t)$.

$$f(z(t) = E(n(t)x(t)x(t)|z(t)) - E(n(t)x(t)|z(t))E(x(t)|z(t))$$
$$= n(t)E(x^2(t) - (E(x(t)|z(t))^2|z(t)),$$

i.e. $f(z(t),t)$ is the conditional variance of $x(t)$.

Now consider a Gaussian solution of 4.5. When $P(a,b,t)$ is a bivariate normal distribution, the conditional variance is not a function of the conditioning variable, i.e. $f(z(t),t) = f(t)$. Hence all the coefficients of 4.5 are linear and if a solution exists it must be Gaussian.

If the only unknown is $f(t)$, which in the linear case has the property of being the mean square error at time t, is it still necessary to solve 4.5? The answer, thanks to Kalman and Bucy, is no. A major contribution of their work

(Ref. 8) is the proof that the evolution of the mean square error is a Riccati equation. Thus the reduction is complete. The filter, f(t) and g(z(t),t) = m(t)z(t) - f(t)n(t)z(t) = g(t)z(t), do indeed coincide exactly with the Kalman filter.

A numerical solution to the simultaneous set of equations:

$$5.1\quad \frac{\partial}{\partial t}P(a,b,t) = -\frac{\partial}{\partial a}(m(a,t)P(a,b,t)) - \frac{\partial}{\partial b}\left[(f(b,t)n(a,t) + g(b,t))P(a,b,t)\right] + \frac{1}{2}\frac{\partial^2}{\partial a^2}(\sigma^2(a,t)P(a,b,t)) + \frac{1}{2}\frac{\partial^2}{\partial b^2}(f^2(b,t)P(a,b,t)),$$

$$5.2\quad g(b,t) = \frac{1}{P(b,t)}\int_{-\infty}^{\infty} m(a,t)P(a,b,t)da - f(b,t)\frac{1}{P(b,t)}\int_{-\infty}^{\infty} n(a,t)P(a,b,t)da,$$

$$5.3\quad f(b,t) = \frac{1}{P(b,t)}\int_{-\infty}^{\infty} an(a,t)P(a,b,t)da - \left(\frac{1}{P(b,t)}\right)^2\int_{-\infty}^{\infty} aP(a,b,t)da \int_{-\infty}^{\infty} n(a,t)P(a,b,t)da,$$

where $P(a,b,t) = \int_{-\infty}^{\infty} P(a,b,t)da$, was attempted in lieu of an analytic example for the algorithm for the sequentially best estimator in a nonlinear situation.

The programming always followed the logic depicted in figure 1. First a two dimensional density function was placed on a square grid; P(a,b,0) is the initial joint density of the state x(0) and estimate z(0). The coefficients and their derivatives which appear in 5.1 were calculated. The calculations of equations 5.2 and 5.3 were carried out for each b coordinate and a sampling was printed out. A

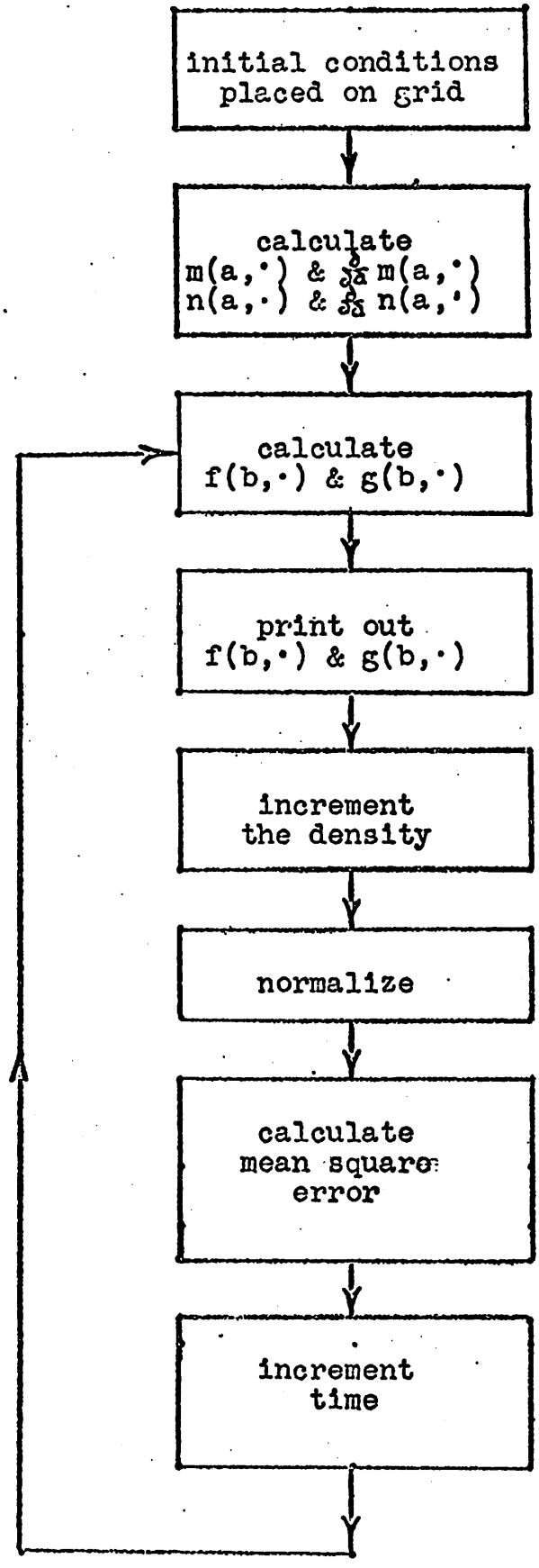

initial conditions placed on grid
calculate m(a,·) & δ_a m(a,·) n(a,·) & δ_a n(a,·)
calculate f(b,·) & g(b,·)
print out f(b,·) & g(b,·)
increment the density
normalize
calculate mean square error
increment time

figure 1

difference equation analog of equation 5.1 was used to increment the density at each point of the grid. The density is normalized and the mean square error $E(x(t) - z(t))^2$ is calculated and printed out. The time variable is increased and the loop closed by going back to calculate the new coefficients of the sequentially best filter.

Some simplifications are already apparent. The coefficients of the state and observation equations were taken to be time invariant and so their calculation is outside the loop. Furthermore, $\sigma(x(t),t)$ was always taken to be a constant.

The major problem with the scheme described in figure 1 is stability. If $f(b,\cdot)$ or $g(b,\cdot)$ has an erroneous fluctuation as a function of b, then the new $P(a,b,\cdot)$ would reflect that error at the same values of b. The next calculation of $f(\cdot,\cdot)$ and $g(\cdot,\cdot)$ would be worse in that area, etc.. The intrinsic sources of the distortion will be discussed later; the result was that beyond six or seven iterations the accumulated error made the results of little value.

The numerical calculations were carried out with two major objectives. The first was to show that the algorithm was reasonably tractable; the second was to demonstrate an improvement over a linearized filter. Efforts in both areas were met with a reasonable degree of success.

Example 2

The purpose of this example is to show that by replacing

equation 5.1 with a difference equation and 5.2 and 5.3 with finite summations, one could start with an arbitrary initial distribution and by alternately updating the density and then $f(\cdot,\cdot)$ and $g(\cdot,\cdot)$, a joint density with high correlation would evolve. It was found that the smooth bivariate normal with zero correlation was the best initial density to start with. The terms in the state and observation equations were taken to be linear so in theory the coefficients of the filter should come out linear, as indeed they do. The resulting $f(\cdot,\cdot)$ and $g(\cdot,\cdot)$ appear in figure 2. They showed very little change of shape as time passed. The mean square error, on the other hand, began a quick descent. See figure 3.

A most interesting phenomenon is the speed with which $z(\cdot)$ tried to align itself with $x(\cdot)$. The density was approximated by a 200X200 point grid representing a plus and minus of ten units for each variable; call it ±10 inches for the sake of discussion. As a result the space increments were 0.1 inches. Time increments as small as .005 seconds were sufficient to see a considerable change from the symmetric, independent, bivariate normal to the skew symmetric, correlated, two dimensional normal in just a few iterations. This is contrasted by the relatively slugish response of the density to changes in the state or observation equations.

Figure 2 is rewarding in that for small values of the space arguement b, $f(b,\cdot)$ and $g(b,\cdot)$ are linear and agree with the Kalman filter. There is more information in the graphs.

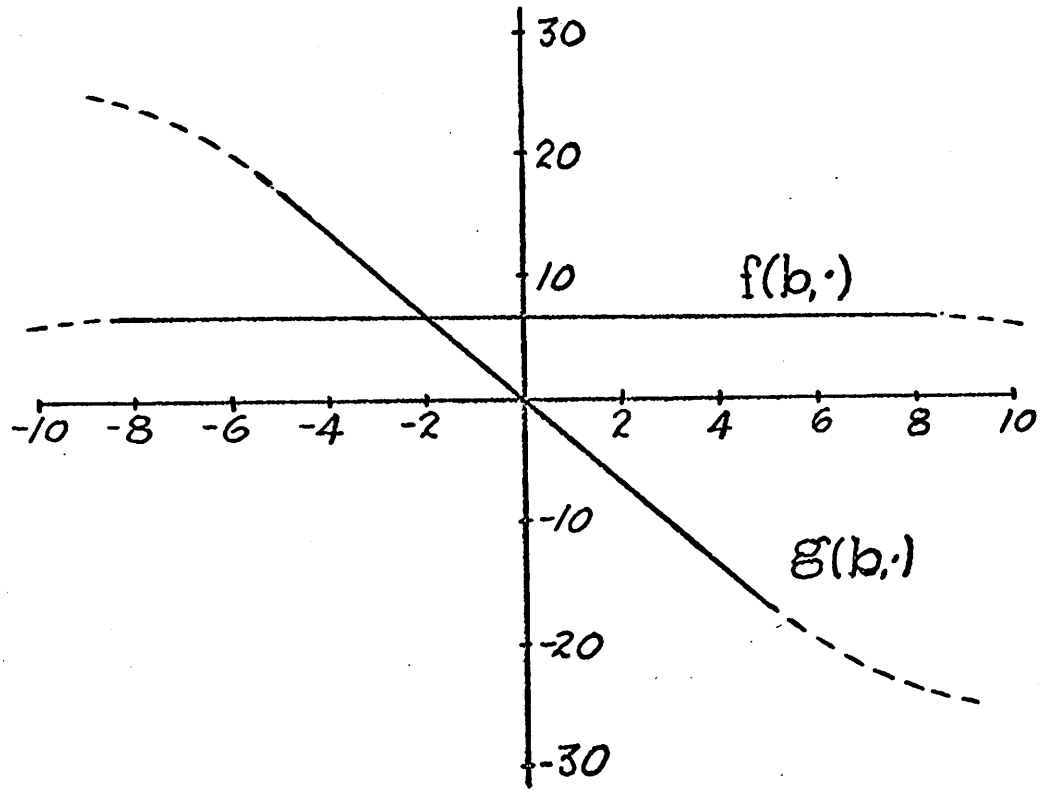


figure 2

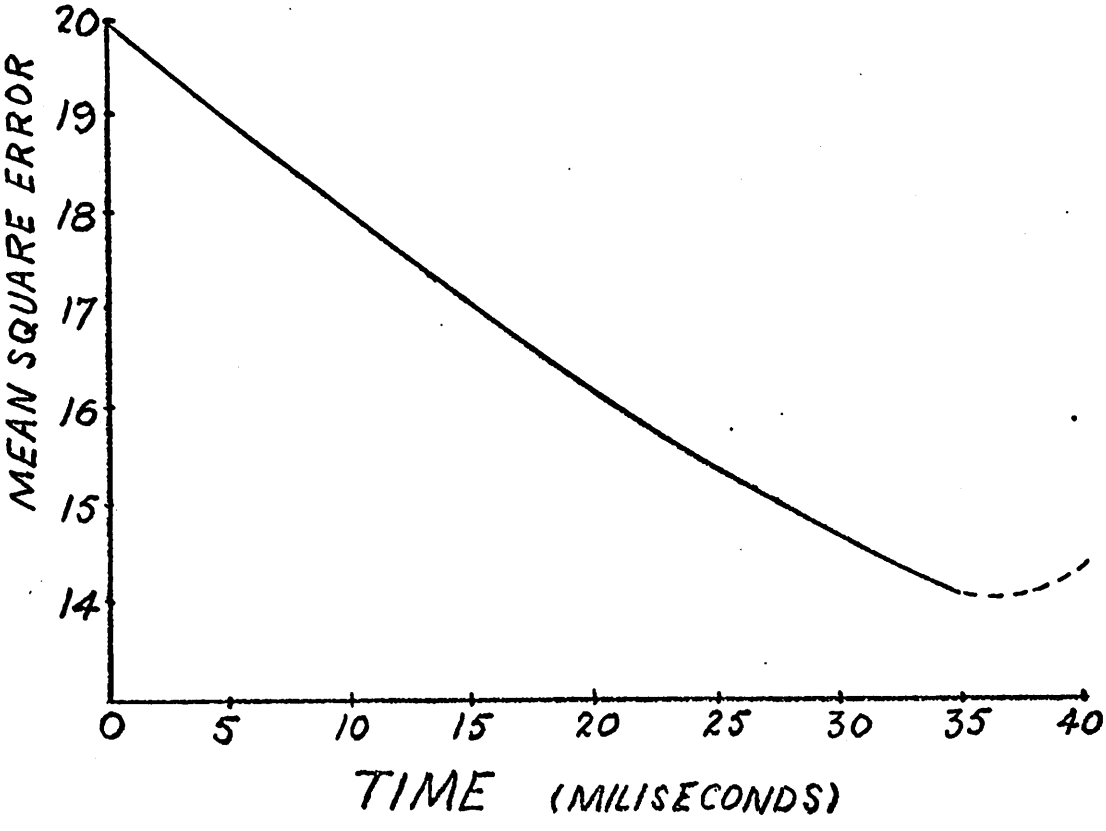


figure 3

If one calculates the conditional expectation for a bivariate normal, the result is a linear function of the conditioning variable; $g(b,\cdot)$ contains such a term. If a two dimensional normal is truncated to a 20 inch square, the conditional moment functions are going to have to bend near the periphery of the domain. This is what is happening in the dotted section of the graphs of $f(\cdot,\cdot)$ and $g(\cdot,\cdot)$. This the major unavoidable error referred to earlier. With each iteration, samples of the density were printed out and one could watch the error accumulate around the periphery of the square. The density was normalized after each updating and the normalizing factor (the sum of all the numbers on the grid) provided an index of the accuracy of the system. The number stayed very close to unity during the reliable part of the evolution, then jumped to a large number with the simultaneous degeneration of the density.

Two things were learned from example 2. The first was that digital computation of a solution of equations 5.1, 5.2, and 5.3 was feasable. By alternately updating the density and $f(\cdot,\cdot)$ and $g(\cdot,\cdot)$, one could indeed grind out the sequentially best filter. Because of the complixity of equation 5.1 only two and three point approximations of the first and second derivatives were used. This was the subject of considerable concern with respect to the error that may be incurred. The second thing learned from the example was that the error introduced by truncating the domain of the density

is the real cause for concern and is the factor that limited the number of iterations before complete degeneration occured.

Example 3

A linear system is assumed to have started at $t = -\infty$. By using the theory of Kalman and Bucy and the proper selection of coefficients it was possible to calculate the stationary distribution of $x(\cdot)$ and its estimator $z(\cdot)$ and find a nice fit to the truncated domain used in the first example. The linear functions in figures 4 and 5 are such that the stationary distribution of $x(\cdot)$ and $z(\cdot)$ is a correlated bivariate normal with error $E(x(t) - z(t))^2 = 0.93$, and very small values for $x(\cdot)$ $z(\cdot)$ equal to plus or minus ten.

Incrementing began at $t = 0$ with $\Delta t = 0.1$ seconds. With everything linear the error was maintained almost constant for seven iterations while the optimal $f(\cdot,t)$ and $g(\cdot,t)$ for $t = 0, .1, .2, \cdots, .7$ were generated and stored on magnetic tape. See figures 6 and 7. Notice that as in the first example the functions are linear near the center of the domain and agree with the Kalman filter.

Having carried out the above for a control group as well as generating the computerized version of the linear filter, nonlinearities were imposed on the system. Coefficients $m(b,\cdot)$ and $n(b,\cdot)$ were changed to those shown in figures 4 and 5. The resulting $f(b,\cdot)$ and $g(b,\cdot)$ are shown in figures 6 and 7. The change of coefficients was accompanied by a growth of mean square error. The errors are compared in

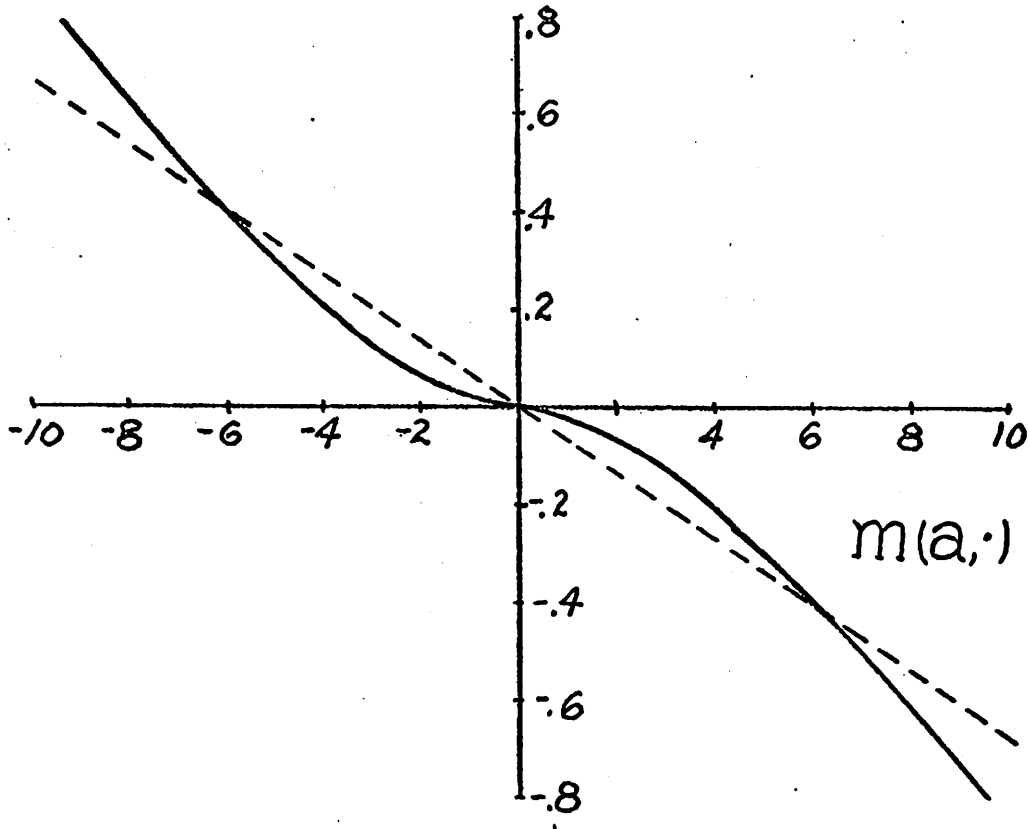


figure 4

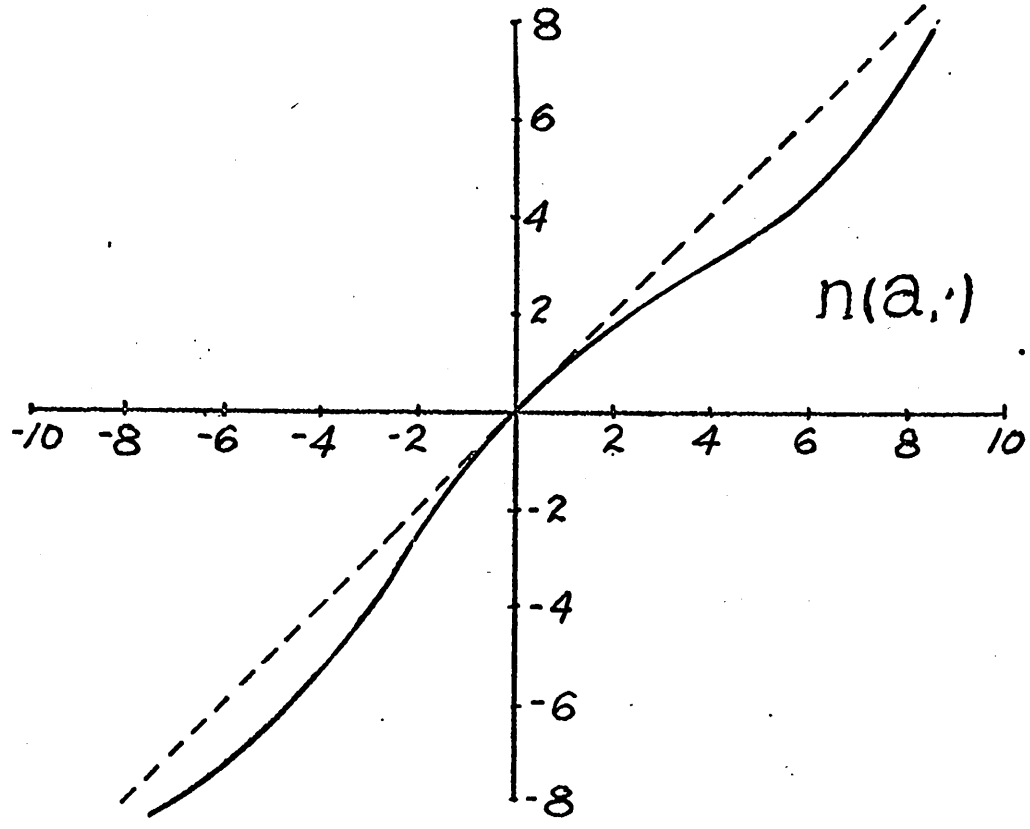


figure 5

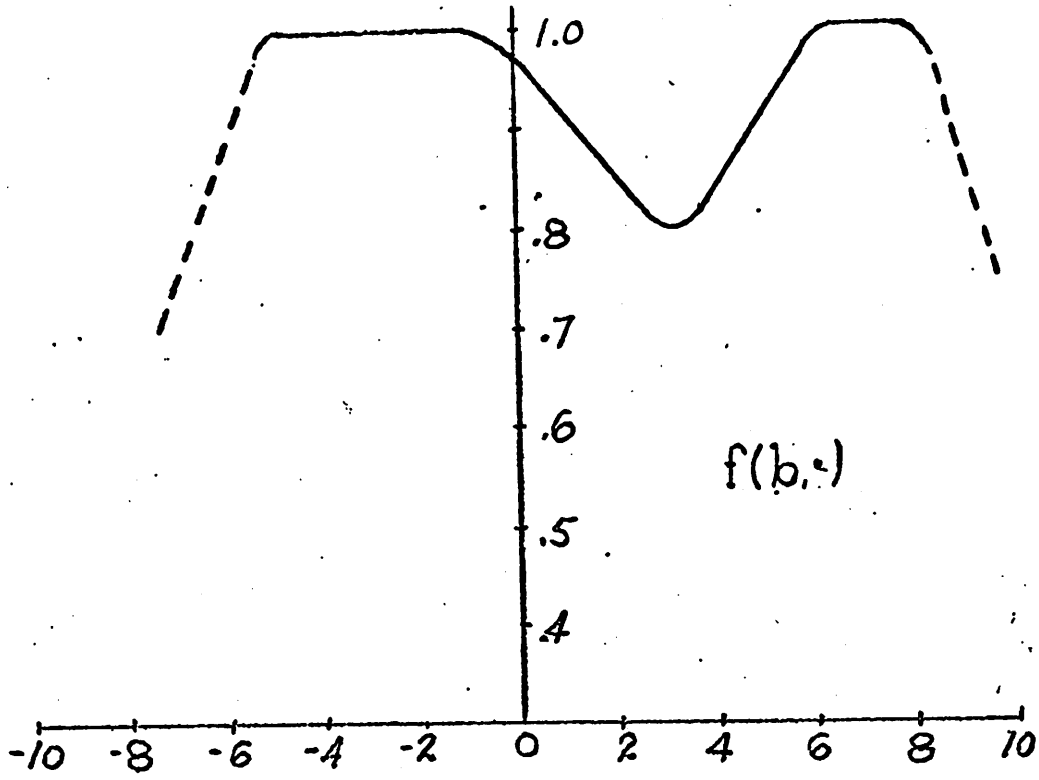


figure 6

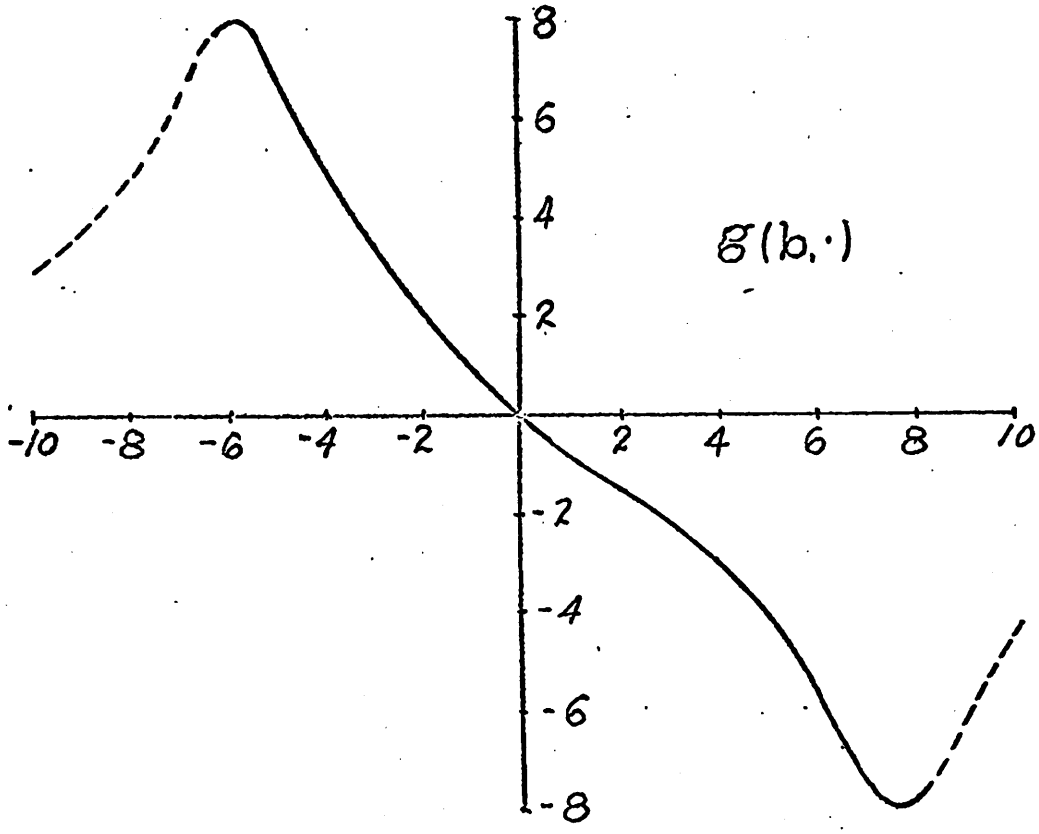
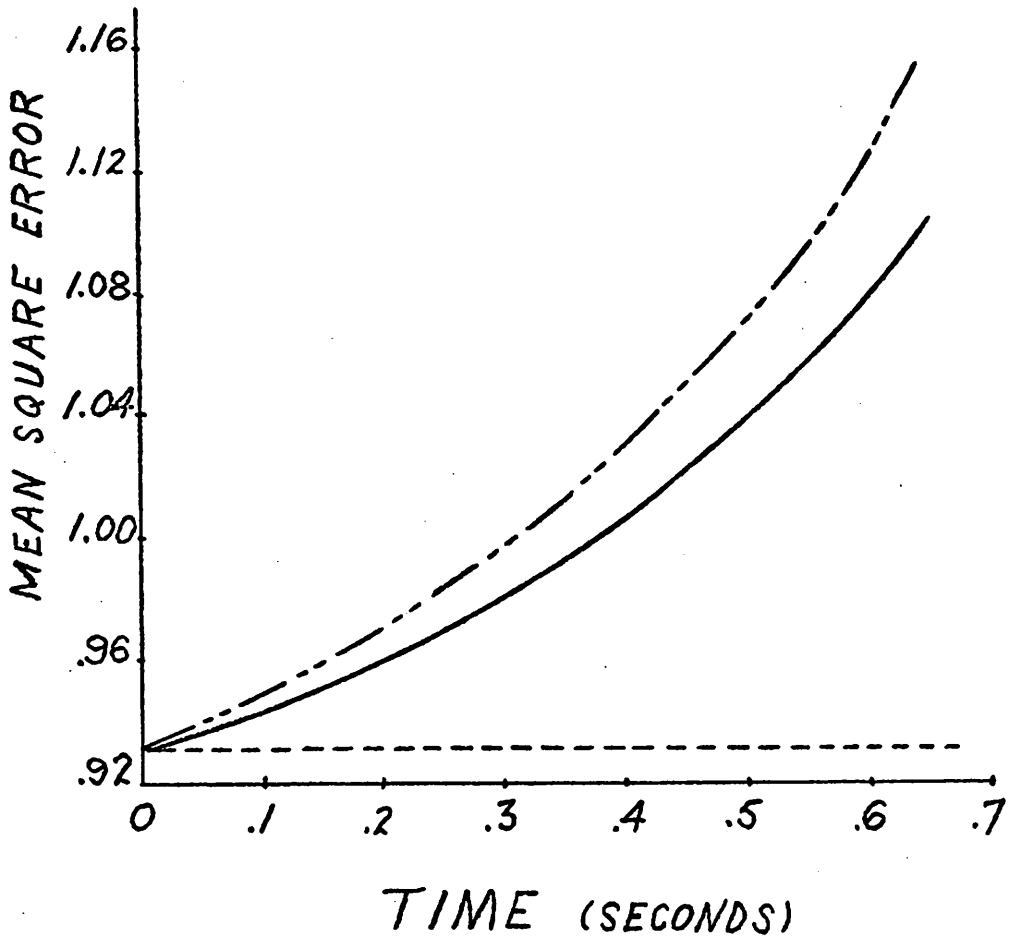


figure 7

figure 8. Lastly the evolution of the density was recomputed with the nonlinear $m(\cdot,\cdot)$ and $n(\cdot,\cdot)$. This time instead of calculating $f(\cdot,\cdot)$ and $g(\cdot,\cdot)$ each time around the loop, the functions generated in the linear case were read in from the tape. This simulated filtration of the nonlinear system using the best Kalman or linear filter. The resulting growth of error also appears in figure 8. The fact that the sequentially best scheme had a smaller error than the linear filter on the nonlinear system was satisfying reward for the time and effort spent in programming.

Example 3 is in truth the last and best of several attempts to force the initially linear model into nonlinear behavior. The difficulty was that if only $m(\cdot,\cdot)$ is perturbed slightly from the linear, the reaction for the density is so slow that errors stopped the process before it could deviate appreciably from the normal. If $m(\cdot,\cdot)$ is changed too violently, the difference equation is so coarse that the increments of the density are erroneous and degeneration again occurs.

In retrospect it is obvious how to improve the numerical analysis to obtain more convincing results with regard to the merit of the sequentially best estimator. However, this would involve a major revamping of the program which is not justified here and now. The major modification would be a redistribution of the grid points. A look at a highly correlated bivariate normal reveals that most of the probability



---------- Linear Case, Linear Filter

---------- Nonlinear Case, Linear Filter

———————— Nonlinear Case, Sequentially Best Filter

figure 8

mass is in the first and third quadrants. A square grid system wastes many points in the extremities of the second and fourth quadrants where the value of the density is negligable. Reapportionment in this case would make more efficient use of the storage locations and have the effect of increasing the capacity of the computer memory. If thereby the instability caused by truncation is reduced, some very interesting and useful results should follow.

CHAPTER VI

SUMMARY

No research result is an isolated point. It is at best an epsilon extension of the vast domain of accumulated knowledge. A valid summary, therefore, is a description of a neighborhood including both the facts and concepts which gave birth to the result and the extensions and improvements eminating from that result.

Chapter II answers a question of primarily academic interest. The Prokhorov density on function space is generalized to include the model of "noisey state-noisey observation." The conclusion that the distribution of the terminal value of the state, conditioned on the total observation, is thence a corollary to the work on the continuity of measures on function spaces.

In Chapter III a sophisticated and rigorous proof is given for the stochastic partial differential equation satisfied by the unnormalized version of the conditional density. Its uniqueness lies in the successful separation of the effects of updating the state variable from updating the observation curve. The mechanics of the proof are of further interest because a change of order of integration takes place. A function space integral and stochastic integral are commuted.

The existence of the space derivatives of the conditional density had to be hypothesized. While no direct proof of

the differentiability of the unconditional density has as yet been accomplished, a knowledge of its smoothness makes the hypothesis plausable. The function space integral and the stochastic differential seem to be the best framework within which one may prove the differentiability of the transition density, then go on to differentiating the conditional density.

Chapter III illuminates the impossibility of realizing a recursive scheme for the best estimator. Chapter IV is an attack on the other flank, to design the best recursive filter. With specifying the form of the filter, the problem becomes illdefined, even with mean square error. The time at which the error is to be minimized becomes significant. This ambiguity is exploited by choosing the time dependence of the error in such a way that the problem becomes more tractable yet maintains enough practicality to still be of interest. The concept of "sequentially best" embodies the goal of updating the estimator not the output of a black box. As a result of this point of view the design of the filter itself becomes recursive. Having accepted these constraints it is not too difficult to find the algorithm whose solution will grind out the sequentially best filter.

The most exciting question raised by the above scheme is: How can the gap between the sequentially best filter and the conditional expectation estimator be bridged? It appears that increasing the dimensionality to simultaneously make the best estimator and to bleed the most information from the

new data is the most promising course.

The computer results of Chapter V exhibit the feasability of off-line computation of the sequentially best filter. The shortcomings of the examples in turn illuminate some of the pitfalls to be avoided while implementing the procedure.

Nonlinear filtration is a tough problem. It is hoped that the contribution of this thesis will be two-fold. One, to guide development of an improvement over existing linearizing schemes, and two, to germinate new and fruitful interpretations and solutions to the question.

REFERENCES

1. Prokhorov, Yu. V. "Convergence of random processes and limit theorems in probability theory;" Theory of Probability and its Applications, Vol.1, No.2, pp. 157-214, 1956

2. Stratonovich, R. L. "Conditional Markov processes," Theory of Probability and its Applications, Vol. 5, No.2, pp. 156-178, 1960

3. Kushner, H. J. "On the differential equations satisfied by conditional probability densities of Markov processes, with applications," SIAM J. on Control, Vol.2, No.1, pp. 106-119, 1964

4. Kashyap, R. L. "On the partial differential equations for the conditional probability distribution for nonlinear dynamic systems with noisey measurements,"Division of Engineering and Applied Physics, Harvard University, Cambridge, Massachusetts.

5. Mortensen, R. E. "Optimal control of continuous time stochastic systems,"Rept. No. ERL-66-1,(1966), University of California, Berkeley.

6. Duncan, T. E. "Probability densities for diffusion processes with applications to nonlinear filtering theory and detection theory," Tech. Rept. No. 7001-4, May 1967, Systems Theory Lab., Stanford University.

7. Zakai, M. "On the optimal filtering of diffusion processes," E.E. Publication No. 80, Oct. 1967, Israel Institute of Technology, Haifa, Israel.

8. Kalman, R. E. & Bucy, R. S. "New results in linear filtering theory," J. Basic Engr. (ASME Trans), 83D, (1961), pp. 95-108.

9. Kushner, H. J. "Dynamical equations for optimal non-linear filtering," J. Differential Equations, 2, (1967), pp.179-190.

10. Kolmogorov, A. "Über die analytischen methoden der wahrscheinlichkeitsrechnung," Math. Ann. 104, (1931), pp. 415-458.

11. Ito, K. Lectures on Stochastic Processes, Tata Inst. for Fundamental Research, Bombay, India, 1961.

12. Gelfand, I. M. & Yaglom, A. M. "Integration in functional spaces and its applications in quantum physics," J. of Math. Physics, Vol.1, No.1, pp 48-69, 1960.

13. Doob, J. L. Stochastic Processes, Wiley, New York, 1953.

14. Loeve, M. Probability Theory, Van Nostrand, New York, 1963.

15. Skorokhod, A. V. Studies in the Theory of Random Processes, Addison-Wesley, Reading, Massachusetts, 1965.

16. Wiener, N. "Differential spaces," J. Mat and Physics, 2, pp. 131-174, 1923.

17. Lee, R. C. K. Optimal Estimation, Identification, and Control, M.I.T. Press, Cambridge, Massachusetts, 1964.